From Who You Know to What You Read: Augmenting Scientific Recommendations with Implicit Social Networks

Hyeonsu B. Kang
Carnegie Mellon University
Pittsburgh, PA, USA
hyeonsuk@cs.cmu.edu

Rafal Kocielnik
University of Washington
Seattle, WA, USA
rkoc@uw.edu

Andrew Head
Allen Institute for AI
Seattle, WA, USA
andrewhead@allenai.org

Jiangjiang Yang
Allen Institute for AI
Seattle, WA, USA
jjyang@allenai.org

Matt Latzke
Allen Institute for AI
Seattle, WA, USA
mattl@allenai.org

Aniket Kittur
Carnegie Mellon University
Pittsburgh, PA, USA
nkittur@cs.cmu.edu

Daniel S. Weld
Allen Institute for AI &
University of Washington
Seattle, WA, USA
danw@allenai.org

Doug Downey
Allen Institute for AI
Seattle, WA, USA
Northwestern University
Evanston, IL, USA
dougd@allenai.org

Jonathan Bragg
Allen Institute for AI
Seattle, WA, USA
jbragg@allenai.org

ABSTRACT

The ever-increasing pace of scientific publication necessitates methods for quickly identifying relevant papers. While neural recommenders trained on user interests can help, they still result in long, monotonous lists of suggested papers. To improve the discovery experience we introduce multiple new methods for *augmenting* recommendations with textual relevance messages that highlight knowledge-graph connections between recommended papers and a user's publication and interaction history. We explore associations mediated by author entities and those using citations alone. In a large-scale, real-world study, we show how our approach significantly increases engagement—and future engagement when mediated by authors—without introducing bias towards highly-cited authors. To expand message coverage for users with less publication or interaction history, we develop a novel method that highlights connections with proxy authors of interest to users and evaluate it in a controlled lab study. Finally, we synthesize design implications for future graph-based messages.

ACM Reference Format:
Hyeonsu B. Kang, Rafal Kocielnik, Andrew Head, Jiangjiang Yang, Matt Latzke, Aniket Kittur, Daniel S. Weld, Doug Downey, and Jonathan Bragg. 2022. From Who You Know to What You Read: Augmenting Scientific Recommendations with Implicit Social Networks. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3491102.3517470

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3517470

1 INTRODUCTION

Keeping on top of the literature while not being overwhelmed is a fundamental yet aspirational goal for many scientists today [47, 88, 123, 125] due to the immense scale of archival knowledge and its continued growth [18, 59, 85, 109]. While paper recommendation systems can help users identify useful papers from the larger literature, users can have difficulty understanding why those papers might be relevant to them or match their interests. Existing approaches for addressing this challenge include providing explanations of the recommender's behavior, or providing relevance messages that supplement recommendations, which may increase the persuasiveness and informativeness of the recommendations [16, 44, 49, 100, 102, 106, 108, 110–112, 117, 119, 126–128].

In this paper we compare the effectiveness of various types of relevance messages for scientific paper recommendations in a large-scale randomized study. Previous work has suggested the potential effectiveness of citation [15]-, knowledge [22, 23, 87]-, and social-[102] graphs for recommendations and relevance messages in various domains of recommendations. However, what type of messages is most impactful for scientific paper recommendations remains an open question. To answer this, we compare two types of messages that expose information about the interaction between the citation graph and the user. The first type of messages finds the intersection between reference papers in new paper recommendations—which are recently published and often do not have any incoming citations or rich metadata available—and papers the user previously published or personally curated (e.g., 'This paper cites 2 papers in your library'). We refer to this approach as citation-based relevance messaging that leverages *relevance via what they read*. Our second kind infers the implicit social network of authors from the citation graph to feature author-focused connections to recommended papers. Specifically, this inferred author network consists of relations targeted to our domain, such as which authors previously co-authored together or who has cited whose work (e.g., 'John Doe authored 3

papers you cited'). We refer to this approach as direct author-based relevance messaging that leverages *relevance via who they know*. Unlike prior work that used citation [15]-, knowledge [22, 23, 87]-, and social-graphs [102] to recommend or explain the recommendations, we use the citation and inferred author networks strictly to generate useful relevance messages that supplement the recommendations, in order to support a broad set of information needs that goes beyond the maximal content relevance [13, 21]. Our relevance messages are generated in a model-agnostic and post-hoc manner, which may generalize to new application contexts independent of the underlying recommender algorithms. To study users' authentic engagement patterns in a real-world scenario, we conducted a field experiment on an existing alert system of a popular scholarly search engine, which sends out emails with personalized paper recommendations to users who have opted in to alerts.

Through an iterative design process, we developed robust message designs for use in our two-month-long study with over 7,000 participants. Comparing the emails featuring direct author-based relevance messages to emails featuring citation-based relevance messages and Control, we saw the largest significant increases in user engagement from the emails augmented with direct author-based messages, which had user click-through rates that were 28% higher than Control. Direct author-based relevance messages also seemed to result in a higher level of future engagement, with 13% overall increase over Control in the email open rates, and 30% increase after the first two-week exposure to the messages. Furthermore, through an analysis of the distribution of clicked paper recommendations, we found that user engagement did not shift to papers written by authors with higher academic status when direct author-based messages were used compared to Control, suggesting that the messages were unlikely to exacerbate the *rich-get-richer* phenomenon [78].

However, the effectiveness of direct author-based relevance messages was limited by their scarcity; they boosted engagement more than citation-based relevance messages despite occurring much less frequently, on 4% of paper recommendations compared to 9%. In follow-up analyses using generalized linear mixed models, we show that substantially increasing the % of paper recommendations featured with direct author-based relevance messages is indeed a mechanism likely effective for further engaging users, even after controlling for potential covariates. To increase the coverage, we designed and implemented a new method for expanding the relevance relations on the implicit social network using *indirect* author-based relevance messages, which borrow from the networks of potentially familiar and trusted middle authors via [author]-[trusted author]-[user] triplet relations (e.g., assuming Dr. Anthony Fauci is a user-trusted middle author, "Catherine Paules has authored 4 papers that Dr. Anthony Fauci cited.; You saved 5 of Dr. Anthony Fauci's papers in the library."). In other words, the indirect author-based relevance incorporates implicit 'endorsement' from a known intermediate author, Dr. Fauci, whom the user may trust and from whom they may appreciate paper recommendations.

In a controlled lab study with fourteen scientists, we show the feasibility of indirect author-based messages for increasing message coverage. In addition, we gained qualitative insights into different types of benefits and challenges involved with different types of relevance messages. At a high-level, the benefits can be classified into two categories. The first category is benefits directly *on* recommendations, such as mobilizing the user's mental models of authors to gain a deeper understanding of recommended papers, or judging the potential usefulness of them. The second category of benefits is rather *around* the recommendations, such as developing awareness of other scientists they care about, understanding connections within academic communities, and understanding one's impact in the academic community. These findings confirm results from prior studies that cast recommendation as a socially embedded process that depends on both trust and the relationship of individuals [70, 92], but also surface new factors and design implications specific to scientific recommendations situated in a broader intellectual community.

In summary, this work makes five contributions. First, we designed and implemented two types of graph-based (citation and inferred author network) relevance messages for augmenting personalized paper recommendations, grounded in an iterative design process and interviews with multiple stakeholders. Second, we present evidence from a large-scale online deployment study showing that our messaging approaches indeed increased user engagement, and that direct author-based relevance messages performed the best, although their full potential was likely not reached due to low coverage. Third, we designed and implemented an additional, indirect author-based message that models endorsement from intermediate authors who are likely known and trusted, in order to further engage users by expanding the relations found on the inferred author network to mitigate the scarcity of author-based relevance messages. Fourth, through a controlled lab study with fourteen scientists, we show the feasibility of indirect author-based relevance messages and present qualitative insights into how different types of relevance messages benefited users. Finally, we present design implications for future augmentation approaches that aim to incorporate graph-based relevance information.

2 RELATED WORK

2.1 Exploratory Search Needs in Scholarly Recommendations

The information environment for today's scientists can be characterized by the immense scale and dynamic changes [18, 54, 59, 77, 85, 98, 109, 122], which make effective allocation of attention [103] an imperative for scientists. The issue of information overload [80] is especially pronounced in the crucial task of staying up-to-date with the relevant literature [65]. Compounded with domain-specific knowledge barriers that make the scholarly reading experience challenging [12, 51, 84], scientists' experience with the literature is often characterized as tedious, scattered, and relying upon chance discovery [20].

Exploratory search and curatorial needs for scholarly recommendations are high in this environment [65]. Additionally, what scientists judge as relevant may not only be logical—such as topical or narrowly defined as papers containing specific terms—but also situational, and depend on the scientist's personal information needs that go beyond the need for maximal content relatedness [13, 21]. To support the exploratory process of curating high-relevance papers, scientists commonly adopt two kinds of often effortful strategies. The first kind can be characterized by its use of *citation* networks which

enable scientists to search forward or backward in time through citation or reference chaining [9, 11, 30, 58, 115, 120]. Chaining in this manner requires scientists to maintain the relevance and coherence between papers as their number grows exponentially. The second kind leverages a different type of network, which is *social* in nature, to support serendipitous sharing and finding of recommendations on social media platforms such as Twitter [37, 55, 63, 81, 114] or cold-emailing high-profile experts in an outside field to receive valuable bibliography [30, 89, 90, 115].

To support these needs while reducing the burden of the laborious process, prior work has explored ways to automatically explain the recommended items' relevance to the users. Such explanations of relevance may be generated after the recommendations are derived [100, 126], and may be focused on providing explanations of the recommendation mechanisms themselves, which can better engage users by enhancing their understanding of the inner workings of the 'black-box' recommender algorithms [106, 108]. In contrast, rather than focusing on providing faithful explanations of the inner algorithm, our work seeks to increase the persuasiveness and informativeness [16, 102] of recommendations by incorporating *external* relevance signals. Existing work in this space has explored ways to incorporate relevant knowledge graph entities [22, 23, 87] or social signals such as local (e.g., showing the user's immediate friends' preference) or global (e.g., overall popularity) relevance [102] to increase persuasiveness. However, open questions remain as to how and what kinds of relevance information may be incorporated into scholarly recommendations to effect large-scale behavioral changes in the scientist user population, and which approach works best.

2.2 Behavior Change and Motivation

To study the large-scale changes among the scientists engaging with scholarly recommendations, our work builds upon the literature on behavior change and persuasion. These areas suggests several techniques, such as principles of authority and social influence [4, 25, 26], which inspired our designs for increasing user engagement. While these theories have been adapted in practice, our work is different from all prior explorations in key aspects.

The vast majority of prior work tested their motivation strategies only with crowd-sourced populations on platforms such as Amazon Mechanical Turk [40, 56, 64, 72]. Furthermore, most message-based behavior change designs have been applied in the context health & well-being [57, 64, 83], civic engagement [40], UI testing [72], or creative, brainstorming tasks [56]. Demographics and incentives of users on crowd-sourcing platforms and in personal contexts can be substantially different from our professional scholarly population [113]. Finally, translating theory-informed or only lab-tested motivation designs into real-world systems has been shown to be a non-trivial task due to feasibility limitations and ethical considerations [29, 40]. Aside from these general limitations of prior work, we further discuss in detail the specific differences in relation to selected works closest to our designs.

Grau et al. [40] designed motivation-supportive messages in the context of a crowd-civic platform, to engage volunteers, yet these designs did not explore leveraging *relevance* or *social graph*, focusing instead on aspects of *controlled* and *autonomous* motivation tested only with a limited sample of the Korean population. Kocielnik and Hsieh [64] designed message triggers that are diversified either by cognitively close concepts to the targeted action or the recipient, and found that the close-to-recipient diversification was more effective. This work offers insights about the importance of *closeness*, yet it was tested in a limited deployment with 27 participants and in a substantially different domain of physical activity. Unfortunately, factors affecting engagement can be substantially different across different settings and populations [14, 36]. McInnis et al. [72] focused on motivating one-time comments in an artificial task of "testing website interface" and hence is substantially different from motivating long-term engagement. Hsieh et al. [56] focused on exploring differences in user populations attracted by different incentives (e.g., monetary reward vs lottery), but did not explore the design of messaging or leveraging social graphs and was limited to motivating single-time study participation.

2.3 Social Network-based Relevance Information

Social recommendation has been a millenia-old mechanism for finding personally relevant items, but the advent of social media enabled its propagation at much greater scale. Social recommendations are also common in the scholarly domain as described above; scientists use platforms such as Twitter to share and access relevant papers to read. Theoretically, important factors of social recommendation include homophily [74], diffusion and influence [67, 105], and trust [69, 71], which originate from social network analysis. Prior work in social network-centric recommendations developed designs targeted at such factors to improve user engagement [42, 101]. However, open questions remain as to how the (social) network information may be incorporated in scholarly recommendations. No explicit network of scientists exists, and the interesting network structure may not even be inherently social in the sense studied in prior research. For example, while several works have studied the notion of *immediate* friends in network-centric recommendations and relevance explanations based on the theory of homophily (e.g., close friend groups may share similar tastes in music) and trust (e.g., 'if my friend likes this album, it must be good') (e.g., [42, 101, 102]), the importance of immediate relationships may fade away rather quickly in the scholarly domain. The insight here is that similar kinds of immediate closeness may not be as useful or even interesting to scientist users, as they may already be familiar with the work by their 'friends' (e.g., through co-authorship) and may instead benefit more from relevant yet novel work mediated through more distant connections. This implies characteristics of relevance diffusion and trust relations within scholarly networks may differ from those of the domains previously studied. Furthermore, important questions arise from the perspective of system-wide fairness, for example whether frequently featuring distant yet trusted authors results in negative externalities such as skewed distribution of work visibility and subsequent downstream reduction of impact from the work made relatively less visible (e.g., the Matthew effect [78]).

In this work, we contribute to the literature of social network-centric relevance messages in the domain of scholarly recommendations. We expand the design space by introducing a novel messaging technique that leverages an intermediate, trusted author to expand the coverage of relevance connections beyond the user's own history, in order to highlight additional papers or authors that may be of interest to the user. We contribute a large-scale deployment study

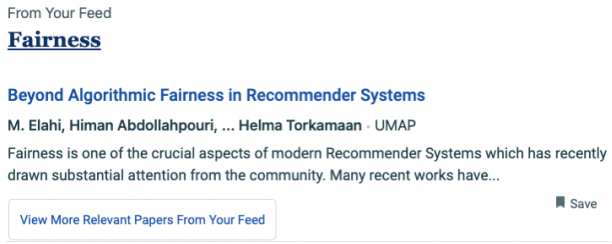


Figure 1: An abbreviated email alert.

and a controlled lab study evaluating the effectiveness of messages on inducing behavior and motivation changes. We further uncover the distinctive usefulness and challenges of social network-centric relevance messages. We end with a discussion of open questions remaining for future work.

3 RELEVANCE MESSAGE DESIGN FOR SCIENTIFIC RECOMMENDATIONS

Following an iterative design process, we designed two strategies for engaging users with email alerts containing paper recommendations: *citation*- and *author-based relevance messages*, which augment the recommendations. In this section, we describe our scientific recommendation setting and message design iterations, and present a detailed description of the designs themselves.

3.1 Email Alerts for Scientific Paper Recommendations

We designed our engagement strategies to work within email alerts sent from Semantic Scholar. The Semantic Scholar platform provided users the ability to search and curate interesting papers for personalized research feeds. Users could explicitly assign a binary positive ('more like this') or negative ('less like this') rating to each paper in their curated research feeds. Users then could opt in to receive alert emails when new relevant papers are found for each feed on a regular basis, with a user-selected frequency (e.g., daily or weekly). Typical alert emails started with the title of the feed, and a list of new relevant papers (usually ranging from 1 to 50; Fig. 1 shows a simplified version). Emails included basic paper metadata and link-based affordances for navigating to and saving papers, managing alert subscriptions, etc. Working within an existing production paper alert system lent ecological validity to our experiments. Due to the email setting, our design space did not include complex interactions such as hovers. The research feeds recommend papers using a neural recommender, trained on the user's individual paper ratings. Importantly, our relevance messages are agnostic to the underlying recommendation model, which was held constant across the conditions.

3.2 Iterative Design Process

Our designs went through four phases of iterations and prototyping.

Phase 1 - Theory & Expert driven brainstorming. Design brainstorming took place among three authors (two with knowledge of scholarly recommendation services and one with expertise in behavior change and persuasion) informed by literature from behavior change & persuasion (e.g., principles of authority and social influence) [4, 25, 26], recommendation systems (e.g., relevance and discovery) [8], and information processing literature (summarizing, balancing and diversifying the information) [60, 93]. Several strategies were eliminated on the grounds of: 1) challenging execution in deployment context, 2) data availability, and 3) ethics (e.g., scarcity, which may present a false impression of limited availability).

Phase 2 - GUI prototyping, data availability & technical feasibility evaluation. Selected strategies were prototyped in high visual and data model fidelities [34]. The visual fidelity prototyping explored text-based and visualization-based designs [38, 64]. The high data model fidelity prototyping ensured: 1) the designs worked as intended in the live system, 2) the complete record of interactions was preserved, and 3) a sufficient number of users were impacted on a daily basis.

Phase 3 - Graphical Design, Marketing & Engineering Feedback. Additional graphic design review ensured the use of an icon, font, and color scheme was consistent with the existing alert emails. Further marketing team feedback resulted in rephrasing parts of the messages to fit factual & information-centric tone (e.g., an early phrase *"Cites:"* was replaced with *"Also cites:"* to more factually reflect the presented numbers). Engineering team feedback led to improved data pooling, freshness and completeness.

Phase 4 - Beta testing. The implemented designs were beta-tested for two weeks with 100 internal users in actual live service to: 1) eliminate any technical issues and 2) collect feedback from users in more naturalistic setting. Interviews were carried out with seven users over Slack and via a video call, which confirmed the general understandability, visibility, and user interest in the presented relevance messages. Several changes were introduced following user feedback: 1) dropping a change in email title, as it was not noticed, 2) replacing the separator between message parts from '+' sign to comma, and 3) clarifying or removing unclear content.

3.3 Message Designs

We present the detailed description of the final designs.

Design 1: Citation-based Relevance Messages. *Goal*: Design 1 conveys potential relevance of an alert paper to the user via direct citations (Fig 2.4). It aims to highlight citations from the alert paper to papers the user has previously explicitly expressed interest in. We do not consider citations of the alert paper, since most alert papers are new and do not yet have many citations. Prior work in scholarly contexts has emphasized the important of citations as a simple measure of relevance [5, 43].

Design: We define *user relevant sources* to be the user's personal *library* and *research feed*, as well as papers the user has *authored*. These sources contain papers in which the user has explicitly expressed interest at some point. We further define a *relevant alert paper* to be a paper that cites one or more of the papers in the user relevant sources. Our *citation-based relevance condition* adds one of the relevance messages presented in Figs 2.2 and 2.3 to any relevant alert paper. If an alert paper does not cite any of the papers in the user relevant sources, no relevance message is added. Alert papers that cite papers from only one user relevant source receive one of the message variations presented in Fig. 2.2, while alert papers citing papers from multiple user relevant sources receive message variations presented in Fig. 2.3. Each message communicates information about the number of cited papers from a particular user relevant source, as well as provenance information about the source

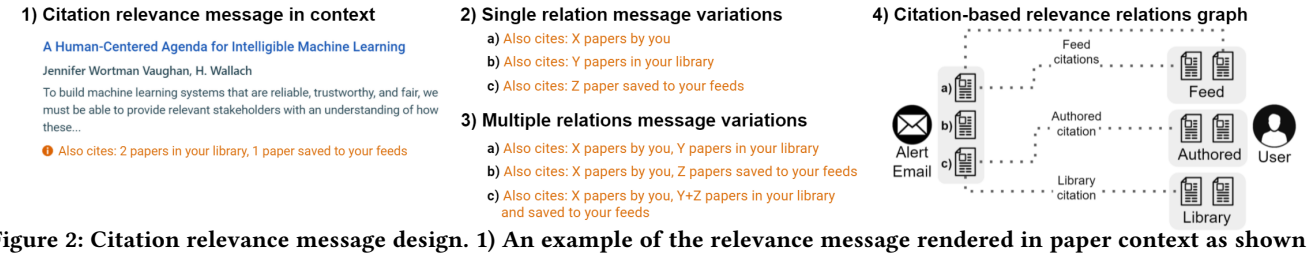


Figure 2: Citation relevance message design. 1) An example of the relevance message rendered in paper context as shown to the user. 2) Examples of messages featuring relevance relations to different user papers: papers a) user authored, b) placed in the user's library, c) added to the user's feed. 3) Examples of messages featuring multiple relations. 4) Graph based depiction of the citation-based relevance messages: an alert paper a) citing two papers from the user's feed, b) with no relevance relation - it will have no message, c) citing both a user authored paper and a paper the user added to their library.

(e.g., library or feed) being cited. In a particular case when a paper cites papers from all the user relevant sources, a shorter variant of the message, which combines the citation counts for library and feed papers, is presented in order to reduce user cognitive load [107] (see Fig. 2.3c).

Design 2: Direct Author-based Relevance Messages. *Goal*: Design 2 conveys potential relevance of an alert paper to the user using their implicit social network of authors (Fig. 3.4). In particular the authors with papers that the user has explicitly expressed interest in are emphasized. The strategy is motivated by several indications from prior work that academics actively search for researchers with similar interests [77], but existing services offer limited support for such discovery [20]. Further motivation for emphasizing author information comes from indications about the importance of author networks in academia [95] and the value of social networks of trusted sources [7, 24] in broader recommendation contexts.

Design: We define a relevant alert paper to be a paper authored by at least one *user relevant author*, defined as having authored one or more papers in the user relevant sources. For this design, we expanded the number of user relevant sources to four by adding papers cited by the user. Early user feedback indicated that this source was meaningful only in Design 2. Our *direct author-based relevance condition* adds one of the relevance messages presented in Figs. 3.2 and 3.3 to any relevant alert paper. If none of the authors on an alert paper are user relevant authors, no message is added. If a paper was authored by more than one user relevant author, only the author with the highest number of papers in the user relevant sources is shown, with ties broken randomly. Alert papers featuring a user relevant author who has authored papers in one or more user relevant sources are featured with one of the message variations presented in Fig. 3.2 or Fig. 3.3, respectively. Each relevance message communicates information about the number of papers the user relevant author has authored in each user relevant source. If a featured author has authored papers in all the user relevant sources, the library and feeds sources are merged to create a shorter variant of the message and reduce user cognitive load [107] (see Fig. 3.3d).

4 STUDY 1 - LARGE-SCALE ONLINE DEPLOYMENT STUDY

4.1 Procedure

We randomly assigned over seven thousand email-alert users (see Section 3.1 for a system overview) to one of the three conditions:

Condition	# of Users	# of Emails out of total featuring at least 1 message (%)
Control	2,248	N/A out of 22,548
Citation	2,474	5,984 out of 23,658 (25%)
Direct Author	2,316	3,895 out of 22,657 (16%)

Table 1: Statistics of the users and emails in our analysis.

Control/No message (status quo experience with the research feed and new paper alerts), Citation (citation-based relevance messages are added), and Direct Author (author-based relevance messages are added). Over a span of about two-months (April 7th – June 13th, 2021), participants received regular alert emails containing new paper recommendations for personally curated research feeds. Table 1 shows descriptive statistics of the dataset collected for analysis. The system logged user engagement (e.g., opening an email or clicking a paper title to view the paper detail page on the search engine). Thus, our measure of user engagement is two-fold, with the click-through rates grouped at the email level as a measure of engagement—i.e., CTR: a binary measure of either 1: the email was opened and at least one included paper recommendation was clicked, or 0: otherwise (includes unopened emails)—and an additional measure of *future* engagement via email open rates. If participants found the emails useful, they would likely open more of them in the future, resulting in higher overall email open rates. In this sense, higher email open rates are indicative of increased future engagement.

4.2 Results

We report results from our studies below. We compare the different messaging *strategies* in our analyses rather than messages alone, i.e., the effect of messages together with the recommendations they augment, by comparing the full email dataset (unless otherwise specified). In such analyses, we use *messages* interchangeably with *messaging strategies*. To denote statistical significance we use the following notations: $\alpha = .05(^{*}), .01(^{**}), .001(^{***}), .0001(^{****})$. Alpha levels were adjusted when appropriate in post-hoc analyses using Bonferroni correction.

4.2.1 Both types of messaging strategies increased CTR, but only direct author-based messaging increased future engagement. We found that both types of relevance messages significantly increased the CTR over Control, with Direct Author ($\mu = 0.058, \bar{\sigma} = 0.2330$, $t(44622.22) = 6.20, p = 5.6 \times 10^{-10}$) and Citation ($\mu = 0.056, \bar{\sigma} = 0.2292$, $t(46065.34) = 5.36, p = 8.3 \times 10^{-8}$) resulting in higher CTRs

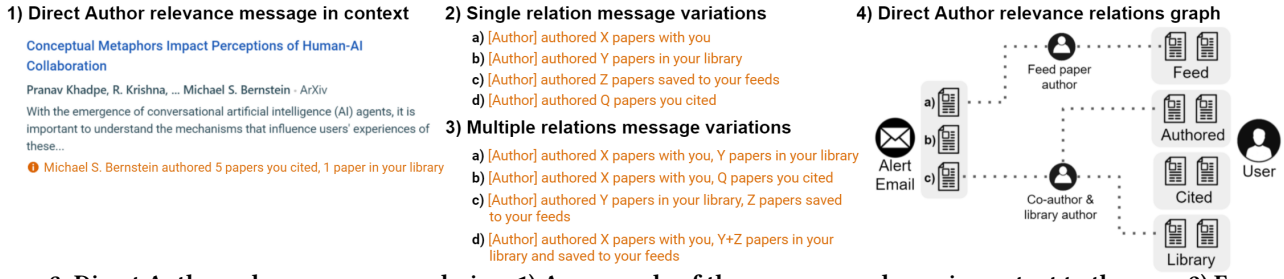
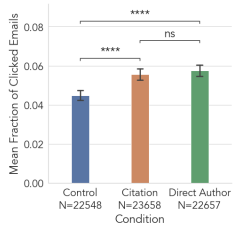
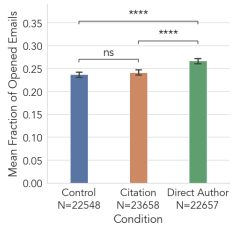


Figure 3: Direct Author relevance message design. 1) An example of the message as shown in context to the user. 2) Examples of messages featuring author relevance relations to different user papers: papers user a) authored, b) added to the library, c) added to their feed, and d) cited. 3) Examples of messages featuring multiple relations. 4) Graph based depiction of the direct author-based messages: an alert paper a) featuring an author of one paper from a user feed, b) with no direct author relevance relation (thus, having no message), and c) featuring an author of both a user-authored paper and a paper in the user's library.

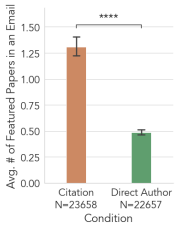


(a) Both direct author ($\mu = 0.058$) and citation ($\mu = 0.056$) messaging resulted in significantly higher CTR than Control ($\mu = 0.045$).

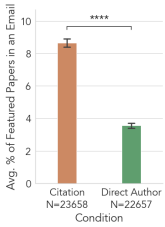


(b) Direct author messaging ($\mu = 0.27$) resulted in the highest email open rates ($\mu = 0.24$ for both Citation and Control).

Figure 4: Analysis of the CTR and email open rates by type.



(a) Direct author messages were significantly rarer ($\mu = 0.5$) than Citation messages ($\mu = 1.3$) in each email.



(b) The rate of messages showed a similar pattern ($\mu = 8\%$ for Citation vs. $\mu = 4\%$ for Direct Author messages).

Figure 5: Analysis of the message prevalence by type.

than Control ($\mu = 0.045, \bar{\sigma} = 0.2068$) (Welch's two-tailed t-test, Fig. 4a). However, the difference was not significant between the Citation and Direct Author conditions ($t(46149.07) = 0.92, p = 0.36$). We validated these results with analyses on potential biases of randomization, such as the average email length (Fig. 13) and message position (Fig. 14) (see Appendix A).

We further examined whether different message types resulted in differences in how often users open the emails. The Direct Author condition showed a significantly higher email open rate than the rest ($\mu = 0.27, \bar{\sigma} = 0.442$ vs. $\mu = 0.24, \bar{\sigma} = 0.428$ in Citation and $\mu = 0.24, \bar{\sigma} = 0.425$ in Control, Fig. 4b, significant at $\alpha = 1.0 \times 10^{-4}$). Because the subjects, headers, and other metadata of emails did not differ between conditions, it is likely that the difference in open rates is attributable to the content of the emails and specifically the type of relevance messages featured in them.

Furthermore, we analyzed the effects before and after the first two-week exposure to messages. The result of the difference-in-differences analysis (Appendix B) shows that users in both Citation and Direct Author groups opened more emails, suggesting habit-forming, but only Direct Author messages significantly boosted the open rates after accounting for the baseline increase in open rates with repeated exposure of alert emails over time (see Fig. 15 in Appendix B).

4.2.2 Importantly, direct author-based messages were significantly rarer. While both citation- and direct author-based messages were effective, their prevalence differed significantly. Specifically, direct author-based messages were much less frequent ($\mu = 0.5$ messages per email, $\bar{\sigma} = 1.85$) than citation-based messages ($\mu = 1.3, \bar{\sigma} = 7.37$, $t(26739.0) = -16.6, p = 0$, Welch's two-tailed t-test) (Fig. 5a) and the pattern remained similar when normalized by the length of emails ($\mu = 0.09, \bar{\sigma} = 0.212$ in Citation vs. $\mu = 0.04, \bar{\sigma} = 0.119$ in Direct Author, $t(37428.4) = 31.9, p = 0$) (Fig. 5b).

Yet the frequency of messages was suggested to be a significant factor on CTR. Locally Weighted Scatterplot Smoothing (LOWESS) suggested a significant inverted U-shaped relationship on CTR by the % of paper recommendations featured in the email (Fig. 16b, Appendix C.). Overall, empirically we found that emails with a greater fraction of treated papers result in dramatically more engagement, up to a point (approximately 25–50% treated) above which engagement falls off. This pattern is conceptually consistent with existing theories of engagement such as the Aristotle's idea of the mean [39], Csikszentmihalyi's optimal difficulty [32], and the relationship between workload and innovative work behavior [82].

4.2.3 Optimizing the frequency of direct author-based messages showed a higher estimated marginal utility than for citation-based messages. In addition to the % of paper recommendations featured with relevance messages in the email (*% Featured*), we also found multiple other covariates suggested to be correlated with user engagement (i.e., CTR) through analysis of descriptive statistics using LOWESS. For example, users with a higher h-index and a claimed profile had more data available in the system that could be used to feature relevance messages on papers—and these users also tended

Receiver ID	Mail ID	Dep. Variable	Predictive Variables			
		CTR	Claimed Profile	Receiver h-index	% Featured	# of Total Papers
1	100	1	1	3	0.38	14
2	101	1	0	11	0.60	20
1	102	0	1	3	0.30	10

Table 2: Sample format of the collected data used for generalized linear mixed-effects modelling. Each row represents a unique user - alert email combination. CTR is the binary dependent variable representing the email-level user click-through outcome (1: whether any paper recommendation included in an opened email was clicked, or 0: no paper recommendation was clicked), *Claimed Profile* shows whether the user has a claimed profile on the search engine, which may indicate an overall high level of engagement. *Receiver h-index* shows the h-index of the user and was normalized by the avg. h-index of users. *% Featured* shows the % of paper recommendations in the email featured with relevance messages. *# of Total Papers* shows the total number of papers included in the email and was also normalized by the avg. email length in the data corpus before the analysis. Users were randomly assigned to either Control, Citation, or Direct Author conditions, and included as random effects in the model.

to show higher baseline engagement (see Fig. 16c in Appendix C). Therefore, we included these covariates as additional predictive variables, and modeled the dependent variable, CTR, using a generalized linear mixed-effects model (GLMM) [68] with the LME4 package in R [10]. GLMMs are often used to analyze (potentially correlated) repeated measures, which in our case corresponds to each user engaging with multiple emails. GLMMs have been used to analyze measurements across many disciplines including medicine, behavioral sciences, and HCI [27, 33, 45, 46].

We developed increasingly sophisticated models for analysis. For example, our first model (Model 1 in Table 8, Appendix. E) simply included the *% Featured* with Direct Author relevant messages and the number of total papers in each email as fixed effects[1]. The result of Model 1 validated the empirical data that showed a curvilinear relationship between CTR and the *% Featured*. Our full model (Model 2) added other empirically significant predictive variables described earlier (i.e., *Claimed Profile* and *Receiver h-index*) as fixed effects along with random effects for users to account for user response level correlation (Table 2 shows the structure of our dataset).

The full regression result showed once again a significant curvilinear relationship between CTR and the *% Featured* with direct author-based messages, even after controlling for other covariates (Table 3). We further estimated the marginal effect of different *% Featured* for each type of message and user segment representing a high (i.e., user with a claimed profile) versus low level (i.e., without a claimed profile) of engagement. The optimal *% Featured* was around 50%, after which the likelihood of CTR dropped off (Fig. 6). The optimal likelihood of click-through was predicted higher for direct author messages (30%) compared to citation messages (24%). The lift from 0-to-optimum *% Featured* was also predicted higher for direct author messages (Δ = +20%) compared to citation messages (Δ = +14%). Taken together, the analysis suggests that for users with less interaction history and fewer featured papers, strategies to increase the coverage of relevance messages to the 40–60% range are a promising avenue to increase engagement, and specifically, direct author-based messages may benefit more from increased coverage compared to citation-based messages. We turn to these strategies in Section 5.

[1]Note that though we only report the results from direct author-based relevance messages here, results from the citation-based relevance messages are similar.

	Coef.	SE	p
(Intercept)	-7.77	0.282	***
% Featured	4.56	0.823	***
(% Featured)^2	-4.95	1.063	***
# of Total Papers	0.02	0.038	0.67
Claimed Profile	2.26	0.349	***
Receiver h-index	9.53	3.476	**
% Featured × Claimed Profile	-3.42	1.283	**
(% Featured)^2 × Claimed Profile	4.30	1.499	**
% Featured × Receiver h-index	-2.25	5.559	0.69
(% Featured)^2 × Receiver h-index	2.00	7.829	0.80

Table 3: Regression analysis with our full model (Model 2) predicted a significant curvilinear effect or % Featured on CTR in the presence of other covariates (i.e., predictive variables) e.g., whether user has claimed a profile, h-index, and their interactions. ***: $p < 0.001$, **: $p < 0.01$.

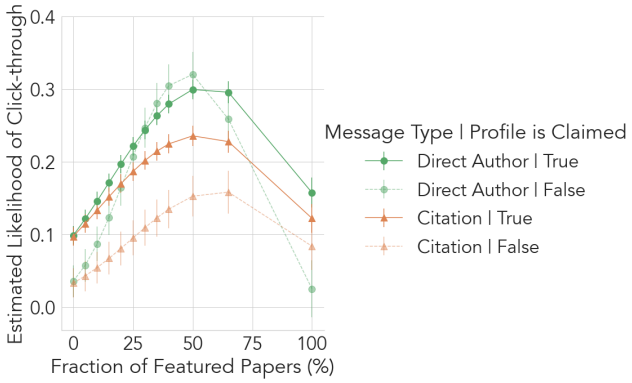
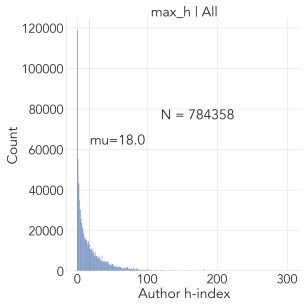
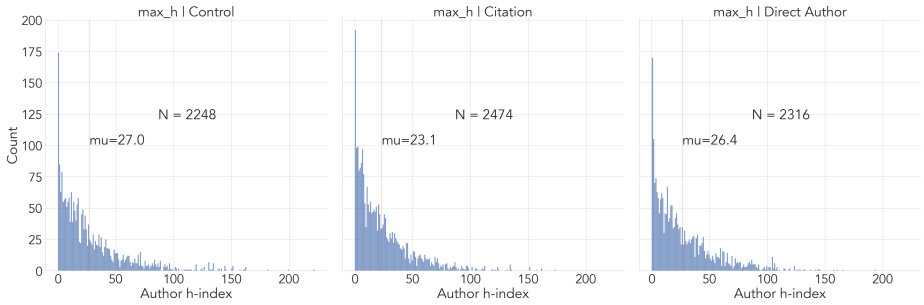


Figure 6: Direct author messages showed a significantly higher ceiling of CTR than citation messages. The overall estimated (marginal) likelihood of CTR peaked when approximately 50% of papers were featured with messages, for all groups. However, increasing the % Featured had a more pronounced effect on CTR for direct author messages (over 20-absolute-percentage-points for both profile-claimed and unclaimed users) than citation messages (15 points or less).



(a) The 'background' distribution of maximum author h-index from all paper recommendations sent out in alert emails. The average maximum author h-index was $\mu = 18.0, \bar{\sigma} = 23.29$ (median=9.0).



(b) The distribution of maximum author h-index of clicked papers in each condition. The average h-index increased to $\mu = 27.0, \bar{\sigma} = 29.33$ in Control (median=17.0); $\mu = 23.1, \bar{\sigma} = 24.24$ in Citation (median=16.0); and $\mu = 26.4, \bar{\sigma} = 26.67$ in Direct Author (median=18.0). The increase of the average h-index over the background distribution was significant, suggesting that users considered high status authors as a signal for deciding whether to click on a paper by default. At the same time, citation messages reduced the h-index relative to control, suggesting its effect of guiding user attention to lesser known authors. Direct author messages and control did not differ significantly.

Figure 7: Analysis shows that users clicked on papers featuring high status authors, but does not provide any evidence that messages further centered user attention to them.

4.2.4 Examining the messages' system-wide impact on fairness. Though effective, featuring paper recommendations with relevance messages may produce unanticipated negative externalities such as boosting only the visibility of papers by authors who are already often featured in the recommender, or being accessible only by users with high data availability, thus further selectively enhancing their engagement with the literature. Therefore, we examined the fairness of our message designs in two respects: a) their effect on work visibility and b) their coverage for different user groups.

Fairness of visibility. While the phenomenon of "rich-get-richer" has been widely studied [17, 28, 31, 35, 62, 66, 76, 78, 94, 116, 118], to our knowledge no study has examined the effect of interventions designed to shift user attention to certain types of papers on visibility via a randomized study on a large-scale, real-world, deployed recommender system.

One way to examine relevance messages' effect on visibility is to take an outcome oriented approach, by measuring changes on the papers users clicked before and after intervention. In our alert emails when a user clicks on a paper link, it brings the user to an interactive paper detail page, which includes full abstract, the link to a full-text file, any figures, and the author information among others. Therefore the click interaction serves as a strong indicator of the user's exposure to the paper's content. Thus we operationalized the visibility of each paper as a binary measure of whether it was clicked; when a paper is clicked for the first time, the paper webpage is made visible to the user. Using this measure of visibility, we investigated whether its distribution over the academic status of the authors of papers changed systematically with the introduction of relevance messages.

To operationalize the academic status of each author, we used the h-index measure. The index, since its introduction by Hirsch in 2005 [52], has been popularized as a metric of academic success. Though limitations of the index exist [91, 104], it is seen as a robust measure [2, 48, 53, 96], and informs high stakes decisions such as hiring, promotion, and funding [1, 50, 73]. It is also widely featured in many scholarly search engines and citation databases. We assigned an h-index to the paper recommendation by taking the max h-index of authors on the paper. This is plausible because the highest status author of the paper may appear salient and easily recognizable to the user at first glance. This produced our baseline h-index distribution of authors (Fig. 7(a)). Next, we similarly computed the h-index of each recommendation but using only the clicked paper recommendations from each condition (Fig. 7(b)). Finally, to examine whether relevance messages led to users clicking papers with high status authors more often, we tested whether the average h-index of clicked papers significantly differed from that of the baseline.

The result shows that while the h-index of clicked papers was significantly higher than the h-index of all papers, this increase was no worse than the baseline, suggesting that users naturally incorporate author identities and status represented as h-indices when deciding whether to click on a paper recommendation (shown in the jump from all authors' h-index to clicked authors' h-index $\mu = 18.0 \rightarrow \mu = 27.0$; Fig. 7(a) and Fig. 7(b) left, $t(2255.13) = -14.55, p = 0$, Welch's two-tailed t-test), and featuring relevance messages did not exacerbate this effect. There was evidence that citation messages guided user attention towards lower status authors' papers more than Control ($t(4371.48) = 5.06, p = 4.36 \times 10^{-7}$). For robustness against any spurious effects from choosing the maximum h-index, we repeated the analysis using the average h-index of authors and found that the patterns remained the same (see Appendix F). Taken together, we conclude that augmenting paper recommendations did not adversely impact fairness of visibility with respect to h-index, and may have shifted towards a fairer distribution of user attention when citation-based messages were used.

Fairness of coverage. While Direct Author messages were rarer than citation messages overall (Fig. 5b), was there any systematic difference in the coverage of user groups with varying academic status? We further divided the users into three groups of h-index (Low, High, and Unknown) using the median (3) h-index from the

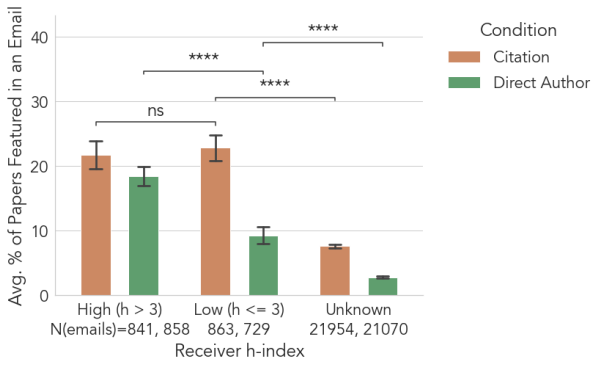


Figure 8: While the frequency of direct author messages decreased significantly, from 18% to 9%, as users' h-index decreased from High to Low, it remained similar for Citation.

authors whose h-indices were available (h-index was known for 115/2,474 users in the Citation condition and 107/2,316 users in the Direct Author condition). We found that groups received significantly fewer Direct Author messages as the h-index got smaller: μ = 18% in the high h-index group, μ = 9% in the low h-index group, and μ = 3% for the group with unknown h-index (the pairwise decreases were significant, $p < .0001$, Fig. 8). This may not be surprising given a user's high h-index may be correlated with an overall higher level of engagement with the search engine that results in richer interaction data and also higher connectedness within the academic network (e.g., a bigger collaborators' network), useful for message generation. However, the frequency of citation-based relevance messages did not significantly differ between the Low and High index groups, suggesting a relative abundance of citation-based relations for users with fewer publications (e.g., our citation message generator could find relevance through lower index users' curated feeds alone); however, the same was not true for author-based messages. This result suggests that increasing coverage of messages may be particularly helpful for improving the fairness of author-based relevance messages, and benefit users who have fewer connections in the academic social network.

5 DESIGN OF INDIRECT AUTHOR-BASED RELEVANCE MESSAGES

5.1 Key Motivations and Related Work

The key motivating results of Study 1 showed that 1) Direct author-based messages were an effective mechanism for increasing future engagement (the overall email open rates were significantly higher in the direct author-based messages condition compared to Control and the citation-based messages condition (Fig. 4b), and the effect was clear after controlling for the baseline effect from habit-forming (the difference-in-differences analysis, Appendix B); 2) This effectiveness was achieved despite the direct author-based messages being significantly rarer than citation-based messages (Section 4.2.2). This relationship may generalize, possibly enabling higher CTR with fewer messages/email. For example, in our GLMM analysis, when the likelihood estimate was held as constant, the estimated % Featured is lower for direct author-based messages than citation-based messages (Fig. 6; 3) In addition, direct author-based messages showed a higher ceiling of the (marginal) likelihood of click-through compared to citation-based messages (Fig. 6). Thus, in order to achieve the author-based relevance messages' full potential, we designed a novel mechanism for expanding the coverage by incorporating indirect relations mediated by trusted intermediate authors between a user and the recommended papers.

Prior work in social network-centric relevance explanations further provides support for expanding author messages via indirect relations. For example, Sharma and Cosley investigated the value of featuring *direct* friendship-based relevance messages on persuasiveness and informativeness in the music recommendation domain [102]. They found relevance through good friends mattered more than random friends, and as such showing the (good) friends' names in messages led to higher informativeness than representing them as aggregate friend popularity (e.g., '3 of your friends liked this album'). Additionally, when messages featured the overall popularity (e.g., '12,211 of Facebook users like this'), the popularity mattered only if the users identified with the crowd. Extrapolating these findings to the design of indirect author messages, we expect to find similarities such as the importance of relation strength and specifying which author and why they were featured in the messages, but also new challenges as to deciding who should be the intermediate authors and how to identify them. We expect the citation network of existing publications and their authors to contain topical relevance and trust relations targeted specifically to our task domain, and this informs our algorithm design for *inferring* an academic 'social' network from it. In this regard, our approach differs from [101, 102] which used existing social networks (e.g., Facebook) and [24] which relied on Twitter. Additionally, the SONAR system combined social information for members within an organization from co-authorships of organizational Wiki articles and user interaction traces such as bookmarking the same pages and usage of same tags [41, 42]. However, this work also differs from our work due to its dependence on an explicit social network and its limited scope to direct relations. In the subsequent sections, we describe the design of indirect author-based messages and its generation algorithm.

5.2 Design Goals

We grounded the design of indirect author-based messages with the following goals:

G1. *Support relevant information needs.* Scientists often turn to trusted sources as a way to curate relevant papers. Yet, this process is often tedious and existing tools provide piecemeal support at best. Therefore, messages should provide benefits similar to receiving personalized bibliography from an expert source.

G2. Support serendipitous discovery, which is an oft-mentioned benefit [20, 77, 95]. To achieve this, the author featured in messages should be likely subjects of serendipitous discovery.

G3. Support design continuity. New message designs should leverage the robust designs used in Study 1, to minimize unanticipated negative consequences from design changes, and to prevent harming the user experience of the platform.

5.3 Message Implementation and Generation

We adopted the overall textual and visual design from earlier message designs, with an additional line of text to accommodate the

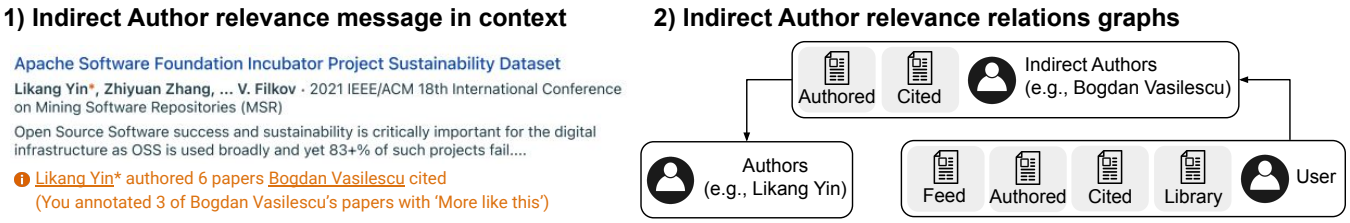


Figure 9: Indirect author relevance message design. 1) A sample message rendered in the recommendation context as shown to the user. The message features two lines of text. The first line features the relation between the author of the paper (Likang Yin, underlined) and the indirect author (Bogdan Vasilescu). In the second line, the user's relation to the indirect author is described. The author names are clikable and linked to profile pages on the search engine. 2) Indirect author candidates are first identified from a set of user papers and interaction data then filtered based on their connection to the authors of the paper.

relational information (Fig. 9, left). We placed an asterisk next to the author's name to differentiate it from the indirect author's. We also added underlines and click interaction to the names to support a discovery experience. The second-line text was added to describe the user's relation to the indirect author for clarity.

To generate the messages, candidate intermediate (indirect) authors were identified from the user's publication and interaction data. Then, each candidate's publication data was used to extract [author]-[indirect author]-[user] triplets for each author of a recommended paper. Finally the first ranked triplet was fed into the message template (Fig. 9, left).

To rank the triplets, let us first define the strength of the [author]-[indirect author] relation as 'Relevance' and the strength of the [indirect author]-[user] relation as 'Influence'. For a candidate (author, indirect author, user) $= (i, j, u)$ triplet, we computed these strengths as

$$\text{Relevance}_{i,j} := a \times \log(\text{co-authored}_{i,j} + 1) + b \times \log(\text{cited}_{i,j} + 1)$$
$$\text{Influence}_{j,u} := \log(\text{engaged}_{j,u}) \times (j\text{'s h-index}),$$

where $\text{co-authored}_{i,j}$ is the number of papers that i and j co-authored, $\text{cited}_{i,j}$ is the number of i's papers j cited, $\text{engaged}_{j,u}$ is the number of j's papers u engaged with using one of the following actions: co-authoring, citing, saving, and annotating with 'more like this' on a personal research feed. The intuition here is that the higher the number of actions i took on j's papers (e.g., citing, saving), the stronger the tie strength. The constants a and b in Relevance control the relative strength of relations, and we set $a = 2, b = 1$ to weight co-authorship twice as strong as citation because the former is believed to indicate stronger relevance. In addition, we take a logarithm of the count of papers to account for the diminishing signal strength (e.g., an extra citation means much less when it is already cited 20 times). For Influence, we multiply the logarithm of engagement counts with the candidate indirect author's h-index to prioritize individuals with higher academic status because they are more likely to be known and trusted by the user. Finally, our ranking objective for a given user u is: $\text{argmax}_{i,j}\left(\text{Relevance}_{i,j} \times \text{Influence}_{j,u}\right)$. This multiplicative objective was designed to prioritize triplets with coherent (rather than lopsided) tie strengths (e.g., a high score on only one of Influence or Relevance but low score on the other may result in an overall irrelevant relation to the user).

6 STUDY 2 - CONTROLLED LAB STUDY

We performed a formal usability study to gain insights into the following questions: How do different types of relevance messages aid scientists' ability to review the recommended research papers in an email alert context? How do scientists make sense of the relevance information conveyed in the messages? What are the challenges and design implications for future author-based relevance messages?

Using a within-participants design, we compared the Indirect Author-based relevance messages to Citation- and Direct Author-based relevance messages and Control (no message). The quantitative and qualitative results were in favor of the relevance messages, and the value of different types of messages seemed complementary. Through open coding of interview transcripts, we discovered different themes of the benefits complementing the results from Study 1. We also uncovered challenges from which we synthesized implications for design.

6.1 Study Design

Participants. 14 scientists were recruited via university and company mailing lists. 1 was an assistant professor, 5 were postdoctoral researchers, 1 was a professional researcher, 1 was a Master's student, and 6 were doctoral students. 6 of the 14 participants identified their discipline as human-computer interaction. Participants were compensated at a $30/hour (USD) rate. The study sessions were between 1-hour- and 1.5-hours-long and held remotely on a video conferencing platform. Participants opened an individualized Google Doc prepared by the interviewer and were asked to share their screen. After obtaining consent from each participant, the interviewer proceeded to record the session.

Stimulus recommendation emails. Personalized paper recommendations were generated for participants using their publication, library, and research feed data. Each participant's recommendations were randomly subdivided into 4 sets (A, B, C, D) of equal length, ranging from 12 to 30 papers per set. Then, relevance messages were generated and added to the corresponding paper recommendations. Following the results from Study 1 (Section 4.2.3), we allowed up to 50% of the papers in each email to be featured, which was the approximate optimal fraction of papers to be featured. The generated emails looked exactly the same as in Study 1, except for the messages and the headers, which were anonymized so as to not

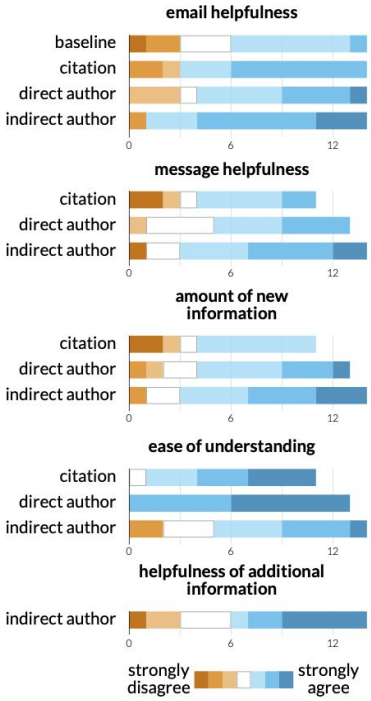


Figure 10: Subjective responses to test questions. Overall, participants responded favorably on the general helpfulness of messages, with indirect author messages being harder to understand. The benefits of different message types seemed complementary (see text).

include any identifying information when recording with screen share. In Control (A), paper recommendations were shown without messages. In the citation- (B), direct author-based (C), and indirect author-based (D) conditions, messages of the corresponding type were added underneath each paper recommendation, when applicable.

Tasks and Assignment. Each session ran as follows: 1) Greeting and obtaining consent for recording; 2) Background questions around how the participant typically obtains research papers to read. If the participant was a user of an alert system, the interviewer also asked questions about their experience with typical new paper alerts; 3) Complete four timed tasks (6 minutes each), each of which was followed by a task-specific questionnaire. Using four 4×4 Latin Square blocks, we assigned each participant to one of the randomly drawn rows. The number of recruited participants (14) resulted in a near but not fully factorial design (two presentation orders had one more participant each).

Measures. For each task, we measured the following (all but the last two measures are on a 7-point Likert scale between 1: Strongly disagree and 7: Strongly agree). In 4 condition blocks (3 in B), participants skipped measures related to relevance messages, as no messages were shown to them. "*Email helpfulness*" is the participant's self-assessed agreement with the following statement: "I found the email helpful."; "*Message helpfulness*" (in B, C, D) indicates the participant's self-assessed agreement with the following statement: "I found the orange text underneath paper recommendations helpful."; "*Novel information*" (in B, C, D) indicates the participant's

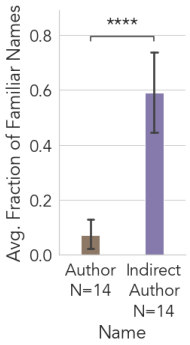


Figure 11: On average, participants recognized indirect author names about 60% of the time, and author names 7% of the time (8 times less than indirect authors).

self-assessed agreement with the following statement: "I found the orange text underneath paper recommendations to contain interesting new information."; "*Ease*" (in B, C, D) indicates the participant's self-assessed agreement with the following statement: "It was easy to understand what the orange text underneath paper recommendations was trying to tell me."; "*Additional information helpfulness*" (in D) indicates the participant's self-assessed agreement with the following statement: "I found the second line of the orange text helpful."; "*% Featured*" (in B, C, D) is the percentage of paper recommendations featured with relevance messages; "*Number of familiar author/indirect author names*" (in D) is the number of author/indirect author names included in indirect author-based relevance messages the user found familiar.

We distinguish between the overall email helpfulness question and the message type-specific questions for two reasons: one, the overall helpfulness measure allows us to compare the usefulness of message-augmented emails with that of the baseline; two, message type-specific measures allow us to see whether participants saw complementary value from the different types of messages, in which case we expect to see no significant difference between the message conditions.

Analysis. For each of the quantitative measures, we ran post-hoc analyses with Welch's two-tailed t-test. For qualitative analysis, the interviews were recorded, transcribed, and coded in four iterations following an open coding approach. The goal of this process was to identify common themes that captured rich qualitative insights grounded in data [19]. Using the transcripts, audio, and video recordings of the interviews, two coders first independently performed open coding of the initial themes. Subsequently, they had in-depth discussions over multiple sessions to merge similar codes and form higher level themes, as well as a larger group discussion with four of the authors on the themes. Finally, these consolidated codes and themes were applied to the transcripts to arrive at the qualitative findings of Study 2. We focused our analysis on uncovering the benefits and challenges of relevance messages, and the results converged on themes presented in Table 4. The detailed descriptions of these themes are provided in Section 6.3.

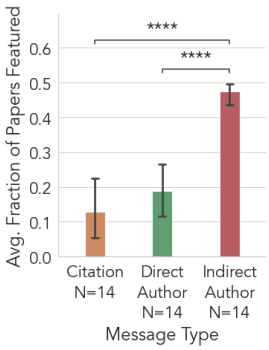


Figure 12: Indirect author messages significantly increased the coverage to 47%, compared to direct author (13%) and citation messages (19%).

6.2 Quantitative Results

The result validated the feasibility of indirect author messages. The % of papers featured with indirect author messages was close to the estimated optimal, at 47% ($\bar{\sigma} = 0.062$, Fig. 12). In comparison, citation- ($\mu = 0.13$) and direct author-based messages ($\mu = 0.19$) were significantly less frequent (at $p < 0.00001$, Welch's two-tailed t-test, Fig. 11). In addition, indirect authors were significantly more recognizable than the authors ($\mu = 0.59, \bar{\sigma} = 0.290$ vs. ($\mu = 0.07, \bar{\sigma} = 0.111$) ($t(16.74) = -6.25, p = 9.4 \times 10^{-6}$, Fig. 11).

Overall, participants perceived indirect author messages favorably. The result of a post-hoc t-test showed that emails augmented with indirect author messages ($\mu = 5.7, \bar{\sigma} = 1.27$) were perceived as significantly more helpful than the baseline ($\mu = 4.1, \bar{\sigma} = 1.46$) ($t(25.49) = -3.04, p = 0.005$), but for citation- ($\mu = 4.8, \bar{\sigma} = 1.66$) and direct author-based messages ($\mu = 4.9, \bar{\sigma} = 1.32$) the difference was not significant (Fig. 10, top). At the same time, participants found indirect author messages the most difficult to understand (Fig. 10, second from the bottom, avg. response was $\mu = 4.9, \bar{\sigma} = 1.21$ for indirect author messages and $\mu = 6.5, \bar{\sigma} = 0.52$ for direct author messages, $t(17.92) = 4.56, p = 0.0002$). Six participants (42%) also responded either in a negative (21%) or neutral (21%) manner to the additional text designed to help with understanding (Fig. 10, bottom). Thus, the message usefulness may have been reduced by the difficulty of understanding. Furthermore, we hypothesized that each message type may bring complementary value to the user when they engage with paper recommendations. For both *Message helpfulness* ($p > 0.14$) and *Novel information* ($p > 0.18$) questions, we found no evidence of significance differences in user responses. P14 commented that "each of these messages feels like it has a different utility than the other two." We further investigate this hypothesis with qualitative analysis of the data below.

6.3 Qualitative Results

Now we present qualitative insights into participants' experience with messages. Using axial coding following our open coding, we organized these insights into six types of benefits and four types of challenges (Table 4) and provide detailed descriptions below. We use a fraction (e.g., 3/14 participants) to represent multiple participants expressing similar thoughts. In participant quotes, we used notations **[A]** and **[IA]** to denote **Author** and **Indirect Author** names featured in the message, respectively.

Types of Benefits. *B1: Noticing and being drawn into relevance messages.* Participants in the interviews commented that the overall design of relevance messages stood out to draw their attention to the recommended papers featured with messages: "It seems to be all the orange text sticks out... So that's why I would scroll through the email and really look out for those" (P9). Participants mentioned that they 'liked' the visual distinction of the message design (3/14 participants), they 'missed' the messages when there was none after moving to the next task in the experiment (P2), they used them as anchor points to structure their experience and attention around relevant items (4/14 participants), and that they were drawn to the personalized connections described in the messages (P1).

B2: Developing awareness of other scientists. Participants were curious and wanted to see work by other scientists they cared about to gain a general understanding of what their most recent research areas were: "Oh, this is what [A] is doing... So now I can be like, 'Hey, I saw the UMAP paper you published' when I meet him in two weeks. So a conversation point" (P4). Author-based relevance messages were appreciated because they surfaced author connections that might otherwise have been unnoticed or easily missed: "This paper is not something interesting to me but the personal connection is. I'm also surprised that my PhD advisor highly cited papers by him. So maybe I should have a look at it more closely?" (P3). In such cases, even divergent topical areas did not seem to deter participants' curiosity: "Not directly relevant to my research but I would probably look at this because I know the author" (P9); "I don't cite [A] anymore. It's more like, just curious what he's up to" (P4). Messages were valuable as a vehicle for getting updated information on how the research direction of an author represented in them was evolving: "I know [A]. I would like to work with her. So maybe I'm going to read this paper because just to know what she's up to" (P6); "interesting that [IA] is working on voice interaction too" (P7).

B3: Discovery of Community-Related Insights. Messages also helped participants recognize community-related connections that go beyond individual author connections in two ways: first, messages helped them make sense of their connections with broader research communities; second, messages helped them understand the impact of their work.

B3-a. Connections in the Intellectual Community. Participants found author-based relevance messages helpful for understanding how they might be connected to larger research communities: "Interesting... because at least in my community it feels a bit like a small world. So it's interesting to know who's cited by who" (P3); "Okay so [IA] and [A] have a deep connection in the research space and I have a weak connection with this space with them" (P6); "It's like, 'Hey, you cited this person before', although I'm not like always trying to cite the same people, but when I did my knowledge tracing work in my first few years, [A] did come up a lot. So being able to see that is cool" (P4). In addition, the count of papers was helpful for quantifying known connections: "[A] authored four papers I cited. Wow, yeah, I knew that I cited him but I didn't have, like, any kind of quantification of it" (P4).

Benefits and Challenges of Relevance Messages Augmenting Scientific Recommendations		
Benefit	B1: Drawing Attention	Design and personalization of messages were sufficient to draw user attention
Benefit	B2: Social	Developing awareness of other scientists
Benefit	B3: Discovery of Community-Related Insights	
	B3-a: Connections in the Intellectual Community	Seeing connections in the community
	B3-b: Scientist's Intellectual Impact	Understanding & tracing their impact
Benefit	B4: Mobilizing Mental Models	
	B4-a: Accessing Readily Available Mental Models	Understanding the recommended research
	B4-b: Applying Mental Models for Transfer	Making sense of connections to new authors
Benefit	B5: Serendipitous Discovery	Discovering new and interesting authors
Benefit	B6: Judging Potential Value	Judging the usefulness and interestingness of recommended papers
Challenge	C1: Interpretability	Sources of misinterpretation: linguistic and semantic mismatches, incongruence between recommendation and relevance messages
Challenge	C2: Context vs. Efficiency	Tension between wanting to see more contextual information and efficiency of reading
Challenge	C3: Scientist's Evolving Identity	Challenges from how scientists' interests change over time and move away from old topics
Challenge	C4: Trust	Challenges due to errors in relevance quantification and increased user sensitivity

Table 4: Different types of benefits and challenges of relevance messages augmenting paper recommendations.

B3-b. Scientist's Intellectual Impact. A specific category of connections to research communities is impact, for example when a scientist's work is cited by others in the community to develop the research area further. Participants found citation-based messages useful for understanding their impact: "When it says 'cites one paper by you', I got excited. It also helped me contextualize this work better" (P7); "It (the recommended paper) cites me... this is something that I cannot immediately figure out (without this message) unless I opened up the paper and, like, control-F'ed my name" (P6); "The messages that talk about how it cites my papers get me really curious. And I'll almost always go look at it, to figure out how my work is being discussed and how my work influenced their work. And maybe think about future work." (P14).

B4. Mobilizing mental models of scientists' work and expertise areas. Participants found author-based relevance messages especially useful when they had prior understanding of the featured authors' work. The mental model of scientists was broadly represented as a combination of their research topics, frequently used epistemological approaches, and their seniority. Using mental models participants could better make sense of the recommended paper's contribution, its broader intellectual context, and inform their filtering decisions as described in the following sections.

B4-a. Readily available mental models of topical areas and understanding broader research context. To many participants author names could be readily mapped as 'specific topical areas' (P9) or 'general research directions' (P14) which can be useful for filtering. Though less frequent than topical indexing, author names also signalled the quality of work: "Papers from [A] are always pretty good, so I probably will read it." (P14). Mental models of authors extended beyond topical associations or quality signals and sometimes even helped participants understand *a priori* in what concrete context the user needs present in recommended papers might arise. P6 described this phenomenon in the following quote: "I don't do ML work, but I can understand how there might be pain points in team communications around the ML model quality... so the [A]'s work I'm familiar with is in documentation and programmer support tools. So I can imagine how that goes for teams with multiple stakeholders, not all of them are technical, especially in places like [large technology corporation] that I believe [A] is at."

B4-b. Mental models for transfer. For indirect author-based messages, participants could transfer their readily available mental models of the known indirect author to make an educated guess about the unfamiliar author. This was perceived as useful for mapping out 'how ideas diffuse' (P2), 'who's building off of the old but important work in the field' (P2, P6) or 'working in an interesting intersection of fields' (P4, P9), which is currently 'a nebulous and difficult task' (P2). Furthermore the unfamiliar author's work may be understood through the mental model of the known author, for example as an indication of 'an important link that is missed' (P12). P4 commented: "[IA] does some cognitive science stuff but he's not a huge cogsci guy... so maybe [A] does similar work to [IA] but like a combination of [IA]'s work with a more of a cogsci spin to it? Which is cool." (P4). Similarly P12 said: "Okay this paper maybe is from a machine learning community I don't follow. But apparently [IA] cited a bunch of [A]'s papers, so maybe I'm missing something." (P12).

B5. Serendipitous author discovery. Author-based relevance messages also led participants to pay attention to author names they otherwise may not have noticed, and serendipitously discover new

and interesting authors 'outside the radar' (P12). P10 said: "A really interesting concept, but I don't recognize this author, let me check his publications... (clicks on the author link to check his profile) Okay wow, this is like a gold mine, I can easily spend an hour or two reading his papers." Discovery of authors also happened in a more structured manner by transferring mental models from the known indirect author to the unfamiliar author. Participants found authors that they had not known before, but felt like they *should* know, given the significant connection through the indirect author. Participants described that they could 'picture where the indirect author is citing the unknown author' (P6), and that a certain number of citations from the indirect author represented 'a good body of work by the unknown author' (P4) and a strong signal of how their research interests are aligned, which indicated the potential value of discovering the new author.

B6. Judging potential value of a paper. Indirect author-based messages were useful for vetting the value of the recommended papers. Participants described how knowing the scientist that they give credit to cited the author of the recommended paper multiple times gave them confidence that the recommended paper would also be useful to them. P14 described that this form of relevance implied a high chance of utility that he, too, could use the recommended paper to support claims in his papers, given how often his advisor (who was the featured indirect author in the message) cited the author. For this reason, it mattered whether the indirect author was someone whose work the scientist knows and gives credit to ("the fact that this mathematician that I give credit to is citing this paper, then that gives credit to the paper." – P3), otherwise the relevance may be ignored. However, one participant commented that he would be interested in seeing a highly selective group of influential authors ("5–10 most cited researchers" – P8) in each subfield to be featured in the messages, regardless of personal connections. Sometimes indirect authors were useful for negative filtering, too e.g., "[A] is in like a CS education-y, a different field than me... I care about what she's doing, but it's a less cool factor for me" (P4); "[A] works with computationally very heavy mechanisms. Not what I'm interested" (P10).

These results uncover qualitative insights that complement the effectiveness of messages in Study 1. In addition, they also show how indirect author-based relevance messages complemented the other two types of messages. Next, we describe the challenges involved with current messages.

Types of Challenges. *C1. Interpretability of indirect author-based messages.* Generating relevance messages in a post-hoc manner could lead to incongruent information. While some participants regarded messages as something that 'doesn't hurt to have' (P4) or 'a weak signal' (P3) even when they were not found to contain particularly useful information, others indicated confusion when encountering a relevance message for a paper that did not seem to relate much to their topical interest ("It gives me more confidence about the recommendation that there's a high chance that I can cite this paper too, but I don't know how it's related to me" – P14). Compared to citation- and direct author-based messages, which featured directly relevant information to users, indirect author-based messages were associated with 'a steep learning curve' (P6) and subject to frequent misinterpretation and re-reading, though participants could get used to them after seeing a few messages of the same type ("The first time was confusing but getting the hang of it now" – P2). This perceived difficulty of interpretation was consistent with participants' subjective ratings on the *Ease* question, in which indirect author-based messages scored significantly lower ($\mu = 4.9, \bar{\sigma} = 1.21$) than direct author-based messages ($\mu = 5.9, \bar{\sigma} = 1.04$) (two-tailed Welch's t-test, $t(22.74) = 2.18, p = 0.04$).

Part of the difficulty was anticipated by our research team given the nature of the message that involved second-degree relations between authors, and was proactively mitigated by the second line text that described the scientists' relation to indirect authors. This was perceived as helpful by 8 out of 14 participants (e.g., "I can tell right away that it's my personal connection" – P1, Fig. 10, bottom), but only when they had some mental model of the indirect author ("I find the text marginally useful when I don't recognize the name" – P1). Participants also wished to see more contextual information that could remind them of 'forgotten connections' (P2) or their own previous interaction data such as which paper they found interesting or had saved in the library, which led to the relevance messages being surfaced.

C2. Tension between more context and efficiency. Unlike names found on a social network where users have real-world relations with each other, names extracted from the *implicit* 'social' network of authors may not be grounded in real-world relations and thus necessitate additional contextual information that helps users understand who the authors are and how they might be related. Participants frequently mentioned wanting to see specific contextual information helpful for making sense of the relevance surfaced from the messages. P7 said: "I think it could definitely use more detail. Because I don't know what I cited, from this context. I know she cited our paper. So now I am more interested, like, oh, what did she say?". For P1, providing author names in the message alone was not nearly enough, given how he needed contextual information even for contacts with personal ties. P2 noted that while most names were unmemorable, their papers might be, and contextualizing and reminding users of their previous behaviors could help ("Is there a way to tell why I thought an author interesting before?" – P2). Other participants also wanted to see fine-grained citation context such as its salience ("it matters more if it was more of a key citation" – P11) or the section information.

Ultimately participants thought that additional context could fulfill their filtering needs, as P6 put it: "I know enough about [IA] to say that he has done some research I'm very interested in and then some that I'm less interested in... I guess he kind of has two camps of research, or two or more at least, and so I'm curious like, this person has authored six papers that [IA] cited... *what camp did this (paper) fall into*, or like what papers is [IA] citing? Are they gender papers or otherwise, cuz that *changes a lot how much I care.*" When participants had a deeper understanding of the expertise areas of an author, they sometimes wanted to filter based on a subset of areas they cared about the most, as relations based on less central areas of interest were less useful or irrelevant. Indirect authors may have done 'a lot of great, but very diverse work' (P7) which may not be interesting to the user.

Taken together, these comments suggest that there is a rich subspace of citation and author information that can be used to further

contextualize the relevance messages, and that it may be most effective when aligned with scientists' task-specific filtering needs. However, while additional context may help scientists reason about the relevance between authors and themselves, it may also increase the chance of information overload [124], creating a tension between desiring more information and also wanting to quickly scan the recommendations in the email. Several participants, including P2, aptly described this potential tension: "I usually want to scan as quickly as possible the titles... I feel like having to parse and switch back and forth between different kinds of relations to indirect authors will tire me."

C3. Evolving research topics and scientist's identities. Another important identified issue involves how scientists' interests change over time, and that they may move away from old topics they themselves previously published in or move back to them with renewed interest later on. Depending on the changes, relevance extracted from citation graphs may become stale over time, and detect a 'pioneering but universally cited' (P11) body of work as highly relevant, or relevance to 'old research interests' (3/14 participants). At the same time, participants also expressed the need to see recommendations related to their 'past selves as scientists' (P14) for specific use cases like accepting review requests on those topics. Taken together, there may be high activation research areas at any time for scientists which constitute their current identities as scientists, with several 'past identities' that consist of somewhat dormant, but occasionally reactivated topical areas. Accurately detecting such identities as scientists may be difficult or even impossible, yet an important aspect for systems aimed at surfacing current relevance to the users nonetheless.

C4. Trust. Participants were more sensitive to potential errors in relevance between authors than in relevance based on citation. Because an important ingredient of effective relevance messages was participants' prior mental models of other scientists, for author-based messages this naturally invited conceptualization of relevance based on people and their intellectual legacy. For example, a high number of citations of an author represented 'a good body of work' for someone (P4) or a potential 'advisor-advisee relationship' (P6), while a low number of citations could have meant 'an up-and-coming' or a 'junior' scientist in the field (P5).

This also meant higher sensitivity to how the relevance between authors was represented in messages, which could erode trust in the system when they suspected the relevance was not accurately quantified. P5 noted that using an exact quantification of how many papers authors have written or cited could be a 'risky statement' because they were falsifiable and the margin of errors was small, especially for scientists that he knew well. Yet, systems that use citation graphs are subject to inevitable sources of errors, given the challenges from the scale and the speed of changes (e.g., papers may be published at different platforms such as ArXiv.org or conference proceedings at different times). One of the participants also noted a case where the relevance message was featuring a very different number of citations from what he expected, which he suspected was due in part to the author's deadname being incorrectly updated. As such, a challenge for systems that aim at inferring implicit social networks of authors from citation graphs is recognizing the increased sensitivity to the accuracy of author-based relevance and sufficiently fail-proofing or communicating the uncertainty of data associated with the representations of relevance to prevent erosion of trust.

7 IMPLICATIONS FOR DESIGN

We further propose three main design areas for future relevance messages that combine citation graphs and implicit social networks of authors to support scientists' sensemaking, filtering, and discovery (overview in Table 5).

Goal-Centric Interpretability. Our participants used relevance messages for a host of diverse yet interrelated benefits. They saw authors featured in the messages as topical representations of their mental models of other scientists, and used them to contextualize the recommended paper and the connections between authors. Messages also helped the participants discover new and interesting authors that they otherwise might not have paid enough attention to notice. These specific use cases suggest that future strategies of relevance messaging need to be tailored to different goals scientists have in mind. In addition, scientists frequently engaged in multiple use cases while reviewing a single email suggesting that support for fluid switching between different goals while interacting with the content in a single email, or throughout multiple emails over a period of time may be important.

One issue with the template-based uniform messaging approach adopted here is that although it may benefit efficient scanning, this efficiency comes largely from the uniformity of the phrases used in the messages. This results in generic, rather than specific descriptions of relations. Participants in our study also alluded to this, by suggesting alternative message designs focused on specific aspects of their needs such as "convey the strength of topical similarity" (3/14 participants), "directly say a new author is worth checking out" (5/14 participants), and "describe the topical context of the author-author connections" (9/14 participants). Furthermore, in exploratory scenarios under time pressure where multiple sources of information compete for user's attention and judgement of relevance, the amount of effort required for goal-specific interpretation of messages may be prohibitively high. Addressing this in future designs could greatly improve the effectiveness of relevance messages by reducing the gap of interpretability.

One potential design space is to suggest possible goals to scientists, and adaptively changing the message templates based on their choice. Because this additional discovery step of available options may add cost to the messaging strategies, recommender systems may proactively present variations of the messages and provide scientists a selection mechanism with minimal cognitive load. In addition to message presentation and selection, alternative measures of relevance may be computed and used to quantify the strength of relevance, supporting scientists' choice of the messaging strategy. For example, the strength of relevance may be computed as a normalized count of connections the recommended paper has to a user-curated feed, and further augmented with topical similarity to each paper on the feed by leveraging readily available NLP techniques.

In sum, an important challenge here for designers is to provide tailored support for a diverse set of goals that newly emerged with relevance messages. Making an assumption about default user goals and supporting only static messages corresponding to them, or

Design Implication	Areas	Summary
Goal-Centric Interpretability	C1, C2, B[2–6]	Messages need to be tailored to scientists' specific goals in order to work as a useful source of signal.
Task-Centric Configurations	C[2–4]	Messages need to support task-specific configuration needs: *here and now* vs. *there and then.*
Dynamic Scientist Identities	C3, B4, B5	Author-based relevance needs to better capture and organize relevance through multiple scientist identities and temporally changing communities.

Table 5: Overview of Design Implications. Using the benefits and challenges identified in Section 6.3, we synthesized design implications for future relevance messaging approaches that combine citation graphs and implicit social networks of authors.

generic messages that pose an interpretability gap may ultimately lead to user frustration and abandonment.

Task-Centric Configurations. Scientists need better support for their specific task context and ways to effectively manage their limited attention. The fundamental challenge is that scientists frequently experience information overload, and while personalized neural recommenders can help they can still result in an overwhelming amount of information to comb through. The three high-level tasks that scientists in our study commonly performed were filtering, sensemaking with mental models of other scientists, and discovery. Scientists expressed needs for additional support in configuring the messages to filter based on blacklisted/whitelisted authors, in prioritizing messages with specific citation context and importance, and in dialing up the salience of important relations between authors featured in the messages.

While participants appreciated existing modalities of longer-term configurations such as tuning the frequency of alert emails (mentioned by 5/14 participants) and the ability to receive new paper recommendations based on a set of curated papers or certain authors, relevance messages also introduced several user needs for additional configurations that apply within an individual email. For example, participants wanted to configure messages to feature additional 'academic status' (3/14 participants) and 'topic cards' (4/14 participants) next to unfamiliar author names to quickly get a sense of their work, and similarly 'citation context' (8/14 participants) next to citation-based messages. We saw that sometimes recommendations in an email acted as a launchpad for separate discovery loops, for example by clicking author links to jump to the author details pages of people they wanted to learn more about (8/14 participants), though there were differences among the participants in terms of when they intended to actively engage in the discovery loops (e.g., as soon as they open the details page (3/14 participants) vs. storing them as open tabs for review at a later time (5/14 participants)). The challenge for designers here is how to notice when filtering-oriented tasks evolve into discovery and vice versa, such that task-specific configurations can be effectively and adaptively supported. The changing nature of the tasks noted here also nods to the findings in information seeking behaviors (e.g., [99, 121]) that involve alternating between broad foraging and focused exploiting phases. From our observations, participants may need two distinctive types of configuration support for tasks that are 'here and now,' which are characteristic to scientists' *in situ* filtering and micro-discovery needs within an alert email, vs. 'there and then,' which are more relevant to configurations expected to last longer thus re-configured less frequently.

Dynamic Scientist Identities. Relevance messages that leverage implicit social networks of authors should better reflect scientists' evolving research interests. One potential design space is in bootstrapping temporal shifts in topical interests by identifying subgroups of scientists from prior publications and interaction data (e.g., curated research feeds and libraries around specific topics). Inferring temporally changing community structures from citation-graphs and implicit social networks of authors may also be useful for better supporting user needs in understanding how scientists are connected in a community and how within- and cross-community impact occurs. Another potential design space is in supporting multiple *scientist identities* for individual authors. While publication history can be a useful source of signal, it also consists of a collection of related research areas that the scientist has previously embarked on, some of which may no longer be relevant to other scientists' interests. Bootstrapping the scientist-topic structures by segmenting time periods with high topical consistency from the publication history may be helpful. In addition, increasing the recency and topical relevance by simple discounting of older publications may be effective.

Support for multiple scientist identities can also help mitigate the message–recommendation incongruence. Featuring messages on paper recommendations that are topically distant can confuse users into thinking that the recommended paper is highly relevant. Yet, due to post-hoc generation, messages that feature relevance through an overall thread of research may augment papers that belong to its topically distant segments. In some cases participants were left wondering how to make sense of the conflicting signals of 'a high confidence of utility signal from the message' (P14) but low perceived topical relevance of the recommendation.

Therefore, designers who wish to use author-based relevance information to augment scientific recommendations need to consider dynamic changes happening within an individual scientist's career trajectory as well as the community-level shifts over time. Capturing and organizing author-based relevance using such structures has the potential to better orient scientists in the multifaceted and dynamic space of relevance.

8 DISCUSSION

Inferring targeted social networks for messages. Previous work on social explanations has shown that social explanations can be persuasive, but they might be only a secondary effect to the people's inherent preferences and quality expectations about the recommended items [102]. This was consistent with our observations

— the engagement benefit of messages for scientists was also influenced by other context such as topical alignment, freshness, and quality cues from the content and metadata of the paper (title, author names, publication venues, etc.).

A prominent difference between the social messages developed in this work and those in prior work lies in how targeted the network leveraged for message generation was for the recommendation task. On one end of the spectrum, explicit social networks leveraged in [101, 102] (e.g., Facebook) may capture real-world friendships among the users, but may also be less targeted for the task of recommending music, as friendship relations between two people encompass more than just similar tastes in music. On the other end, our inferred network of authors may lack the real-world relationships among the users, but may still capture task-specific relations about what different scientists read and how they build on each other's work. Our study results confirmed that the relations featured in messages indeed were relevant in this regard, and also useful because they activated existing mental models of the familiar authors featured in messages to improve their understanding of the recommended papers or the novel authors. There may also be a hybrid of networks for generating social messages. For example, [42] explored one approach to aggregating two types of relations: those from an explicit, public social network and others from a more targeted, within-organization intranet, in order to find most relevant and trustworthy co-workers to support the given recommendation task. The use of different types of social networks raises interesting open questions as to how they differ in terms of message persuasiveness (i.e., how much explanations boost user engagement) and informativeness (i.e., user satisfaction from consuming the recommended item). Answering these questions requires studies that directly measure user perceptions before and after consuming the recommendations augmented with various types of messages pulled from different kinds of networked relations, that go beyond the persuasiveness focus (operationalized by click-through rates) of this work. Understanding these questions may bring new insights as to how different types of social network relations may be combined or used to facet one another to improve the persuasiveness and informativeness of recommendations.

Revisiting the contrast between informativeness and persuasiveness. Prior work has also contrasted the informativeness and persuasiveness of recommendations [16], and investigated how applying social explanations often did not simultaneously increase both [102]. This contrast is interesting to revisit in light of our findings for a few reasons. First, the structure of trust may be more targeted in the inferred scholarly network than in explicit social networks. For example, [102] found that users trusted good friends more and perceived social explanations featuring them as more persuasive than when featuring random friends, but this persuasiveness was not highly correlated with their ultimate liking of the music after consumption. However, this finding may be due to the specific tie strength used to categorize friends in the work (e.g., good vs. random) that was less targeted to the music recommendation task. In contrast, using an inferred scholarly network of authors from publications and their citation network may have a benefit of surfacing more targeted relations useful for judging the ultimate utility of the recommendations, even though the relations may not be grounded in the real-world social relations. Future work may test this hypothesis, and also further explore our findings around the challenges related to interpreting and trusting the inferred relations.

Second, the ultimate informativeness measure may be more nuanced in the domain of scholarly recommendations due to potential differences in their utility curve. Unlike other domains such as music or movie recommendations, scientists are both consumers and producers of the recommendations, rather than consumers alone. Therefore, the users are incentivized by and actively look for recommendations with both immediate (e.g., a reference to cite in the manuscript currently writing) and longer term utility goals (e.g., papers in new domains for future research) in mind. This suggests depending on the task context, each user may exhibit a different informativeness utility curve of recommendations and more or less openness to diverse recommendations that have different kinds of informativeness. Lastly, the scientific community consists of both lateral (e.g., peers) and vertical relations (e.g., direct advisory relations, or distant advisory relations through academic lineage) that users may leverage to interpret messages. Therefore, the notion of trust here is not limited to the immediate closeness among the friends as leveraged in previous work, and supports the use of intermediate author relations (e.g., 'I want to know what my peers or my advisor's former collaborators are working on', 'I trust B's recommendations because he's a well-known expert in the field'). Our findings from indirect author-based messages mediated by middle authors opened a new design space and also uncovered challenges for identifying trusted and preferred middle authors from whom users might appreciate recommendations. Important questions remain open for future investigation, such as how the systematic choice of middle authors shifts the system-wide visibility of work (e.g., 'is the effect similar to crowning an author with a prestigious, status-conferring prize?' [97]), whether it concentrates, reinforces, or creates new pathways to the power of persuasion by specific authors (e.g., 'how does the dynamic differ or align with the megaphone effect [75]?'), and if so, how the goals of trust and fairness may be balanced in the selection algorithm.

Limitations and broader impacts. The methods developed here were evaluated on a particular search engine and within the email alert context. However, given that our approaches are agnostic to the underlying recommender algorithms, they may be applied to other settings, for example, to supplement recommendations found on social media such as Twitter. Future work investigating whether our findings generalize to such settings may be important. More investigation is also needed to understand the full impact of interventions like ours to ensure fairness with respect to other important attributes, including author gender and ethnicity, and the full scope of the impact of the increased engagement, including unintended consequences such as coming at the expense of other important activities. In addition, whether our messages influence diversity of consumption (e.g., [8]) remains an important open question.

Our operationalization of author reputation is also limited in several ways. While we looked at multiple ways of aggregating h-index across authors of a paper (max and average), other aspects

of authorship such as the reputation of the lead author or the last author may be important to consider. We also considered only the changes in papers clicked directly from the email, whereas our intervention could potentially also influence downstream papers clicked in the future.

Finally, the real-world nature of Study 1 introduced some unavoidable measurement error. For example, due to the wide range of devices and event context, it was not possible to ensure that we accurately detected all events such as email opens. It is also very challenging to automatically extract knowledge graph elements such as citations and disambiguate authors, particularly for newly published papers which may not be accompanied by publisher metadata, and errors in this process introduced additional noise into the deployment. While it is unlikely that significant biases were introduced into experimental conditions in a systematic way because participation was randomized, improving the data coverage and accuracy from the deployment setting will be valuable for surfacing the full scopes of effects.

9 CONCLUSION

Scientists today are faced with a daunting yet fundamental task of staying on top of the large, rapidly growing literature. With little support from existing tools to know why the recommended papers might be worth their attention to read, scientists are forced to wrangle long, monotonous lists of recommendations—and perhaps quit in the process. To better support scientists' broad information needs and mitigate the issue of scarce relevance signals, we designed and empirically tested two kinds of graph-based relevance messages by finding connections from who they know to what they read. Our large-scale, real-world online deployment study revealed empirical evidence that relevance messages are an effective means for engaging scientists. To further increase their benefits, we designed and implemented a third kind of messages via an inferred scholarly network and featuring relations mediated by middle authors that the user may trust. From a controlled lab study with 14 scientists we gained qualitative insights into the utility of our relevance messages as well as their challenges. Finally, we synthesized a set of future implications for design, which aim to use inferred social network relevance to engage scientists. We envision a future in which scientists are delighted to keep up with academic literature through personalized paper recommendations that help them attend to authors they know, discover new interesting authors found from what they have read or interacted with in the past, and provide many additional engaging and helpful signals that feed into a positive loop of further improving the recommender systems.

ACKNOWLEDGMENTS

This project is supported in part by NSF Grant OIA-2033558, NSF RAPID award 2040196, ONR grant N00014-21-1-2707, and the Allen Institute for Artificial Intelligence (AI2). The authors thank Alex Schokking, Alex Buraczynski, Paul Sayre, Cecile Nguyen, Sebastian Kohlmeier, and Rodney Kinney for their advice on and help with engineering and Kyle Lo for his advice on the analysis of experimental results. We also thank the anonymous reviewers for their constructive feedback. Finally, this work would not have been possible without our study participants.

REFERENCES

[1] Alison Abbott, David Cyranoski, Nicola Jones, Brendan Maher, Quirin Schiermeier, and Richard Van Noorden. 2010. Metrics: Do metrics matter? *Nature News* 465, 7300 (2010), 860–862.
[2] Daniel E Acuna, Stefano Allesina, and Konrad P Kording. 2012. Predicting scientific success. *Nature* 489, 7415 (2012), 201–202.
[3] Chunrong Ai and Edward C Norton. 2003. Interaction terms in logit and probit models. *Economics letters* 80, 1 (2003), 123–129.
[4] Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes* 50, 2 (1991), 179–211.
[5] Dag W Aksnes and Arie Rip. 2009. Researchers' perceptions of citations. *Research Policy* 38, 6 (2009), 895–905.
[6] Paul David Allison. 1999. *Multiple Regression: A Primer (Pine Forge Press Series in Research Methods and Statistics)*. Unspecified.
[7] Reid Andersen, Christian Borgs, Jennifer Chayes, Uriel Feige, Abraham Flaxman, Adam Kalai, Vahab Mirrokni, and Moshe Tennenholtz. 2008. Trust-based recommendation systems: an axiomatic approach. In *Proceedings of the 17th international conference on World Wide Web*. 199–208.
[8] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. 2020. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference 2020*. 2155–2165.
[9] Andy Barrett. 2005. The information-seeking habits of graduate student researchers in the humanities. *The Journal of Academic Librarianship* 31, 4 (2005), 324–331.
[10] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).
[11] Marcia J Bates. 2002. Speculations on browsing, directed searching, and linking in relation to the Bradford distribution. In *Emerging frameworks and methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS 4)*. Libraries Unlimited Greenwood Village, CO, 137–150.
[12] Charles Bazerman. 1985. Physicists reading physics: Schema-laden purposes and purpose-laden schema. *Written communication* 2, 1 (1985), 3–23.
[13] Nicholas J Belkin and Alina Vickery. 1985. *Interaction in information systems: A review of research from document retrieval to knowledge-based systems*. Number 025.04 BEL. CIMMYT.
[14] Adar Ben-Eliyahu, Debra Moore, Rena Dorph, and Christian D Schunn. 2018. Investigating the multidimensionality of engagement: Affective, behavioral, and cognitive engagement across science activities and contexts. *Contemporary Educational Psychology* 53 (2018), 87–105.
[15] Steven Bethard and Dan Jurafsky. 2010. Who should I cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 609–618.
[16] Mustafa Bilgic and Raymond J Mooney. 2005. w. In *Beyond personalization workshop, IUI*, Vol. 5. 153.
[17] Thijs Bol, Mathijs de Vaan, and Arnout van de Rijt. 2018. The Matthew effect in science funding. *Proceedings of the National Academy of Sciences* 115, 19 (2018), 4887–4890.
[18] Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology* 66, 11 (2015), 2215–2222.
[19] Richard E Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development*. sage.
[20] Corinna Breitinger, Patrick Wortner, Bela Gipp, and Harald Reiterer. 2019. 'Too Late to Collaborate': Challenges to the Discovery of in-Progress Research. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 134–137.
[21] Donald O Case and Lisa M Given. 2016. Looking for information: A survey of research on information seeking, needs, and behavior. (2016).
[22] Rose Catherine and William Cohen. 2016. Personalized recommendations using knowledge graphs: A probabilistic logic programming approach. In *Proceedings of the 10th ACM conference on recommender systems*. 325–332.
[23] Sneha Chaudhari, Amos Azaria, and Tom Mitchell. 2017. An entity graph based recommender system. *AI Communications* 30, 2 (2017), 141–149.
[24] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. 2010. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1185–1194.
[25] Robert B Cialdini. 1987. *Influence*. Vol. 3. A. Michel Port Harcourt.
[26] Robert B Cialdini and Noah J Goldstein. 2004. Social influence: Compliance and conformity. *Annu. Rev. Psychol.* 55 (2004), 591–621.
[27] Avital Cnaan, Nan M Laird, and Peter Slasor. 1997. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in medicine* 16, 20 (1997), 2349–2380.
[28] Stephen Cole and Jonathan R Cole. 1968. Visibility and the structural bases of awareness of scientific research. *American sociological review* (1968), 397–413.

[29] Lucas Colusso, Ridley Jones, Sean A Munson, and Gary Hsieh. 2019. A translational science model for HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[30] Lisa M Covi. 1999. Material mastery: situating digital library use in university research practices. *Information Processing & Management* 35, 3 (1999), 293–316.

[31] Diana Crane. 1965. Scientists at major and minor universities: A study of productivity and recognition. *American sociological review* (1965), 699–714.

[32] Mihaly Csikszentmihalyi and Mihaly Csikzentmihaly. 1990. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row New York.

[33] Robert Cudeck. 1996. Mixed-effects models in the study of individual differences with repeated measures data. *Multivariate behavioral research* 31, 3 (1996), 371–403.

[34] Beant Dhillon, Peter Banach, Rafal Kocielnik, Jorge Peregrin Emparanza, Ioannis Politis, A Rączewska, and Panos Markopoulos. 2011. Visual fidelity of video prototypes and user feedback: a case study. In *Proceedings of HCI 2011 The 25th BCS Conference on Human Computer Interaction 25*. 139–144.

[35] Thomas A DiPrete and Gregory M Eirich. 2006. Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annu. Rev. Sociol.* 32 (2006), 271–297.

[36] Catherine A Durham and Diego Andrade. 2005. *Health vs. environmental motivation in organic preferences and purchases*. Technical Report.

[37] Zhichao Fang, Rodrigo Costas, Wencan Tian, Xianwen Wang, and Paul Wouters. 2021. How is science clicked on Twitter? Click metrics for Bitly short links to scientific publications. *Journal of the Association for Information Science and Technology* 72, 7 (2021), 918–932.

[38] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[39] Adam M Grant and Barry Schwartz. 2011. Too much of a good thing: The challenge and opportunity of the inverted U. *Perspectives on psychological science* 6, 1 (2011), 61–76.

[40] Paul Grau, Babak Naderi, and Juho Kim. 2018. Personalized Motivation-supportive Messages for Increasing Participation in Crowd-civic Systems. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22.

[41] Ido Guy, Michal Jacovi, Elad Shahar, Noga Meshulam, Vladimir Soroka, and Stephen Farrell. 2008. Harvesting with SONAR: the value of aggregating social network information. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1017–1026.

[42] Ido Guy, Naama Zwerdling, David Carmel, Inbal Ronen, Erel Uziel, Sivan Yogev, and Shila Ofek-Koifman. 2009. Personalized recommendation of social software items based on social relations. In *Proceedings of the third ACM conference on Recommender systems*. 53–60.

[43] Saeed-Ul Hassan, Anam Akram, and Peter Haddawy. 2017. Identifying important citations using contextual information from full text. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 1–8.

[44] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. Trirank: Review-aware explainable recommendation by modeling aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 1661–1670.

[45] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S Weld, and Marti A Hearst. 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[46] Marti A Hearst, Emily Pedersen, Lekha Patil, Elsie Lee, Paul Laskowski, and Steven Franconeri. 2019. An evaluation of semantically grouped word cloud designs. *IEEE transactions on visualization and computer graphics* 26, 9 (2019), 2748–2761.

[47] Paul Hemp. 2009. Death by information overload. *Harvard business review* 87, 9 (2009), 82–9.

[48] Monika Henzinger, Jacob Suñol, and Ingmar Weber. 2010. The stability of the h-index. *Scientometrics* 84, 2 (2010), 465–479.

[49] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 241–250.

[50] Diana Hicks, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. 2015. Bibliometrics: the Leiden Manifesto for research metrics. *Nature News* 520, 7548 (2015), 429.

[51] Terje Hillesund. 2010. Digital reading spaces: How expert readers handle books, the Web and electronic paper. (2010).

[52] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences* 102, 46 (2005), 16569–16572.

[53] Jorge E Hirsch. 2007. Does the h index have predictive power? *Proceedings of the National Academy of Sciences* 104, 49 (2007), 19193–19198.

[54] Jeffrey D Hole. 2008. *Email overload in academia*. Rochester Institute of Technology.

[55] Kim Holmberg and Mike Thelwall. 2014. Disciplinary differences in Twitter scholarly communication. *Scientometrics* 101, 2 (2014), 1027–1042.

[56] Gary Hsieh and Rafał Kocielnik. 2016. You get who you pay for: The impact of incentives on participation bias. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 823–835.

[57] Jinkyu Jang and Jinwoo Kim. 2020. Healthier life with digital companions: Effects of reflection-level and statement-type of messages on behavior change via a perceived companion. *International Journal of Human–Computer Interaction* 36, 2 (2020), 172–189.

[58] Haofeng Jia and Erik Saule. 2017. An analysis of citation recommender systems: Beyond the obvious. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. 216–223.

[59] Arif E Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing* 23, 3 (2010), 258–263.

[60] Clay A Johnson. 2015. *The information diet: A case for conscious comsumption*. " O'Reilly Media, Inc.".

[61] Ron Johnston, Kelvyn Jones, and David Manley. 2018. Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. *Quality & quantity* 52, 4 (2018), 1957–1976.

[62] Timothy A Judge, Daniel M Cable, Amy E Colbert, and Sara L Rynes. 2007. What causes a management article to be cited—article, author, or journal? *Academy of management journal* 50, 3 (2007), 491–506.

[63] Samara Klar, Yanna Krupnikov, John Barry Ryan, Kathleen Searles, and Yotam Shmargad. 2020. Using social media to promote academic research: Identifying the benefits of twitter for sharing academic work. *PloS one* 15, 4 (2020), e0229446.

[64] Rafal Kocielnik and Gary Hsieh. 2017. Send me a different message: utilizing cognitive space to create engaging message triggers. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2193–2207.

[65] Esther Landhuis. 2016. Scientific literature: Information overload. *Nature* 535, 7612 (2016), 457–458.

[66] Vincent Larivière and Yves Gingras. 2010. The impact factor's Matthew Effect: A natural experiment in bibliometrics. *Journal of the American Society for Information Science and Technology* 61, 2 (2010), 424–427.

[67] Jure Leskovec, Ajit Singh, and Jon Kleinberg. 2006. Patterns of influence in a recommendation network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 380–389.

[68] Mary J Lindstrom and Douglas M Bates. 1990. Nonlinear mixed effects models for repeated measures data. *Biometrics* (1990), 673–687.

[69] Bin Liu and Zheng Yuan. 2010. Incorporating social networks and user opinions for collaborative recommendation: local trust network based method. In *Proceedings of the workshop on context-aware movie recommendation*. 53–56.

[70] Christopher Lueg. 1997. Social filtering and social reality. In *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*. ERCIM Press, 77–81.

[71] Hao Ma, Irwin King, and Michael R Lyu. 2009. Learning to recommend with social trust ensemble. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 203–210.

[72] Brian James McInnis, Elizabeth Lindley Murnane, Dmitry Epstein, Dan Cosley, and Gilly Leshed. 2016. One and Done: Factors affecting one-time contributors to ad-hoc online communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 609–623.

[73] Marcia McNutt. 2014. The measure of research merit.

[74] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.

[75] Edward F McQuarrie, Jessica Miller, and Barbara J Phillips. 2013. The megaphone effect: Taste and audience in fashion blogging. *Journal of Consumer Research* 40, 1 (2013), 136–158.

[76] Marshall H Medoff. 2006. Evidence of a Harvard and Chicago Matthew effect. *Journal of Economic Methodology* 13, 4 (2006), 485–506.

[77] Anamika Megwalu. 2015. Academic social networking: a case study on users' information behavior. In *Current Issues in Libraries, Information Science and Related Fields*. Emerald Group Publishing Limited.

[78] Robert K Merton. 1968. The Matthew effect in science: The reward and communication systems of science are considered. *Science* 159, 3810 (1968), 56–63.

[79] Breed D Meyer. 1995. Natural and quasi-experiments in economics. *Journal of business & economic statistics* 13, 2 (1995), 151–161.

[80] James G Miller. 1960. Information input overload and psychopathology. *American journal of psychiatry* 116, 8 (1960), 695–704.

[81] Ehsan Mohammadi, Mike Thelwall, Mary Kwasny, and Kristi L Holmes. 2018. Academic information on Twitter: A user survey. *PloS one* 13, 5 (2018), e0197265.

[82] Francesco Montani, Christian Vandenberghe, Anis Khedhaouria, and François Courcy. 2020. Examining the inverted U-shaped relationship between workload and innovative work behavior: The role of work engagement and mindfulness. *Human Relations* 73, 1 (2020), 59–93.

[83] Frederick Muench and Amit Baumel. 2017. More than a text message: dismantling digital triggers to curate behavior change in patient-centered health interventions. *Journal of medical Internet research* 19, 5 (2017), e147.
[84] David Nicholas, Peter Williams, Ian Rowlands, and Hamid R Jamali. 2010. Researchers'e-journal use and information seeking behaviour. *Journal of Information Science* 36, 4 (2010), 494–516.
[85] Richar Van Noorden. 2014. Global scientific output doubles every nine years. http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html
[86] Robert M O'Brien. 2017. Dropping highly collinear variables from a model: why it typically is not a good idea. *Social Science Quarterly* 98, 1 (2017), 360–375.
[87] Vito Claudio Ostuni, Tommaso Di Noia, Eugenio Di Sciascio, and Roberto Mirizzi. 2013. Top-n recommendations from implicit feedback leveraging linked open data. In *Proceedings of the 7th ACM conference on Recommender systems*. 85–92.
[88] Elisabeth Pain. 2016. How to keep up with the scientific literature. *Science Careers* 30 (2016).
[89] Carole L Palmer. 2005. Scholarly work and the shaping of digital access. *Journal of the American Society for Information Science and Technology* 56, 11 (2005), 1140–1153.
[90] Carole L Palmer, Lauren C Teffeau, and Carrie M Pirmann. 2009. Scholarly information practices in the online environment. *Report commissioned by OCLC Research. Published online at: www. oclc. org/programs/publications/reports/2009-02. pdf* (2009).
[91] John Panaretos and Chrisovaladis Malesios. 2009. Assessing scientific research performance and impact with single indices. *Scientometrics* 81, 3 (2009), 635–670.
[92] Saverio Perugini, Marcos André Gonçalves, and Edward A Fox. 2004. Recommender systems research: A connection-centric survey. *Journal of Intelligent Information Systems* 23, 2 (2004), 107–143.
[93] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.
[94] Derek de Solla Price. 1976. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science* 27, 5 (1976), 292–306.
[95] Marie L Radford, Vanessa Kitzie, Stephanie Mikitish, Diana Floegel, Gary P Radford, and Lynn Silipigni Connaway. 2020. "People are reading your work," scholarly identity and social networking sites. *Journal of Documentation* (2020).
[96] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. 2008. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences* 105, 45 (2008), 17268–17272.
[97] Brian P Reschke, Pierre Azoulay, and Toby E Stuart. 2018. Status spillovers: The effect of status-conferring prizes on the allocation of attention. *Administrative Science Quarterly* 63, 4 (2018), 819–847.
[98] Stephen Rowland. 2002. Overcoming fragmentation in professional life: The challenge for academic development. *Higher education quarterly* 56, 1 (2002), 52–64.
[99] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. ACM, 269–276.
[100] Masahiro Sato, Shin Kawai, and Hajime Nobuhara. 2019. Action-triggering recommenders: Uplift optimization and persuasive explanation. In *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 1060–1069.
[101] Amit Sharma and Dan Cosley. 2011. Network-centric recommendation: Personalization with and in social networks. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, 282–289.
[102] Amit Sharma and Dan Cosley. 2013. Do social explanations work? Studying and modeling the effects of social explanations in recommender systems. In *Proceedings of the 22nd international conference on World Wide Web*. 1133–1144.
[103] Herbert A Simon. 1996. Designing organizations for an information-rich world. *International Library of Critical Writings in Economics* 70 (1996), 187–202.
[104] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. 2016. Quantifying the evolution of individual scientific impact. *Science* 354, 6312 (2016).
[105] Parag Singla and Matthew Richardson. 2008. Yes, there is a correlation: -from social networks to personal behavior on the web. In *Proceedings of the 17th international conference on World Wide Web*. 655–664.
[106] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[107] John Sweller. 2011. Cognitive load theory. In *Psychology of learning and motivation*. Vol. 55. Elsevier, 37–76.
[108] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. 2008. Providing justifications in recommender systems. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 38, 6 (2008), 1262–1272.
[109] Carol Tenopir, Donald W King, Sheri Edwards, and Lei Wu. 2009. Electronic journals and changes in scholarly article seeking and reading patterns. In *Aslib proceedings*. Emerald Group Publishing Limited.
[110] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop*. IEEE, 801–810.
[111] Nava Tintarev and Judith Masthoff. 2008. The effectiveness of personalized movie explanations: An experiment using commercial meta-data. In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*. Springer, 204–213.
[112] George Toderici, Hrishikesh Aradhye, Marius Pasca, Luciano Sbaiz, and Jay Yagnik. 2010. Finding meaning on youtube: Tag recommendation and category discovery. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 3447–3454.
[113] Martin Traunmueller, Paul Marshall, and Licia Capra. 2015. Crowdsourcing safety perceptions of people: Opportunities and limitations. In *International Conference on Social Informatics*. Springer, 120–135.
[114] Julia Vainio and Kim Holmberg. 2017. Highly tweeted science articles: who tweets them? An analysis of Twitter user profile descriptions. *Scientometrics* 112, 1 (2017), 345–366.
[115] Pertti Vakkari and Sanna Talja. 2006. Searching for Electronic Journal Articles to Support Academic Tasks. A Case Study of the Use of the Finnish National Electronic Library (FinELib). *Information Research: An International Electronic Journal* 12, 1 (2006), n1.
[116] Arnout Van de Rijt, Soong Moon Kang, Michael Restivo, and Akshay Patil. 2014. Field experiments of success-breeds-success dynamics. *Proceedings of the National Academy of Sciences* 111, 19 (2014), 6934–6939.
[117] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2018. Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 417–426.
[118] Jian Wang. 2014. Unpacking the Matthew effect in citations. *Journal of Informetrics* 8, 2 (2014), 329–339.
[119] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable recommendation via multi-task learning in opinionated text data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 165–174.
[120] Lynn Westbrook. 2003. Information needs and experiences of scholars in women's studies: Problems and solutions. *College & Research Libraries* 64, 3 (2003), 192–209.
[121] Ryen W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–98.
[122] Steve Whittaker and Candace Sidner. 1996. Email overload: exploring personal information management of email. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 276–283.
[123] Simon Williams. 2019. Postgraduate Research Experience Survey. *London: Advance HE* (2019).
[124] Max L Wilson et al. 2008. Improving exploratory search interfaces: Adding value or information overload? (2008).
[125] Chris Woolston. 2019. PhDs: the tortuous truth.
[126] Yongfeng Zhang and Xu Chen. 2018. Explainable recommendation: A survey and new perspectives. *arXiv preprint arXiv:1804.11192* (2018).
[127] Guoshuai Zhao, Hao Fu, Ruihua Song, Tetsuya Sakai, Zhongxia Chen, Xing Xie, and Xueming Qian. 2019. Personalized reason generation for explainable song recommendation. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 4 (2019), 1–21.
[128] Xin Wayne Zhao, Yanwei Guo, Yulan He, Han Jiang, Yuexin Wu, and Xiaoming Li. 2014. We know what you want to buy: a demographic-based system for product recommendation on microblogs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1935–1944.

Appendix A. Post-hoc analyses on the potential sources of biases of randomization

We further ran post-hoc analyses on two potential sources of biases – the average length of emails and the average position of relevance messages – to ensure the validity of random assignment. Systemic differences in these biases between conditions may significantly change user engagement, and have potential to invalidate the interpretation of the results on engagement. For example, longer emails may fatigue users and cause them to drop off more easily.

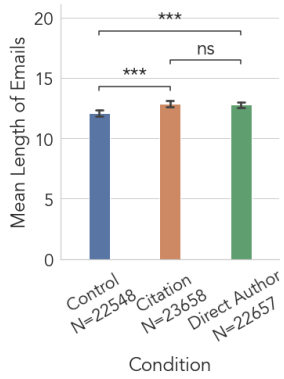


Figure 13: The avg. length of emails were slightly shorter in Control ($\mu = 12.1, \bar{\sigma} = 20.30$) than the other two conditions ($\mu = 12.9, \bar{\sigma} = 21.74$ in Citation and $\mu = 12.8, \bar{\sigma} = 17.36$ in Direct Author), and the difference was less than one paper.

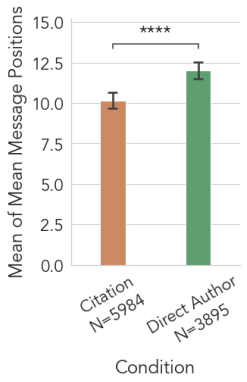


Figure 14: The mean of mean rank positions of messages in emails was significantly higher in Direct Author than in Citation, suggesting receivers of direct author messages had to read more through the list in the email to see the messages.

At the same time, longer emails may also provide users more paper recommendations to look at, and as a result an increased click engagement. On the other hand, the average position of relevance messages would positively impact user discovery and attention, because messages placed higher (towards the top of the email) are more easily discovered.

These potential sources of biases appear unlikely to have changed the directions of the observed effects on engagement (Fig. 13 and 14). Though the length of emails *did* significantly differ between conditions and on average the emails in the Citation and Direct Author conditions were longer than in Control, this difference was less than a paper in both cases ($\Delta_{\text{Citation}-\text{Control}} = 0.78$ and $\Delta_{\text{Direct Author}-\text{Control}} = 0.69$). The average number of paper recommendations clicked per email was $\mu = 0.10, \bar{\sigma} = 0.680$; the expected number of clicked emails from the difference is then $0.78 \times 0.10 = 0.078$ for the Citation condition and $0.69 \times 0.10 = 0.069$ for the Direct Author condition. However, the average number of clicked emails in both treatment conditions was still higher after accounting for the expected surpluses of clicked papers, $\mu = 0.24, \bar{\sigma} = 0.936$ (Citation) and $\mu = 0.36, \bar{\sigma} = 1.262$ (Direct Author). Between the Citation and Direct Author conditions, the average position of relevance messages indexed from the top of the email was smaller (i.e. closer to the top of the email) in the Citation condition ($\mu = 10.1, \bar{\sigma} = 20.55$) than in the Direct Author condition ($\mu = 12.0, \bar{\sigma} = 17.72$). This suggests that if the discoverability of relevance messages were controlled for, Direct Author messages may lead to an even greater increase in user engagement.

Appendix B. Difference-in-Differences (DiD) analysis on the email open rates

We further examined the effect of messaging strategies on future engagement, while accounting for the effect of familiarity and habit-forming with email alerts over time. Though Fig. 4b showed the overall differences in email open rates between conditions, it is possible that there was a significant time effect on users' decision to open an incoming alert email, given their previous experiences with them, such as habitual opening. This would likely create a positive trend-line for the baseline email open rates over time, thus making the estimation of true effects of our messaging strategies harder to obtain. We applied Difference in differences analysis to account for such effect. *DiD* is a technique often used in econometrics and social sciences to derive causal inferences from observational, panel data [79]. We used efficient linear probability estimation method based on linear regression models. This has an additional benefit of producing directly interpretable coefficients [3].

We applied two symmetric linear regression models, each corresponding to one messaging strategy (i.e. Citation or Direct Author). The regression model's dependent variable was the binary open rates (1: Email was opened, 0: Otherwise), and the predictive variables included whether the email was sent within the first two weeks of the experiment (Early Exposure), whether the email was in the treatment condition (vs. Control), and the interaction of the two. The number of emails in the Early Exposure vs. Late Exposure groups was roughly equal, with the biggest difference occurring in

	Coef.	SE	p
(Intercept)	0.256	0.004	***
Early Exposure	-0.037	0.006	***
Message (*Citation*)	0.006	0.006	0.29
Interaction	-0.004	0.008	0.64

Table 6: The result of linear regression with Citation messages and Control. Citation messages did not increase email open rates after accounting for the baseline increase over time ($p = 0.64$). ***: $p < 0.001$, **: $p < 0.01$.

	Coef.	SE	p
(Intercept)	0.256	0.004	***
Early Exposure	-0.037	0.006	***
Message (*Direct Author*)	0.045	0.006	***
Interaction	-0.031	0.008	***

Table 7: The result of linear regression with Direct Author messages and Control. Direct Author messages significantly increased email open rates in addition to the baseline increase over time ($p < 0.001$). ***: $p < 0.001$, **: $p < 0.01$.

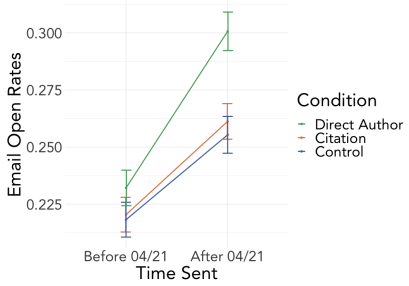


Figure 15: Mean email open rates by condition shows a significantly steeper slope in Direct Author over Control, suggesting its effectiveness in boosting the open rate after accounting for the baseline increase over time.

the Citation condition (11,459 emails in the early exposure group vs. 12,199 in the late exposure group).

The result of our analyses showed that indeed there was a significant time effect on the open rates in both conditions. Specifically, Early Exposure was significantly *negatively* associated with the open rates, suggesting that users became increasingly 'habitual' with opening new alert emails over time (Table 6 and 7). After accounting for this time effect, the effect of messaging was significant only in the Direct Author condition ($p < 0.001$), but not in the Citation condition ($p = 0.64$), further lending support to the effectiveness of the author-based messaging strategy (Fig. 15). Specifically, later Direct Author messages induced a 30% increase in email open rates relative to the first two-week exposure period.

Appendix C. Empirical relationship between the predictive variables and click-through rates (CTR)

The Locally Weighted Scatterplot Smoothing (LOWESS) plot showed a roughly inverted U-shaped curve between CTR the % of paper recommendations featured with *any* type of relevance message, with a greater fraction of treated papers resulting in dramatically more engagement, up to a point (between 25-50% treated) above which engagement falls off (Fig. 16(b)). Furthermore, it also showed CTR was roughly positively correlated with the receiver's h-index (Fig. 16(a)), and whether they claimed a profile (Fig. 16(c)), showing users with claimed profiles showing overall higher levels of CTR.

Appendix D. Variance Inflation Factor analysis to determine factors that should be pruned due to significant collinearity with others

We performed Variance Inflation Factor (VIF) analysis [6] on the predictor variables (*Claimed Profile*, *% Featured*, and *# of Total Papers*) because they are assumed to have parallel causal influences on *Email Clicked*. The VIF value for each predictor variable is typically obtained by first regressing it against all others in the set, then computing the $1/(1-R^2)$. For example a VIF of 1.5 tells us that the variance of the predictor variable is 50% greater than would be the case if no collinearity was present. Following [61], we used 2.5 as a threshold for the existence of significant collinearity among the predictor variables. The result showed that the highest VIF from the predictor variables was 1.14 (*Claimed Profile*) hence we proceeded without pruning any variable from the model.

	Coef.	SE	p
(Intercept)	-8.20	0.278	***
% Featured	3.17	0.600	***
(% Featured)^2	-3.00	0.680	***
# of Total Papers	0.01	0.038	0.81

Table 8: Regression analysis using Model 1 predicted significant curvilinear changes on CTR from the % of papers featured with author-based relevance messages while the total number of recommendations in the email was not a significant predictor. * indicates significance at $p < 0.001$.**

Appendix E. Likelihood Ratio Test for modeling the relationship between CTR and % of recommendations featured with messages

We hypothesized that an optimal % of papers within an email featured with author-based relevance messages (*% Featured*) for click-through is neither too few (which may lead to under-utilization), nor too many (may lead to overwhelming the users). Investing this curvilinear relation requires inclusion of a quadratic variable in the model. The descriptive pattern from the data also suggested this hypothesis (Fig. 16(b)). Using locally weighted scatterplot smoothing (LOWESS), we observed that the peak click-through happened somewhere between 25% and 50% of the papers featured with author-based relevance messages, and quickly decreased outside this range. Along with the descriptive pattern, we also performed a pairwise Likelihood Ratio Test (LRT) on two models, one with only the linear % Featured term and the other with only the quadratic (% Featured)^2 term, in order to further test the soundness of introducing the quadratic term. The result showed that the reduction of deviance from introducing the quadratic term is more than what we would have expected to see if the beta coefficients for them were 0 ($\chi^2(1) = 19.9, p = 8.1 \times 10^{-6}$), thus we proceeded with introducing the quadratic term and building Model 1:[2].

$$g(E[y]) = \beta_0 + \gamma_j + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$
$$g(E[y]) = \beta_0 + \gamma_j + \beta_4 {x_1}^2 + \beta_5 x_2 + \beta_6 x_3$$

where y=CTR, β_1 representing the fixed effects from the *% Featured* (x_1), β_2 representing the fixed effects from the *Claimed Profile* variable (x_2), β_3 representing the fixed effects from the *# of Total Papers* (x_3), β_4 representing the fixed effects from the quadratic (*% Featured*)^2, and similarly for β_5 and β_6. Random intercepts $\gamma_j \sim \mathcal{N}(0, \sigma_\gamma^2)$ were introduced for participants j's. We used the logit link $g(p) = \log(p/(1-p))$ to model the engagement as a Bernoulli variable.

The result of Model 1 is shown in Table 8. This regression tells us whether the probability of CTR changes significantly if we increased the % of papers featured with author-based relevance messages, or changed the number of paper recommendations included in each email. Consistent with our hypothesis, we found a significant

[2]We did not perform an additional check on the VIF when including the quadratic term, despite its potentially significant collinearity with the linear term. See [86] for more discussion on the relevant topic.

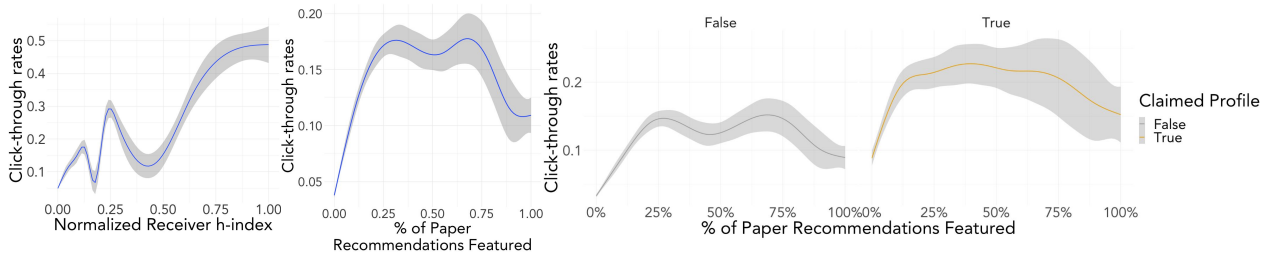
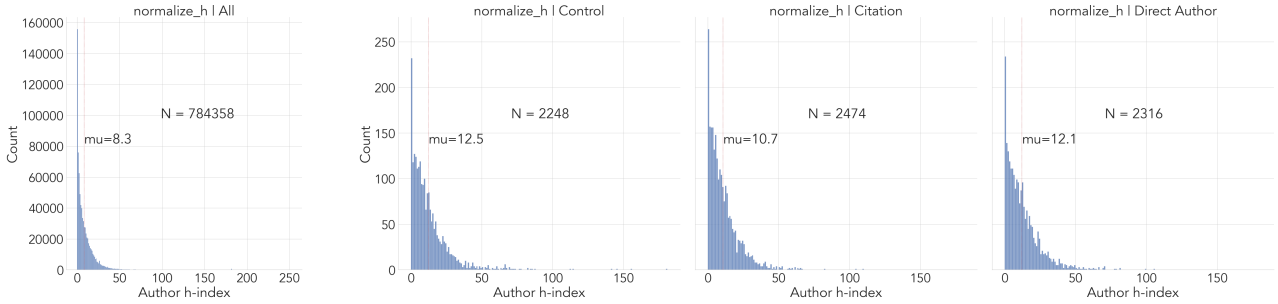


Figure 16: (a) The LOWESS plot suggested an overall positive correlation between receiver h-index and CTR; (b) The LOWESS plot suggested an inverted U-shaped curve relationship between % of paper recommendation featured and CTR, (c) with an overall higher level of CTR for claimed profile users.



(a) The 'background' distribution of normalized author h-indices of all paper recommendations sent out in all alert emails in all conditions. $\mu = 8.3, \bar{\sigma} = 11.14$ (median=5.0).

(b) The distribution of normalized author h-indices of clicked papers in each condition. The average increased to $\mu = 12.5, \bar{\sigma} = 14.25$ (Control, median=8.7); $\mu = 10.7, \bar{\sigma} = 10.87$ (Citation, median=8.0); and $\mu = 12.1, \bar{\sigma} = 12.11$ (Direct Author, median=9.0). The increase of the average h-index over the background distribution was significant, suggesting that users considered high status authors as a signal for deciding whether to click on a paper by default. At the same time, citation messages reduced the h-index relative to control, suggesting its effect of guiding user attention to lesser known authors. Direct author messages and control did not differ significantly.

Figure 17: Analysis showed users on average clicked on papers with authors with higher average h-index, and featuring relevance messages did not shift user attention to only the papers with high-profile authors.

curvilinear relationship between the % of papers featured in the email and user engagement. The coefficients of both the linear % Featured and its quadratic terms were significant ($p < 0.001$), and the signs were in the opposite direction, with a negative beta coefficient for the quadratic term. This validates the empirically observed inverted U-shaped curve relationship on CTR. We also found that the normalized length of email was not a significant predictor of user engagement, which suggests the conflicting effects of longer emails: overwhelming users (negative) vs. providing more paper links to click (positive). We extend Model 1 with additional predictive variables and account for them in our regression analysis using Model 2 (Section 4.2.3).

Appendix F. Repeat analysis on the fairness of visibility (Section 4.2.4) using normalized author h-index

The repeat analyses of author h-index distributional shifts using normalized author h-index per paper recommendation (Fig. 17(b); see also Section. 4.2.4).