Are Best Papers Really the Best?

Sophia Kolak
sophiakolak@cmu.edu
CMU
Pittsburgh, United States

Aidan Yang
aidan@cmu.edu
CMU
Pittsburgh, United States

ABSTRACT

Best papers are papers chosen in computer science conferences as the top papers in a given year. How best papers are selected have not been adequately explored by prior research. We conduct a qualitative study to find out who and how best papers are selected. Furthermore, we use time series and classification techniques to investigate the correlation between best papers and long term impact. We find that best papers are not often selected by a committee, but rather selected by PC general chairs. Our qualitative study results imply that best papers are not chosen due to impact but other paper qualities, such as writing quality and rigor of research evaluation.

1 INTRODUCTION

For the past decade, computing research has been emerging at an accelerated pace. Peer reviewed academic conferences, currently the field's most popular venue for publication, publish nearly twice as many papers as they did a decade ago [8]. This trend has made conferences integral to research trends and career trajectories in academic computing, but it has also contributed scientific information overload [1]. "Over-optimization" of publication metrics further contributes to this process, with many individuals seeking to maximize publications and citation counts [4]. The end result of both shorter publication cycles and an increased volume of papers that researchers are expected to discern both a paper's quality and its relevance at a substantially faster rate.

As in evaluating other large networks of information [3], academic researchers too may rely on *signals* to determine whether which of many papers is worth reading and/or referencing [5]. Citations counts, h-index, and other publicly aggregated bibliometrics, are some examples of signals that researchers might take advantage of. To a hiring committee, for instance, one paper with 500 citations may signal a researcher's merit, whereas 500 papers each with one citation may signal risk. Correspondingly, a large number of citations can signal legitimacy and impact to a fellow researcher, and vice versa.

At the time of publication, however, citation counts and other conventionally accepted signals of a paper's merit are not yet available, meaning alternative signals (if any) must be used to determine a paper's worth and relevance quickly.

To this end, we turn our attention to distinguished papers, a title awarded to a small subset of publications at a given conference, typically emphasized in proceedings and highlighted during the live venue. Unlike bibliometrics, which often can only be assessed in the months and years post-publication, the "distinguished paper award" is a signal available at a paper's first appearance in academic circles. These awards, in theory, have the potential to serve as relevant indicators, both in helping conference attendees turn their attention to the most significant research produced that year, and in furthering the professional development of those who receive them. From the perspective of the program committee, they may also serve as quality standards that future submissions should attempt to follow.

Still, distinguished paper awards are not well understood. Although many computing conferences have moved towards visibility with respect to their reviewing process [1], the criteria PC members actually use to award distinguished papers is unclear and typically varies by conference. What the award actually suggests to other researchers, and the extent to which these perceptions and realities correspond, is similarly opaque. In an attempt to clarify these questions, we perform a mixed-methods empirical study of distinguished paper awards. We first conduct interviews with previous members of best paper award committees to isolate specific features selected for at review time. We then pose similar questions to a random sample of conference attendees, in order to learn their expectations of best papers. Finally, using a database of best papers from 1996 to the present, we empirically study the extent to which the assumptions of both parties are corroborated by features in the data set, as well as with conventional bibliometric signals of impact over time.

- RQ1: What qualities do selectors look for when awarding distinguished papers?
- RQ2: What qualities do researchers expect to find in distinguished papers?
- RQ3: To what extent do distinguished papers correlate with the qualities that selectors and researchers expect?
- RQ4: To what extent do they correlate with bibliometric signals of impact?

2 RELATED WORK

2.1 Characteristics of papers

Pham et al. [8] studied the knowledge network using citation linkage to identify the development of sub-disciplines. Pham et al. built a knowledge network based on relatedness of conference venues, and a citation network based on citation counts. Based on their data from DBLP and CiteSeerX, they claimed that conferences constitute social structures, and computer science is becoming more interdisciplinary. Mubin et al. [7] studied the readability of award winning full papers at CHI from 382 full papers. Their results show that award winning papers have lower readability compared to non-award winning papers. Mubin et al. further studied the impacts of the authors' demographic attributes on likeliness of paper awards as well as their readability. Our work performs a mixed method study on the characteristics of best papers, and compare those characteristics across different disciplines.

[1] https://openreview.net/

2.2 Citation count studies

Chu et al. [2] performed a large scale study on citation patterns across 1 billion citations among 57 million papers over 54 years. Chu et al. found that as a field grows larger, the turnover of field paradigms slowed down and the top-cited papers receive disproportionately more citations than new papers with novel ideas. Lee et al. [6] studied the predictive power of conference venue related attributes to paper citation counts using regression models. They concluded that a higher age of a conference and a higher selectivity are the best predictors of a higher citation count for the same conference's papers. In contrast to comparing canon papers and new papers or characteristics of different conferences, our work focuses on the relationship between best papers and their long term impact as compared to non-best papers across all top conferences. Wainer et al.[9] used data from 12 conferences across 14 years to make observations on best paper awards and citation count. Wainer et al. collected citation data from Google Scholar on 19,421 papers, of which 168 are best papers. To evaluate whether a best paper has higher probability of being cited, Wainer et al. used a custom probability metric across the sum of all best papers in their dataset, and found that a best paper will receive more citations than a randomly selected non-best paper with 78% chance. Our work is the first to perform a qualitative study on how best papers are selected. Moreover, our work performs a time-series analysis of citation counts, and considers other impact factors in addition to citation counts.

3 METHODOLOGY

3.1 Interview Study

In order to provide a thorough analysis of our research questions, we perform a two-part study. For the first part, we conduct semi-structured interviews with academics who have

- Served on a program committee
- Served as Chair of a program committee
- Received at least one distinguished paper award

During our interviews, we focus on three general themes, namely

- The process of selecting distinguished papers (reviewer perspective)
- The experience of receiving distinguished papers (recipient perspective)
- The experience of reading distinguished papers (general perspective)

Through these interviews, we obtain a detailed understanding of the process by which distinguished papers are selected, the ways in which this process varies by venue, and the major factors reviewers select for. We also learn the ways in which authors themselves interpreted the award, which often varied between papers (when the interviewee had won multiple awards). Lastly, we document attitudes toward the award from the perspective of a generic conference attendee, and the ways in which this may affect its long-term impact.

We manually parse PC committee members from conferences of four sub-fields: human computer interaction (e.g., CHI), machine learning (e.g., ICLR, NeurIPS), programming languages (e.g., PLDI), and software engineering (e.g., ICSE, ASE). We sent out personalized invitation emails to 12 individuals fitting our criteria, and received 3 responses (i.e., 25% response rate). The 3 participants cover the sub-fields of machine learning, programming languages, and software engineering. All our participants have over 5+ years of paper reviewing experience, and 2/3 have served on best paper selection organizations.

3.2 Data Analysis

The second part of our study is a quantitative analysis, informed by our initial qualitative findings. Given limited time, we conducted only two interviews for this study, however, both subjects had satisfied all three of the aforementioned criteria (PC member, awardee, and conference attendee) over the course of their careers. As such, we honed in on the conferences they had described at the greatest length, namely, ICSE and PLDI. In our analysis, we pay special attention to these two conferences.

3.2.1 Citation analysis. We begin with a generic analysis of publication rates over the last few decades, establishing our initial claim that the raw number of academic documents is growing exponentially by year. Following this, we study how the number of distinguished papers has changed over time, as well as the way in which this relates to their relative impact. We then look at the rate at which distinguished papers accrued publications for ICSE and PLDI over time, to learn whether or not winning the award has a significant impact on a paper's citation rate. We scrape semantic scholar, dblp records, and googlescholar, to create a database of best paper awards, impact awards, the number of citations they have at present, and the number of citations they received per year, which we make publicly available for future research at the following link [2]

3.2.2 Impact prediction. To understand the impact of best papers, we collect title, author information (i.e., author name, author affiliation), conference information (i.e., conference name, conference sub-field, conference age) from best papers across 30 computer science conferences since 1996. We also collect additional papers that are not best papers. In total, we collect data from 1265 papers. We then collect Google Scholar citation count for each of the 1265 papers. we label each paper on whether it is a 10-year impact award winner or not. We then treat 10-year impact award as a binary, dependent variable, and all other collected paper information (e.g., sub-field, best paper award, number of authors) as independent, predictor variables. In other words, we model 10-year impact award as the target variable for binary classification. We then measure the amount of The goal of this model is to investigate if winning best paper is the most significant predictor for the 10-year impact award. If so, then we can more confidently say that best paper awards are correlated with long-term impact, as compared to other factors. To analyze the predictors with the highest explanatory power, we use a random forest binary classification model. We then take out a selected sample of predictors, and compute the decrease in model accuracy. We measure the accuracy of our random forest model using the model's Area Under the Curve (AUC). AUC is the area under the curve of the plotted valued from true positive rate and false positive rate. An AUC value close to 1 means that our predictors can strongly discriminate our target variable (i.e., 10-year

[2] link

impact award). Finally, a large AUC decrease after eliminating a predictor indicates a higher predictive power associated with that predictor. **Due to insufficient time and readily available data for 10 year impact awards, the classification model was not completed at time of submission**.

4 INTERVIEW RESULTS

From our interviews, we gathered there were two parts to best paper selection. The first is an *initial selection process*, where PC members nominate or shortlist a certain subset of accepted papers to be considered for the award. The second part is the *final selection*, typically managed by PC chairs in our case. Each of these processes is described in further detail below:

4.1 Initial selection process

The selection process varies by conference and sub-field. Both interview subjects had served on multiple PC committees, and described a few different methods of initial selection they had encountered.

4.1.1 Method 1: Voting. All of the PC members are asked to vote for a distinguished paper. From the votes, a list is created, which the subjects believed was used for the final determination, but were not sure how.

4.1.2 Method 2: Nomination Flag. Each reviewer has the option to digitally flag a paper they believe should be nominated within a reviewing software. These nominations may be used to make a final determination, but the interviewees were again not certain how they factored in. Both subjects mentioned having either never or only once selected this option.

4.1.3 Method 3: Shortlist & Nomination Flag. Initially, all papers are reviewed and scored. The top 10% of papers are then added to a shortlist, which is further deliberated on by the PC chairs until a final selection is made. Overall, this seemed to be the most commonly employed method. Both interview subjects mentioned that each time they had received the award, their paper had been received three strong accepts (perfect scores).

4.2 Final Selection

As mentioned, the final selection was typically handled by PC chairs, and was somewhat ambiguous to non-chair PC members. When asked more about these deliberations, the subject who had served as PC chair said that, given time constraints, they would typically choose between the papers on the shortlist based on the discussion the reviewing PC members had had surrounding the paper, as well as whether or not anyone had provided a nomination flag to any of the papers (noted to be uncommon), but would rarely attempt to read them. The subject emphasized particular challenges of this approach. For one, reviews would rarely state explicitly that a paper seemed "award worthy", so chairs often relied on numerical data—in particular, they prioritized those papers that had received three strong accepts. Discussions from other reviewers that happened to highlighting significance or impact were given heavy weight, but otherwise, the subject noted they rarely tried to select for impact directly. Instead, they searched for papers that were well-written with a compelling result, and ideally, some clever novelty.

4.3 Receiving Distinguished Paper Awards

Both subjects had won multiple distinguished paper awards, and described varied experiences upon receiving them. In some cases, they mentioned feeling the award was well-deserved, and expecting it prior to the notification. In other cases, however, they described quite the opposite. One subject mentioned being completely shocked by a recently awarded paper's selection, as they thought it had a narrow scope and was unlikely to be impactful. When asked further about why they though it received the award anyway, the subject noted that it was agreeable ("nothing to argue with"), and had mildly interesting results. Overall, he attributed the award to the paper's sound writing and lack of controversy.

4.4 Reading Distinguished Papers

Both subjects noted having slightly elevated expectations for distinguished papers, in terms of the quality of analysis, and that they were more likely to read them during or before a conference. At a conference, both subjects noted that they were much more likely to attend the talk for a distinguished paper than for an arbitrary acceptance, and were especially likely to attend and engage with the paper more deeply if it was from within their sub-field.

5 QUANTITATIVE RESULTS

5.1 Publication Rates

In our interviews, both subjects mentioned having increasingly little time to review a substantial number of submissions. To this end, we start our analysis by measuring the number of documents added to dblp [3], the digital computer science bibliography, per conference per year. Figure 1 shows this result.

Although dblp data includes both co-located conferences, workshops, and non-technical track archives, the records per-year provide general insight into the rate at which the volume of peer-reviewed academic material is growing. As shown, the overall number of records is growing exponentially, and has nearly tripled in the last three years alone. Of particular note is the unequal trend with respect to various venues. Where some conferences like WWW have a relatively stagnant rate of growth, others like AAAI and CHI are nearly doubling their size each year.

While this trend may be due to other factors, such as an increase in the number of overall submissions or the expansion of the field in general, the labor of peer-review and the information overload for academics remains troubling.

5.2 Citations & Awards

Perhaps due the aforementioned trend of higher publication rates in general, the number of distinguished papers awarded per year has also grown for most conferences. Figure 2 shows the number of papers that received a distinguished paper award per year by conference. In general, the increase in number of awards tended to follow the raw growth of the field as per Figure 1. Still, some venues continuously provide the same number of awards regardless of their growing publication rates, likely due to variation in award selection practices and/or conference culture.

[3] https://dblp.org/

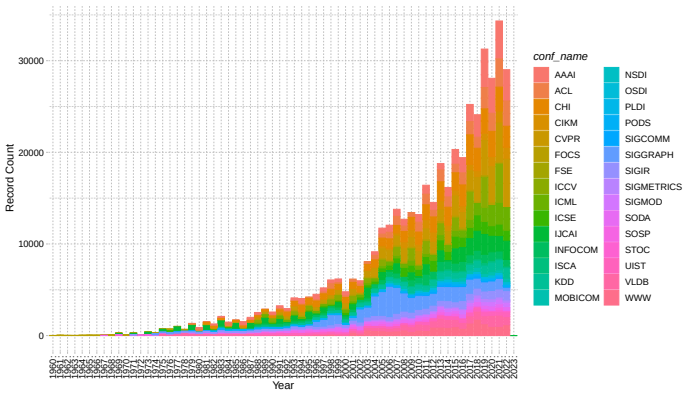

Figure 1: The number of records available on dblp by conference, per year (for each the 32 conferences included in this study).

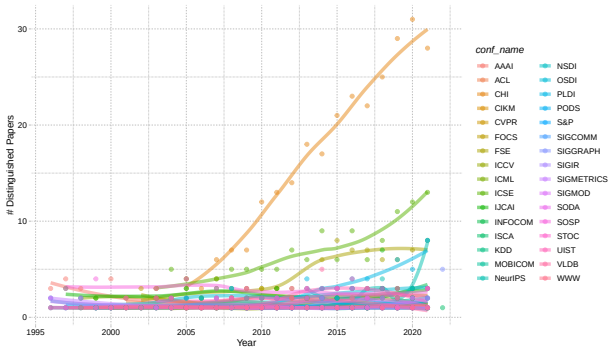

Figure 2: The number of distinguished papers awarded by conference, per year.

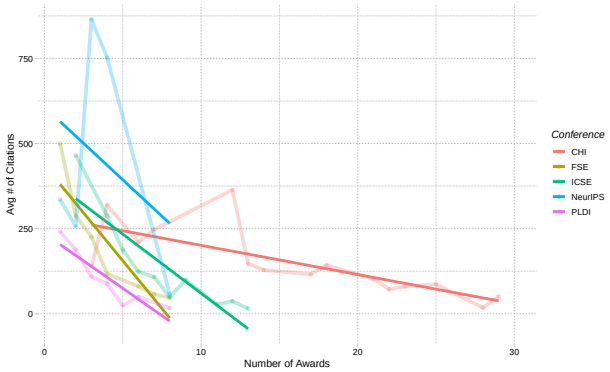

Figure 3: Relationship between the number of distinguished papers at a given conference at how many citations they received (on average), shown for five conferences.

Despite their ever growing prevalence, however, there is a distinctly negative relationship between the number of distinguished paper awards given at a conference and the average number of citations per paper. Figure 3 shows this inverse correlation for the five conferences with the greatest increase in the number of awards over time (CHI, ICSE, PLDI, NeurIPS, and FSE). Table 4 provides further detail, showing the Pearson's correlation coefficient for each of the five venues above, which reaches as high as -.92 for PLDI.

Conference	Correlation
ICSE	-.88
CHI	-.77
FSE	-.91
PLDI	-.92
NeurIPS	-.34

Figure 4: Pearson's correlation between the number of distinguished paper awards at a conference and their average number of citations.

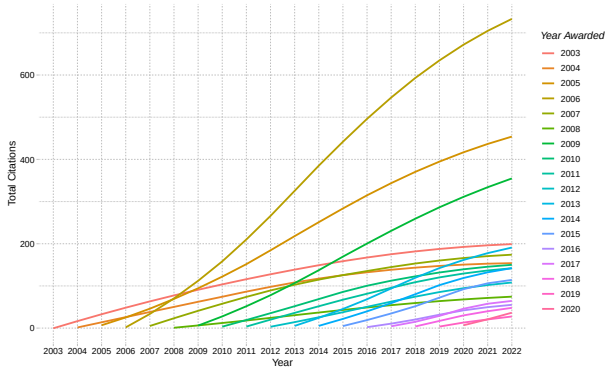

Figure 5: Citations by year for distinguished papers at ICSE

There are many possible explanations for this high degree of negative correlation. One plausible explanation is that, conferences typically increase the number of awards provided linearly with time. As such, papers which received the award when there were more awards distributed are likely more recent, and thus, have had less time to accrue citations.

Another plausible explanation, however, is that the increase in quantity has decreased the overall quality of submissions. This is supported by the fact that growing numbers of awards tended to parallel growing submission rates overall. With more submissions reviewed per PC member, and more papers published, both reviewers and readers may pay less attention to distinguished papers, and thus, do not interact with them as deeply as in years where there were both fewer publications overall and fewer awardees.

5.3 Citations over time

As our interview subjects were most familiar with the reviewing and submission processed for ICSE and PLDI (respectively), we perform an in-depth analysis of these two conferences for the following two sections.

First, we look at the rate at which distinguished papers for these two venues have accrued citations over time. Figures 5 and 6 show the total number of citations for distinguished papers at each venue (averaged per year of award).

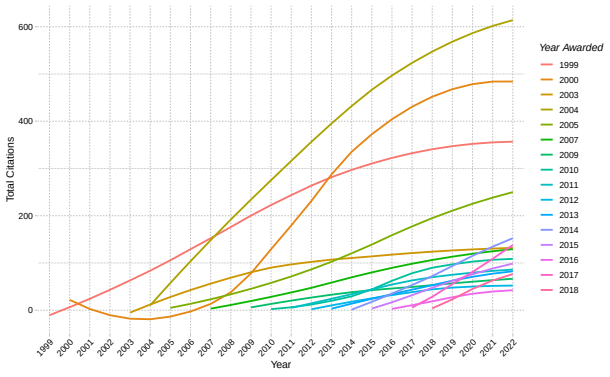

Figure 6: Citations by year for distinguished papers at PLDI

Overall, both plots show a relatively similar pattern. Still there are slight nuances. For ICSE, there is a noticeable difference in the rate of growth for the years 2006 and 2009, paralleled by some of the early 2010's (shown in blue). Whereas these years are best fit by an exponential curve, the remaining years are closer to a logarithmic growth rate. Of these steep sloped years, both 2006 and 2009, were years in which the 10-year impact awards (for 2016 and 2019) matched a distinguished paper awardee from the original conference date.

For PLDI, there is an interesting anomaly for the year 2000, which was slow to gain citations but then shot up around 2010. Interestingly, PLDI (until recently), was among those conferences to award the fewest number of distinguished paper awards per year, typically selecting 2,1, or even no awardees for a given year. In 2010, however, the singular distinguished paper awardee from 2000 was selected for the 10 year impact award, which likely explains its immense growth in citations more than a decade later. Possibly due to its small pool of awardees, PLDI papers tend to sustain a less stable growth rate (compared to ICSE, which tended to distribute more awards).

5.4 10-year Impact Awards

As pointed out in the previous subsection, the variable of a 10-year impact award (also sometimes referred to as the most influential paper award) had a noticeable effect on the citation rate, even among distinguished paper. To shed further light on the relationship between 10-year impact awards and distinguished papers, we document them here.

The 10-year impact award at ICSE and PLDI, as per the name, are given to the paper determined to have been most influential at the same conference 10 years prior. Unlike distinguished papers at these conferences, which tended to have decentralized or opaque determinations, the influence award is chosen by a committee specifically dedicated to choosing a singular paper that should receive the award. Unlike trying to gauge impact at review time, however, these committees have the benefit of time to inform their decision making. As such, the 10-year impact award was typically perceived by interview subjects as more meaningful.

Figures 7 and 8 allow us to measure, informally, how well reviewers can predict impact before a paper has stood the test of time. The years in which the paper awarded 10-year impact was also selected as a distinguished paper at its acceptance are shown in blue for both venues for all years in which there is available data. Comparatively, ICSE reviewers were quite apt at selecting distinguished papers that would go on to win the impact award ten years later, having done so in 4/10 cases.

PLDI awards, on the other hand, matched up much less frequently, with only 1/14 years corresponding. As noted previously, PLDI does award fewer distinguished papers per year. Still, on average for the years shown, ICSE awarded between 2 and 3, where PLDI awarded 1, making the difference relatively negligible.

6 THREATS & LIMITATIONS

6.1 Construct Validity

Construct threats to validity are threats from data collection biases. Given google's anti-scraping technology, we were not able to obtain the year in which all citing papers referenced a distinguished paper. In our data set, which consisted of 1260 best papers (between 1996 and 2022), approx. 200 (15%) contained a google scholar link (rather than a semantic scholar link). Although, given more time, semantic scholar links could have been extracted, we instead opted to only gather the total number of citations for these papers. As such, these papers are omitted from our last two plots.

6.2 Internal Validity

Internal threats to validity correspond with the relationships between our independent variables (i.e., paper related attributes) and the dependent variable (i.e., 10-year impact award). Although we tried to include, to the best of our knowledge, all paper related attributes, some attributes may have been missed that could be correlated to long term paper impact. For example, median author age or distance between all authors on a paper. However, personal author information is too difficult to obtain.

6.3 External Validity

External threats to validity are concerned with how well we can generalize our results. We were only able to have three participants to answer our RQ1 and RQ2. Although our participants cover a wide range of conferences and computer science fields, it is difficult to generalize their experiences to the greater computer science community.

7 DISCUSSION & CONCLUSION

In this paper, we studied both the selection and long-term impact of distinguished papers, with a particular emphasis on PLDI and ICSE. In general, our results show that the way in which these papers get selected is clouded in ambiguity (even for most reviewers), and that the density of papers often requires PC chairs to look rather narrowly at a paper when choosing a final awardee from a shortlist.

In contrast, 10-year impact awards have a more thorough review process from a dedicated committee, and the award is typically perceived as more meaningful by other academics.

Overall, our study suggests that the impact of distinguished papers is relatively varied. As more papers are published per venue

Title	Year of Award	Year Published	# Citations	Distinguished
On the Naturalness of Software	2022	2012	216	No
A practical guide for using statistical tests to assess randomized algorithms	2021	2011	516	No
Oracle-guided component-based program synthesis	2020	2010	42	No
Automatically finding patches using genetic programming	2019	2009	411	**Yes**
Debugging reinvented	2018	2008	65	**Yes**
Feedback-Directed Random Test Generation	2017	2007	457	No
Who should fix this bug?	2016	2006	660	**Yes**
Is mutation an appropriate tool for testing experiments? [software testing]	2015	2005	149	**Yes**
Mining Version Histories to Guide Software Changes	2014	2004	262	No
Hipikat: recommending pertinent software development artifacts	2013	2003	125	No

Figure 7: Most Influential Papers Awarded at ICSE from 2013-2022 along with their citation count, and whether or not they received Distinguished paper at publication time (10 years prior). Note that ICSE did not begin awarding distinguished papers until 2003.

Title	Year of Award	Year Published	# Citations	Distinguished
Test-Case Reduction for C Compiler Bugs	2022	2012	117	No
Finding and understanding bugs in C compilers	2021	2011	565	No
Green: A Framework for Supporting Energy-Conscious Programming...	2020	2010	379	No
FastTrack: Efficient and Precise Dynamic Race Detection	2019	2009	633	No
A Practical Automatic Polyhedral Parallelizer and Locality Optimizer	2018	2008	746	No
Valgrind: a framework for heavyweight dynamic binary instrumentation	2017	2007	1724	No
DieHard: probabilistic memory safety for unsafe languages	2016	2006	374	No
Pin: building customized program analysis tools with...	2015	2005	3099	No
Scalable Lock-Free Dynamic Memory Allocation	2014	2004	141	No
The nesC language: A holistic approach to networked embedded systems	2013	2003	1196	No
Extended Static Checking for Java	2012	2002	927	No
Automatic predicate abstraction of C programs	2011	2001	621	No
Dynamo: A Transparent Dynamic Optimization System	2010	2000	787	**Yes**
A Fast Fourier Transform Compiler	2009	1999	331	No

Figure 8: Most Influential Papers Awarded at PLDI from 2009-2022 along with their citation count, and whether or not they received Distinguished paper at publication time (10 years prior).

per year, the number of distinguished papers also increases, however, as the number of awardees increases, the number of citations per awardee tended to decrease. Expanding on this trend, the significance of distinguished papers, both to those who receive them and the community, may continue to decline as time goes on. As per our quantitative and qualitative analysis, enacting a more transparent and consistent strategy for nominating and awarding distinguished papers, selecting fewer per year, and perhaps limiting submissions and/or acceptances overall is one way in which we might slow the academic information overload.

REFERENCES

[1]
[2] Chu, J. S., and Evans, J. Too many papers? slowed canonical progress in large fields of science.
[3] Donath, J. Signals in social supernets. *Journal of Computer-Mediated Communication 13*, 1 (2007), 231–251.
[4] Fire, M., and Guestrin, C. Over-optimization of academic publishing metrics: observing Goodhart's Law in action. *GigaScience 8*, 6 (05 2019). giz053.
[5] GARNER, C. A. Academic publication, market signaling, and scientific research decisions. *Economic Inquiry 17*, 4 (1979), 575–584.
[6] Lee, D. H. Predictive power of conference-related factors on citation rates of conference papers. *Scientometrics 118*, 1 (2019), 281–304.
[7] Mubin, O., Tejlavwala, D., Arsalan, M., Ahmad, M., and Simoff, S. An assessment into the characteristics of award winning papers at chi. *Scientometrics 116*, 2 (2018), 1181–1201.
[8] Pham, M. C., Klamma, R., and Jarke, M. Development of computer science disciplines: a social network analysis approach. *Social Network Analysis and Mining 1*, 4 (2011), 321–340.
[9] Wainer, J., Eckmann, M., and Rocha, A. Peer-selected "best papers"—are they really that "good"? *PLOS ONE 10*, 3 (03 2015), 1–12.