Findings and Implications from Data Mining the IMC Review Process

Robert Beverly
Naval Postgraduate School
rbeverly@nps.edu

Mark Allman
International Computer Science Institute
mallman@icir.org

ABSTRACT

The computer science research paper review process is largely human and time-intensive. More worrisome, review processes are frequently questioned, and often non-transparent. This work advocates applying computer science methods and tools to the computer science review process. As an initial exploration, we data mine the submissions, bids, reviews, and decisions from a recent top-tier computer networking conference. We empirically test several common hypotheses, including the existence of readability, citation, call-for-paper adherence, and topical bias. From our findings, we hypothesize review process methods to improve fairness, efficiency, and transparency.

Categories and Subject Descriptors

A.1 [**General Literature**]: Introductory and Survey; C.2.m [**Computer Communication Networks**]: Miscellaneous

Keywords

Conference review, review bias, paper review process

1. INTRODUCTION

Conference publication unquestionably plays a vital role in computer science for the timely dissemination of results to peers. Further, computer science conferences are now often viewed as the "go to" venues for our best and most polished work [8]—a role traditionally filled by journals across many scientific disciplines. This preference for conferences over journals naturally leads to a number of implications.

- First, competition is fierce at many computer science conferences, with acceptance rates frequently $\leq 25\%$ [5] for the top venues[1]. This is natural as conferences are crucial to gaining visibility for work and have a large impact on researchers' professional development.
- Second, review cycles are short and focused. While journals have the luxury of time to allow a conversation of sorts between authors and a stable set of reviewers, top conferences are focused on making binary accept/reject decisions quickly, often within three months or less. Time constraints and the sheer number of submissions can lead to less than ideal reviewing. Load is further compounded by the lack of shared state: conferences have little knowledge about previous versions of rejected submissions.
- Finally, given their importance, the decisions made by conference committees are held up to much scrutiny. This is exacerbated by a (necessary) lack of transparency into how a particular submission was dealt with and discussed within a program committee.

[1]e.g. a recent CCR accepted none of 13 submissions [16].

Recent discussions well-illustrate these points [2, 25, 24, 19, 10]. The community is often left feeling as though the processes for selecting conference papers need to be improved—even if there is little consensus on how to improve. However, ensuring fairness, improving efficiency, and increasing review transparency, all while maintaining high quality conferences, is a shared goal of authors, reviewers, and conferences.

Many suggestions have been made about how to improve the process of assembling a conference program. Below we offer our own suggestions. However, before delving into possible solutions we outline two constraints. First, the review process is fundamentally a human endeavor and therefore disagreement on outcomes will always be part of the equation. For instance, prior work shows that score distributions have many equally "good" papers near the boundary between accepted and rejected submissions [6]. This illustrates the inherent ambiguity in attributing value to papers due to human factors and preferences: individual reviewer differences, reviewer group dynamics, current hot topics, and conference topic emphasis [3, 4].

The second constraint is that the process cannot be fully transparent. While perhaps the most transparent experiment to date has been to publicly identify reviewers [12], this still does not capture the discussions about individual submissions or the set of submissions. Often authors are left wondering why their submission was not included in the final program. In the absence of information about the process, authors often cling to unsupported notions (e.g. "the committee is biased against non-fluent English writers" [9] or other bias [21]). While the reasoning of the program committee may be sound, there is currently no *quantitative* means to express or communicate that soundness. Therefore, we should strive to be as open as possible with authors such that they come away feeling confident that their paper was treated fairly (even if not agreeing with the result).

Within these constraints many suggestions have been made, and experiments tried, across many conferences. For example, to combat the suggestion that reviewers do not have the requisite expertise we stock program committees with well-known researchers. To help the community understand the process, the PC chair(s) will often lay out the particulars of the path from submission pool to final program in the proceedings. To combat feelings of bias we ensure that

reviewers with conflicts of interest with authors have no say in the decisions made about those authors' submissions. We sometimes use double blind review whereby authors are not exposed to reviewers. Further, to help authors understand the decision making, conferences ask reviewers to answer pointed questions about each submission. In addition to a free-form review, these questions can give authors more information about precisely how a paper was read by a reviewer. Finally, some conferences return a summary of the program committee's discussion to the authors in an effort to illuminate precisely which issues were sticking points.

In this paper we argue for a way to potentially improve the computer science review process: use the tools of computer science to analyze the data naturally generated by a program committee. In particular, we believe that data analysis has benefits: (*i*) during review to help the program committee chairs and members detect and address issues (e.g. bias) directly within the process; and (*ii*) after review by exposing aggregate properties and decisions of the process (e.g. to show no bias along some axis).

As an exemplar of this approach, we analyze data from the 2010 ACM Internet Measurement Conference (IMC) [1] review process. Our analyses focus on how such data might be used to *improve* the process by examining readability, citations, bids, and topics. We do not claim these analyses to be an exhaustive set, but rather use them to illustrate both useful tools and the approach. Finally, we note that data from a single conference is not enough to draw sweeping conclusions about our techniques. We stress that this research is initial work on a promising approach, and not the last word on the analysis or metrics.

2. IMC REVIEW PROCESS

The IMC 2010 process was driven by a program committee of 26 community members. The authors of the present analysis in this paper include a PC member and the PC chair. PC members were asked to "do all or nearly all of the paper reviewing" themselves. Our data set includes the full submissions, PC bids on the submissions, review assignments, reviews, scores, and final disposition of the 211 papers submitted. IMC accepts two types of submissions. IMC 2010 received 102 short papers which were to convey "less mature but promising work" [1] and were at most six pages, with a seventh page of references. Additionally, IMC 2010 received 109 long papers of up to 14 pages which describe "original research with succinctness appropriate to the topics they discuss." Submissions exposed author information to the reviewers and were written in English. Of the 211 submissions, 47 papers were accepted for the final program (24 long and 23 short). The review process revolved around the HotCRP [18] review system.

The exact review process varies by conference, but in our experience the IMC 2010 is typical of large top-tier venues in that it was iterative. The process started with bidding, in which PC members express a relative interest in each submission based on abstracts and titles. Additionally, at this time the PC members also reported conflicts of interest and were recused from any further involvement with conflicting submissions. The bids facilitate the next phase of the process in which the PC chair assigned each submission two reviewers. After the first two reviews for a paper were submitted, a quick discussion between the reviewers and PC chair was initiated to decide whether the paper should be rejected—with roughly 40% of the submissions being rejected at this point. If both reviews were quite negative and the PC chair concurred, the paper was removed from further consideration. Otherwise, another reviewer was assigned. After the third review was completed, another discussion ensued to decide whether the paper (*i*) was so strong it could be accepted without discussion in the PC meeting, (*ii*) was so weak it did not merit discussion in the PC meeting or (*iii*) was a reasonable candidate for inclusion in the program and hence should be discussed at the PC meeting. In a small number of cases ($\approx 5\%$) this discussion lead to soliciting an additional review. The final step of the process was a day-long PC meeting—attended by all but one PC member—in which final decisions were made.

Each PC member reviewed 19–23 submissions. In several cases the PC utilized external reviewers where additional expertise was needed. Further, all papers for which the PC chair had a conflict of interest were handled outside the HotCRP system by three senior members of the PC such that the PC chair had no visibility into the process. These reviews were eventually entered into HotCRP under a single "anonymous" pseudonym. We omit external reviews and anonymous reviews from our analysis below.

3. METHODOLOGY

We require the raw ASCII text of each submission to perform analysis. While all submissions were in PDF format, the myriad PDF versions and distillations rendered automatic extraction unfeasible. We therefore manually extract ASCII text from each submission, collate each paper's references, and verify correctness.

We note that we do not attempt to validate the accept or reject decisions of the PC for our dataset, or otherwise determine paper quality in a technical sense. Instead, we take the PC's decisions as ground-truth and evaluate our various metrics with respect to these final decisions. Therefore, no correlation for some metric between accepted and rejected papers only indicates that the metric was not discriminatory *in our dataset* (i.e., with respect to the IMC 2010 PC's decisions). Past "shadow PC" experiments show that a different PC would have arrived at a different program [14]. Thus, any extant bias may be attributable to the PC, or may be due to the distribution of paper quality among topics. Assessing paper quality is subjective and even with an independent evaluation of quality, it is difficult to tease out the ultimate source of bias. Our goal in this work is simply to discover and report any such discriminators. In future work, we plan to perform a more general analysis across a large cross-section of conferences.

We use several metrics and techniques to test our various hypotheses. We consider the population of submissions divided among long and short papers, and those accepted or rejected. To determine if there exists a dependence between a particular metric and a paper's accept or reject decision, we first compute the Pearson correlation coefficient [22]. The Pearson correlation varies from [-1,1] with 1 indicating exact linear dependence between two variables, -1 exact inverse linear dependence, and 0 indicating no relationship. We also calculate the permutation test (p-value) for any correlations. The p-value, ranging from [0,1.0], is a confidence measure that indicates the probability of two uncorrelated systems generating as high of a correlation coefficient by chance. A large p-value reduces the confidence

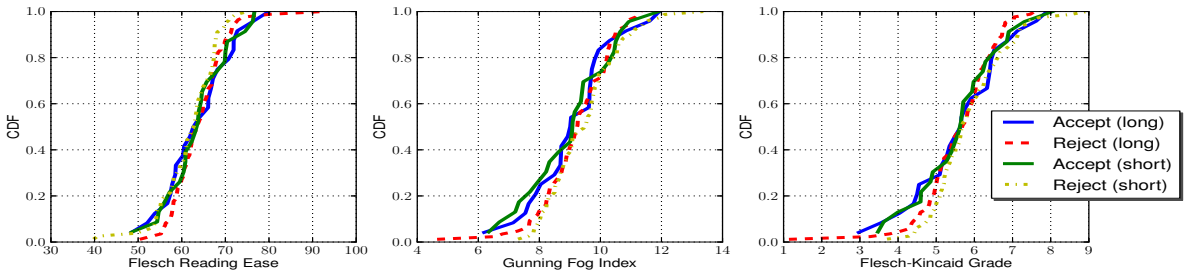

Figure 1: Distribution of readability metrics.

in the observed correlation. For topical emphasis and word discrimination, we employ a mutual information score, while a vector space model provides a document similarity metric; these discriminators are detailed in §4.4 and §4.5.

When considering distributions, we employ the two-sample Kolmogorov-Smirnov (K-S) test [20]. The K-S test forms a null hypothesis that samples from two populations (e.g. accepted and rejected short papers) are drawn from the same distribution. The test returns a K-S statistic from [0,1] and a two-tailed p-value from [0,1] indicating the probability of the null hypothesis. When the K-S statistic is small, or the p-value is high, we cannot reject the hypothesis that the distributions are the same (i.e. the metric under consideration does not discriminate between accept and reject).

4. MINING FOR BIAS

Amid a deluge of papers [15], even the most well-meaning, fastidious reviewer may introduce *unconscious* bias. For instance, a reviewer may unintentionally discriminate against particular topics, writing style, or methodology. Conceivably, paper discrimination may be intentional. Our goal is to facilitate automated mechanisms that expose potential biases so that the PC can evaluate, in real-time, whether they are intended and warranted, the result of the underlying quality of the submissions, or unintentional mistakes that should be addressed as part of the process. Further, statistics about bias could be exposed to the community to improve the transparency of the process.

While bias comes in a myriad of forms, we test for four specific types based on our experience and public discourse: readability, citations, topics, and keywords. We do not claim that these tests are inclusive of all sources of bias. Rather, they provide a starting point to rigorously examine common perceptions. Our methodology can be readily extended to assess additional forms of bias.

4.1 Readability

A paper is only as good as its ability to convey its contribution. In our discussions with authors, a common bit of folklore is that technically sound, but poorly written papers have a lower chance of acceptance—and, hence non-fluent English speakers are at a disadvantage. A 2008 SIGCOMM blog post exposed many of these same beliefs [9]. To explore this hypothesized bias we consider two metrics: vocabulary size and writing complexity.

As a first step, we seek to ascertain whether vocabulary size influences paper acceptance. We tokenize each paper in our dataset and compute the distribution of unique word counts across populations. In addition, we construct distributions after root-word stemming and stop-word removal.

We find that the minimum number of unique words for rejected long papers is lower than the minimum among accepted papers. For both long and short submissions, the maximum number of unique words is higher for accepted than rejected papers. However, the K-S test over vocabulary size between accepted and rejected papers yields p-values of 0.74 and 0.56 for long and short papers respectively—suggesting that vocabulary size was not a factor in IMC 2010 decisions. Using stopping and stemming does not qualitatively affect the results.

For a deeper understanding of writing level, we examine several widely-accepted readability tests [11] in Figure 1. The Flesch Reading Ease score is a measure of contemporary academic English comprehension difficulty, with higher scores indicating more easily understood writing and scores under 30 indicative of university-level text. The Gunning Fog [11] and Flesch-Kincaid indices[17] are similar measures that use the number of words, sentences, syllables, and complex syllables to determine writing grade level.

In our dataset, 60% of all papers have a reading ease score between 60 and 70 indicating relatively simple English. The Pearson's correlation between reading ease and acceptance (-0.01 and 0.12 for long and short) is not statistically significant and indicates that there is no reading ease bias. Further, the K-S p-value between accept and reject reading ease distributions is 0.71 and 0.45 for long and short papers respectively, indicating that the distributions are similar, and there is no evidence of bias.

Commensurate with the reading ease, we see relatively low inferred writing levels. The Gunning Fog K-S p-value is 0.89 and 0.34 for long and short papers, respectively. The Flesch-Kincaid p-values are 0.80 and 0.58 for long and short papers, respectively. The correlation coefficients in all cases are small with large p-values.

We note that readability metrics do not fully capture nuances particular to technical academic writing, e.g. precise and direct sentences. While there is no statistically significant bias evident for these three readability metrics, in future work we wish to examine other metrics such as grammatical correctness.

4.2 Citation Discrimination

A crucial component of any research is understanding prior work and the current state-of-the-art in the field. Citations serve as both requisite knowledge to understand a paper, and to distinguish submissions from similar prior research. Here we consider citation bias. Figure 2 displays the cumulative fraction of accepted and rejected papers versus the number of references contained in the paper. Unsurprisingly, short papers generally contain fewer references than long papers. A larger fraction of accepted short papers have

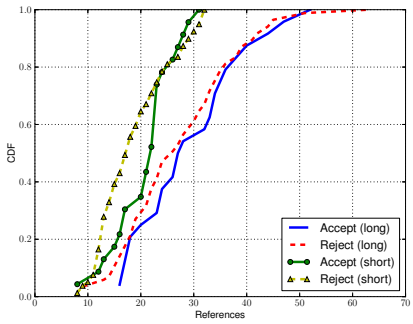

Figure 2: Distribution of references in paper.

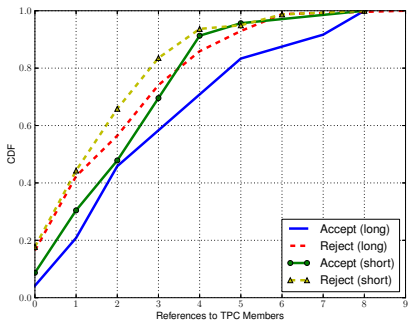

Figure 3: Dist. of references to TPC members.

more than 20 references as compared to rejected shorts (65% vs. 35%). All long papers with fewer than 15 citations were rejected, manual inspection of these found that half were less than the 14 page limit, suggesting that the authors ran out of time rather than space. We find a K-S p-value of 0.91 between the distributions of accepted and rejected long papers, indicating no bias. However, we find a weak positive correlation between longer reference lists and short paper acceptance: a K-S value of 0.30 with a p-value of 0.07, and a correlation coefficient of 0.14 with a p-value of 0.17.

Next, we consider whether citing work by members of the Technical Program Committee (TPC) is important for acceptance. For instance, one paper writing "strategy" is to attempt to engender reviewer favor by gratuitously citing papers authored by likely reviewers. Figure 3 shows the cumulative fraction of paper type versus the number of references to the IMC TPC members. Accepted papers, on the whole, contain more references to papers by TPC members than rejected papers. For long papers, the correlation is 0.21 which was statistically significant ($p< 0.03$), with a somewhat weaker correlation of 0.15 for short papers at a p-value of 0.12. The corresponding K-S p-values are 0.31 and 0.57. One explanation for this disparity is authors successfully biasing reviews positively.

To better understand the effect, we ran the same analysis of IMC 2010 submissions, but this time use the IMC 2009 TPC members who were not also IMC 2010 TPC members. There is less strategic reason to expect papers from 2010 to cite TPC members from 2009. Again, accepted 2010 papers generally contain more references to the 2009 TPC members than rejected papers. However, the effect is weaker: there is a 0.13 correlation for long papers (with a p-value of 0.19) and a -0.04 correlation for short papers (with a p-value of 0.66). We also find that 35% of accepted 2010 short papers reference none of the 2009 TPC members as compared to only 9% that referenced none of the 2010 TPC members.

Table 1: Example topic bias evaluation: $P(Y|token)$, where $P(Y) = 0.223$

Token:	**"wireless"** (unbiased)		**"ipv6"** (unbiased)	
	present	**absent**	**present**	**absent**
accept	0.216	0.226	0.235	0.222
reject	0.784	0.774	0.765	0.778
Token:	**"p2p"** (mild bias)		**"qos"** (bias)	
	present	**absent**	**present**	**absent**
accept	0.145	0.255	0.115	0.238
reject	0.855	0.745	0.885	0.762

While the Pearson correlations are stronger for the 2010 TPC citations than the 2009 TPC citations, the K-S scores show a stronger relationship for 2009 TPC citations. Future work includes analyzing more conferences to determine the extent of this effect. For instance, the correlations we observe may be due to citation bias, or may simply be because the TPC members are experts in their research domains – and therefore more likely to have authored important works that are thus likely to be cited by accepted submissions.

4.3 Querying for Topical Bias

While biasing the review process against topics that are out of scope for a given venue is natural and expected, other forms of bias may be detrimental and should be exposed in order to permit the PC to understand its source and effect.

As a first step toward understanding topical bias, we employ a simple single word-based token model to investigate whether particular keywords were of significance in the final accept or reject decision. This subsection examines bias across all submissions without regard for whether they are short or long. Each document is tokenized for alpha-numeric characters separated by any type of whitespace, and then converted to lower-case. Let ϕ_i be an indicator variable for the presence of word i in a given submission. Define class labels $Y = \pm 1$ as "accept" and "reject." The acceptance rate, or class prior, is thus: $P(Y{=}1) = 0.223$. We count the token prior $P(\phi_i)$ and the conditional probability: $P(\phi_i|Y)$. To compute a diagnosis from these causal probabilities, we employ Bayes' rule: $P(Y|\phi_i) = \frac{P(\phi_i|Y)P(Y)}{P(\phi_i)}$. Words common across accepted and rejected papers impart no discrimination, e.g. "the": $P(the|Y{=}1) = P(the|Y{=}-1) = 1.0$, i.e. $P(Y|the) = P(Y)$.

Conditional probabilities provide a powerful tool to form targeted queries. For example, a TPC might consider IPv6 or wireless new and exciting, and wish to understand if there exists a bias in the current set of candidate paper decisions. Table 1 shows the conditional probability for four topical tokens we imagined might experience bias, based on our knowledge of current networking research. We see that the conditional probabilities of "ipv6" and "wireless" are close to the class prior (0.223), indicating that they impart no bias. Next, consider the presence of peer-to-peer via the token "p2p." The peer-to-peer token is present in 9 accepted and 54 rejected submissions and presents a mild negative bias: the acceptance ratio of 0.145 is less than the acceptance class' prior. Finally, some words have a clear bias. Empirically, we found that "qos," or quality-of-service, which appears in 3 accepted and 23 rejected submissions, has a stronger negative effect on outcome when present.

Such discrimination may or may not be intended, as reviewers may simply be tired of such topics or the papers in such a well-studied area may be making smaller contributions. Alternatively, the quality of papers submitted that

Table 2: Most interesting 10 tokens based on MI

ϕ_i	$P(accept\|\phi_i)$	$P(accept\|not\ \phi_i)$
revisit	0.611	0.187
allows	0.293	0.085
simulated	0.000	0.250
globecom	0.000	0.244
lot	0.078	0.269
nxdomain	1.000	0.212
iptps	0.000	0.242
traceroutes	0.500	0.194
"author"	0.800	0.209
discover	0.344	0.170
alert	0.000	0.239

contain these topical terms may be skewed. We advocate *exposing* topic bias as an integral part of the review process, but we do not take a position on what a PC or the chairs should do about such biases when they are uncovered.

4.4 Word Discrimination

While conditional probability is a simple and powerful means to query for topical bias, it does not take into account the fact that a token may be highly discriminatory for acceptance or rejection and yet only appear in a single paper, limiting our ability to generalize across all words. We therefore turn to the Mutual Information (MI) score [13] to determine the information value of each word:

$$I(\phi_i;Y) = \sum_{\phi_i\in\{0,1\}} \sum_{Y\in\pm 1} P(\phi_i,Y)\log_2 \frac{P(\phi_i,Y)}{P(\phi_i)P(Y)} \qquad (1)$$

The basic intuition of MI is that, when ϕ_i and Y are independent for some token i: $P(\phi_i,Y) = P(\phi_i)P(Y)$ and $I(\phi_i;Y) = 0$. In contrast, if ϕ_i is strongly correlated with the accept or reject decision, its MI will be 1.

We compute $I(\phi_i;Y)$ over all tokens i and rank order the results. Of the top 100 discriminatory tokens from MI, we list 10 of the most interesting terms, discovered via manual inspection in a process we envision similar to how a PC might use this data, in Table 2. As with other analysis in this work, we emphasize that correlation does not imply causation; there are many possible explanations to our observed data. Again, we seek to *expose* information.

Among IMC 2010 submissions, the second and third most discriminatory word tokens are "revisit" and "allows." We were at first surprised to find non-technical terms dominating the MI, however technical terms are often sparsely distributed. In contrast, the word "revisit" is apropos to IMC where the CFP explicitly asks for "reappraisal of previous findings." Thus, if a submitted paper contains the word "revisit," or other, similar words—e.g. "repeat," "breadth," and "updated"—it was more likely to be accepted.

If "allows" is present, the paper has a near prior probability of acceptance (i.e. same as the overall acceptance rate). However, if "allows" does not appear, there is a very high (91%) chance that the paper is not accepted. A random sampling of five rejected papers without "allows" revealed that all five were authored by non-fluent English speakers, suggesting that such papers contained poor English. Other similar words had a similar effect, e.g. "lot" where there was only a 7.8% chance of acceptance.

As we include citations in our tokens, references and authors appear in our top token list. Two terms that identify other conferences, "globecom" and "IPTPS" have high MI; no papers containing these terms were accepted. Similarity, three authors (anonymized in Table 2) proved discriminatory, with a higher-than-prior chance of acceptance.

Finally, among technical terms, "nxdomain," "traceroutes," "alert," and "simulated" were all in the top 100 highest MI terms. Fittingly for an applied Internet measurement conference, no papers containing "simulated" were accepted.

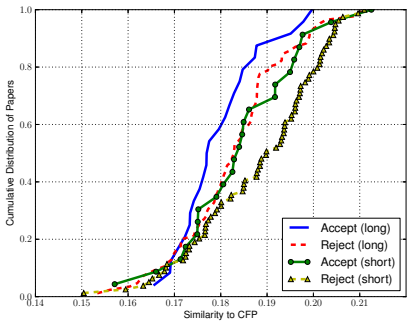

Figure 4: Similarity between submissions and CFP.

4.5 Adherence to CFP

A Call For Papers (CFP) attempts to outline the spirit of a conference. While CFPs often list specific topics (e.g. "peer-to-peer") and general approaches to be considered (e.g. "Internet measurement"), often such enumerations are explicitly labeled as non-exhaustive, e.g. to accommodate a new or fresh area not envisioned a priori (topics in the IMC 2010 CFP are framed as "examples.") That said, we assess how well the IMC 2010 submissions adhere to the CFP, which can help in identifying those with scope issues and how well the PC is evaluating with respect to the CFP.

To measure similarity between two strings, we employ cosine similarity, also known as the vector space model (VSM) [7]. Let the tokens of a given paper and the CFP be P and C respectively. We take the CFP verbatim, minus submission instructions, TPC members, and other non-pertinent details. Let $U = P \bigcup C$ define the universe of tokens under consideration, with $n = |U|$. Define two vectors of length n: $\vec{p}$ and $\vec{c}$ where each element represents the frequency of that token in the paper and the CFP respectively. The similarity between the vectors is the distance between them in an n-dimensional space. The dot product of the length-normalized vectors determines the cosine similarity:

$$sim(P,C) = \frac{\vec{p}\cdot\vec{c}}{|\vec{p}||\vec{c}|} \qquad (2)$$

Figure 4 depicts the cumulative fraction of submitted papers versus CFP similarity score. The results are surprising: accepted papers generally have *lower* similarity to the CFP than rejected papers. The K-S p-values between distributions of accepted and rejected paper similarity are 0.22 and 0.19, indicating an approximately 80% chance that they are different. Further, the correlation coefficients are both -0.11, indicating an inverse relationship with a p-value of 0.25. We conjecture that this similarity difference is due to bias toward new work rather than topics in the CFP. A second explanation is that lower-quality work is attempting to over-fit to the CFP. Again, our goal is to provide more information to the PC during the process and the community afterwards: we are more interested in exposing potential biases during the process than trying to post-facto determine why these occurred or judge them to be somehow "wrong."

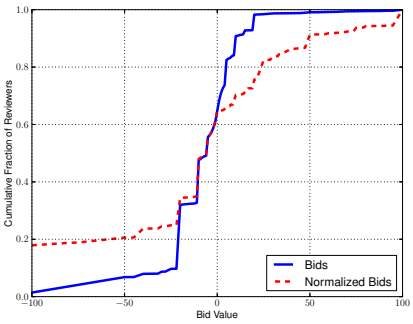

Figure 5: Distribution of reviewer bids.

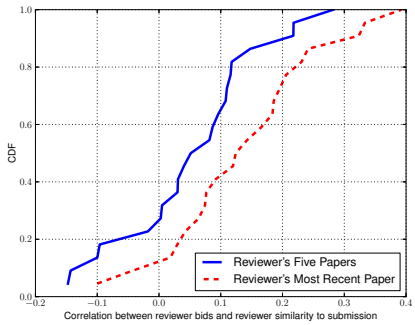

Figure 6: Distribution of correlation between reviewer's bids and reviewer similarity to submissions.

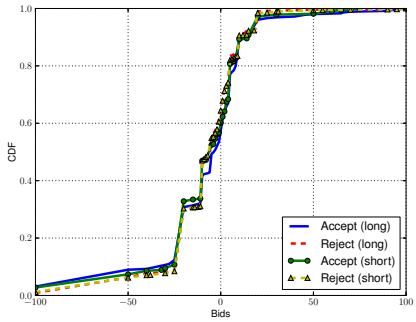

Figure 7: Distribution of paper type and disposition vs. reviewer bid scores.

5. UNDERSTANDING THE REVIEWER

Next, we aim to gain insight into the reviewers and the review process in order to improve the system.

Each reviewer has her own "lens" through which papers are viewed – adding healthy diversity and stimulating discussion within the process. One crucial job of the PC chair is therefore the assignment of papers to reviewers. The assignment objective is to find persons with sufficient technical expertise to evaluate a paper, while ensuring a variety of perspectives on each submission. Automation of the assignment process is feasible today. HotCRP [18] includes a bidding facility whereby each PC member reviews the titles and abstracts of submitted papers and indicates their level of (dis)interest. HotCRP can use bids to produce a candidate assignment schedule that balances the load across the PC. In our experience, however, it is important to retain a human in the loop and use the bids as a starting point as there are a variety of subtle factors to consider, e.g. ensuring that each paper is reviewed by a seasoned PC member. For IMC 2010, the chair collected bids, but did not use HotCRP's automated assignment generator.

The bidding process has become common across conferences in our recent experience. We also observe that bidding is a time-consuming affair, with PC members sifting through hundreds of abstracts to indicate their preferences. This is further exacerbated because bidding typically occurs before submission. IMC 2010 is normal in this regard, requiring paper titles, abstracts, and authors to be registered one week before the submission deadline. Bidding took place during this week and so PC members were forced to consider the 295 papers registered rather than only the 211 papers ultimately submitted. The burden of this task is evident in that only slightly more than half (14) the PC bid on all registered papers, with five members bidding on fewer than 100 papers and one PC member bidding on only 11 papers.

Figure 5 shows bid values as a cumulative fraction of reviewers. We also normalize bids per-reviewer, as individual reviewers can use vastly different scales. Note that a bid of -100 indicates a conflict of interest. We observe that only 32% of the papers received a positive bid, demonstrating that bids are primarily used to indicate disinterest.

We aim to understand the degree to which the bidding process could be automated to alleviate some of the aforementioned burdens. Rather than manual bidding, each reviewer could provide a small corpus of their own research papers. The system could then infer a level of interest in each submission, and produce a candidate bid.

To evaluate the feasibility of such automation, we manually selected a corpus of five published papers for each PC member, including one paper that is as closely related to IMC as possible. As with the submitted papers (P), we tokenize the reviewer's own papers to obtain R (we consider R inclusive of all tokens across the PC member's own five papers and most recent paper). Let $P_{i,j}$ be the i'th submission for potential reviewer j. We use VSM to compute pairwise similarity of each IMC submission to each reviewer's set of authored papers, i.e. $sim_{i,j} = sim(P_{i,j}, R_j)$. Figure 6 examines the correlation between the reviewer's bids and the similarity of the papers to her own work, i.e. $corr(\vec{sim}_j, \vec{bid}_j) \forall j$.

When considering only a PC member's most recent measurement paper, approximately 90% of the PC members submit bids that are positively correlated with the inferred similarity between the submission and the member's previous work. We find a similar trend when computing similarity against each PC member's five recent papers, but with less correlation, indicating that the most recent paper is most reflective of the reviewer's interests. This positive correlation suggests that automating the bid process may increase the efficiency of bidding while ensuring that all papers are bid on appropriately. Naturally, reviewers could override the suggested bids, for instance to express preference for a submission outside of the reviewer's traditional domain.

Finally, we examine whether a PC member's bid on a paper is correlated with that paper's eventual acceptance or rejection. We wish to determine whether reviewers are somehow predisposed to give positive or negative scores based on the bid values. Figure 7 shows the cumulative fraction of paper type and disposition versus the bid score. We see that the distributions are close across the range of bid scores, with a K-S test statistic of 0.08 (longs) and 0.07 (shorts), indicating that the two distributions are the same. Thus, we see no statistical evidence of bias based on the PC bids.

6. CONCLUSIONS

This initial work seeks to illustrate the potential power of deeper introspection into the computer science review process using data-mining techniques. We believe that expanding our understanding of the process has significant merit in ensuring fairness, organizational consistency, and promoting transparency. First, automated data-mining can provide valuable information to overworked PCs, and serve to enhance the system of checks and balances that ensure high-quality venues. While we performed all of the analysis in this paper off-line, the same techniques could be utilized to provide iterative or continual feedback to the chairs and TPC during, and within, the review process. Second, publicly exposing aggregate conference review statistics post-facto makes the process more transparent, thereby improving the integrity of the system. Our hope is to incorporate our techniques into popular open-source conference organization and review systems, e.g. EDAS [23] and HotCRP [18]. Tagging capabilities already present in HotCRP suggest an attractive means to expose some of the metrics we have presented.

We have only scratched the surface of possible hypotheses to investigate, and data-mining to perform. Our intent is to advocate a general technique and demonstrate instances of its application to a conference; space constraints preclude a more exhaustive analysis of many more equally valid questions. Additionally, the size of our dataset precludes any larger conclusions from our analysis. Future work will investigate more than a single instance of a conference, to understand how these techniques generalize across both conferences and time.

Acknowledgments

We thank Steven Bauer, kc claffy, Ryan Craven, Mark Gondree, Ratul Mahajan, Vern Paxson, and the anonymous reviewers for insightful critiques that greatly improved our analysis. Views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. government.

7. REFERENCES

[1] ACM SIGCOMM Internet Measurement Conference, 2010. http://www.sigcomm.org/events/imc-conference.

[2] IEEE ComSoc Technical Committee on Computer Communications mailing list archives, Aug. 2010. https://lists.cs.columbia.edu/pipermail/tccc/2010-August/thread.html.

[3] M. Allman. Thoughts on reviewing. *SIGCOMM Comput. Commun. Rev.*, 38(2):47–50, 2008.

[4] M. Allman. What ought a program committee to do? In *WOWCS*, 2008.

[5] K. Almeroth. Networking Conferences Statistics, 2012. http://www.cs.ucsb.edu/~almeroth/conf/stats/.

[6] T. E. Anderson. Towards a model of computer systems research. In *WOWCS*, 2008.

[7] N. Belkin and W. Croft. Retrieval techniques. *Annual Review of Information Science and Technology (ARIST)*, 22:109–145, 1987.

[8] K. Birman and F. B. Schneider. Viewpoint program committee overload in systems. *Commun. ACM*, 52(5):34–37, 2009.

[9] M. Crovella. Openness of the SIGCOMM conference, 2008. http://blog.sigcomm.org/2008/09/openness_of_the_sigcomm_confer.html.

[10] J. Crowcroft, S. Keshav, and N. McKeown. Viewpoint: Scaling the academic publication process to internet scale. *Commun. ACM*, 52(1), Jan. 2009.

[11] W. H. DuBay. The principles of readability. 2004. http://www.nald.ca/library/research/readab/readab.pdf.

[12] M. Faloutsos. IEEE Global Internet Symposium Open Review Process, 2007. http://netsec.cs.uoregon.edu/gi2007/.

[13] R. Fano. *Transmission of Information*. The MIT Press, Cambridge, MA, 1961.

[14] A. Feldmann. Experiences from the Sigcomm 2005 European shadow PC experiment. *SIGCOMM Comput. Commun. Rev.*, 35(3):97–102, July 2005.

[15] P. Francis. Thoughts on improving review quality. In *WOWCS*, 2008.

[16] S. Keshav. July 2011 editor's message. *SIGCOMM Comput. Commun. Rev.*, 41(3), 2011.

[17] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas. Technical report, Naval Air Station Memphis, 1975.

[18] E. Kohler. HotCRP Conference Management Software, 2012. http://www.read.seas.harvard.edu/~kohler/hotcrp/.

[19] H. F. Korth, P. A. Bernstein, M. Fernandez, L. Gruenwald, P. G. Kolaitis, K. McKinley, and T. Ozsu. Paper and proposal reviews: is the process flawed? *SIGMOD Rec.*, 37:36–39, September 2008.

[20] H. W. Lilliefors. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*, 62(318), 1967.

[21] K. Papagiannaki. Author feedback experiment at PAM. *SIGCOMM Comput. Commun. Rev.*, 37(3):73–78, 2007.

[22] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):pp. 59–66, 1988.

[23] H. Schulzrinne. EDAS Conference Management System, 2012. http://edas.info/doc/.

[24] D. S. Wallach. Rebooting the CS publication process. *Commun. ACM*, 54:32–35, October 2011.

[25] J. M. Wing and E. H. Chi. Reviewing peer review. *Commun. ACM*, 54:10–11, July 2011.