# ComLittee: Literature Discovery with Personal Elected Author Committees

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

Nouran Soliman
MIT CSAIL
Cambridge, MA, USA
nouran@mit.edu

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

Joseph Chee Chang
Allen Institute for AI
Seattle, WA, USA
josephc@allenai.org

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

## ABSTRACT

In order to help scholars understand and follow a research topic, significant research has been devoted to creating systems that help scholars discover relevant papers and authors. Recent approaches have shown the usefulness of highlighting relevant authors while scholars engage in paper discovery. However, these systems do not capture and utilize users' evolving knowledge of authors. We reflect on the design space and introduce ComLittee, a literature discovery system that supports author-centric exploration. In contrast to paper-centric interaction in prior systems, ComLittee's author-centric interaction supports curating research threads from individual authors, finding new authors and papers using combined signals from a paper recommender and the curated authors' authorship graphs, and understanding them in the context of those signals. In a within-subjects experiment that compares to a paper-centric discovery system with author-highlighting, we demonstrate how ComLittee improves author and paper discovery.

## CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; User studies.

## KEYWORDS

Scholarly discovery systems, paper and author recommendations, author-augmented literature discovery, interpretable relevance explanations, interactive machine learning

**ACM Reference Format:**
Hyeonsu B. Kang, Nouran Soliman, Matt Latzke, Joseph Chee Chang, and Jonathan Bragg. 2023. ComLittee: Literature Discovery with Personal Elected Author Committees. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3544548.3581371

*Work completed during a research internship at Semantic Scholar Research, Allen Institute for AI.

This work is licensed under a Creative Commons Attribution International 4.0 License.

*CHI '23, April 23–28, 2023, Hamburg, Germany*
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581371

## 1 INTRODUCTION

In order to understand a research area and keep up with the exponentially growing rate of scientific publication (*cf.* [4, 18, 53]), scientists expend significant effort searching for relevant papers and threads of research. To help scientists in different stages of the discovery process many systems have been developed to support finding, triaging, and reading papers (*cf.* [8, 42, 44, 49, 55, 57]). While useful, such systems often limit interaction to the modality of papers and do not incorporate useful relevance signals from authorship graphs. However, identifying authors that work in a research area can further facilitate the discovery process, since authors often publish multiple papers on related topics as they pursue a research agenda, which can be discovered together. Further, the publication networks of these authors could also help with identifying additional relevant authors and papers, for example, through their frequent co-authors who also work on the topic, the relevant papers they often cite, and authors of these relevant papers.

While much research has been devoted to creating literature discovery tools, given these benefits of identifying relevant authors, recent research has begun to develop a new kind of discovery tool that augments encountered papers with highlights of potentially-relevant authors on those papers [20, 23]. Users can use this information to help them decide whether to read encountered papers [20], or explore additional papers written by highlighted authors, which may be relevant [23]. These highlights have been demonstrated to be effective in multiple literature discovery contexts including exploratory paper search [23] and paper recommendation [20]. We term this emerging paradigm of literature discovery systems that support author discovery *author-augmented literature discovery*.

While highlighting relevant authors provides useful context for users of literature discovery systems, current systems give users limited ability to interact with authors. First, because these systems augment encountered papers but do not alter the underlying methods for recommending papers, they have limited ability to help users explore and discover authors beyond those who have authored the set of surfaced papers. Users may miss important threads of research because the returned papers are not optimized to yield good coverage of relevant authors; these systems typically return papers prioritized by predicted relevance, from a sample

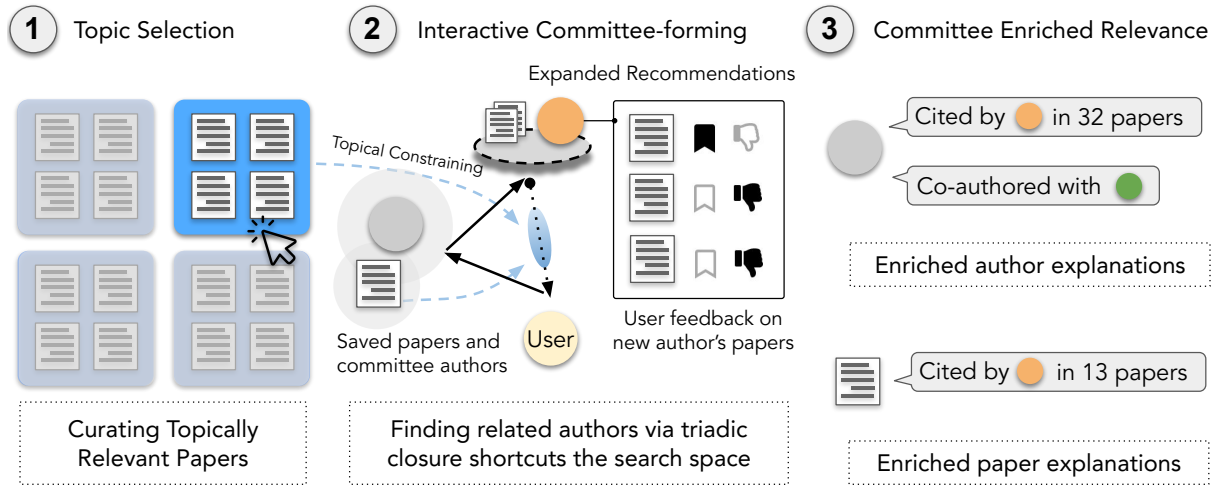



**Figure 1: Construction and application of relevance signals on ComLittee: ① The user saves seed papers relevant to a topic. ② ComLittee recommends an initial set of authors and papers relevant to the selected folder's content. When the user starts forming a committee of authors, ComLittee expands the recommended set of authors and papers using the committee authors' citation network as well as by sourcing authors who published similar papers, based on a prediction of relevance analogous to the phenomenon of triadic closure [11] observed in social networks. During the expansion, topical relevance represented as a set of user feedback on papers is used to constrain the expansion. ③ The curated and maintained committee is applied to enrich relevance signals in subsequent author and paper recommendations. In the figure, no-index circles represent authors.**

of a larger corpus, and may fail to target a group of authors that builds on each other's work, which would provide valuable context about research contributions. Second, these systems do not let users save authors they have discovered and know to be relevant, which limits the system's ability to help users make connections to new papers and authors through these known authors over time. Without saving functionality, systems can only recommend authors that are predicted to be relevant. And because users may have limited interest in or knowledge of these recommended authors, they may be confused or discouraged from using the recommended authors as rich mechanisms for discovering and explaining the relevance of additional papers and authors [20].

Here we propose ComLittee (Figure 1), an author-augmented literature discovery system that promotes rich author-centric interaction. In order to inform the design of ComLittee, we formulated a design space for author-augmented literature discovery (Table 1), in which ComLittee covers new ground along several dimensions. In terms of the workflow, users on ComLittee can 1) select a topic and initialize the system with personally curated seed papers; 2) receive author recommendations produced from a larger set of recommendation sources compared to prior work, using a new approach that combines paper recommender scores and publication network relations with previously-saved authors; 3) save (elect) relevant authors over time to a personal committee, which iteratively updates the system to inform future exploration and enable rich explanations of the relevance of both the recommended author and the papers contained in the recommended author card. In addition, ComLittee renders the relevance explanations as interactive filters useful for quickly homing in on specific papers that contribute to the relevance between authors. We instantiate these interaction and the author-centric workflow features as a list-based discovery system, enabling comparison to related systems [20, 23], and evaluate it in a controlled laboratory study to uncover their feasibility, value, and implications for future design.

In summary, this work makes the following contributions:

- We present a design space for author-augmented literature discovery, with seven interaction and presentation primitives, to situate the current work within the literature and to inform future designs in this emerging space.
- We propose ComLittee, a novel interactive author-centric system for author-augmented literature discovery.
- We evaluate ComLittee in a within-subjects study ($N = 16$), and demonstrate its value over a strong paper-centric baseline based on a prior system [23]. Through detailed quantitative, behavioral, and qualitative analyses, we report how ComLittee led to gains in discovery efficiency (for both authors and papers), novelty (for authors), and interestingness (for papers). In addition, we provide implications for design for future systems in author-augmented literature discovery.

## 2 RELATED WORK

Below we review existing paper-centric systems (Section 2.1); how theoretical and empirical studies of expertise-finding systems suggest the importance of author modality in the literature discovery process (Section 2.2); selected work from interactive machine learning, which guides our research questions (Section 2.3); and various examples of relevance explanation, which guide our system design (Section 2.4).

### 2.1 Paper-centric Literature Discovery Systems

To help scientists and professionals [35] in various stages of the literature discovery process, significant research effort has been

**Table 1: The design space for author-augmented literature discovery. We situate ComLittee and the closest prior work in terms of seven design axes as follows (see text for descriptions). Who: Curated or Inferred. What: Paper (P) or Author (A). When: Sequential (S) or Batched (B) effects. Where: Push or Highlight. Why: Explanation (Exp) or Recommendation (Rec). How (Relevance Source): Paper recommender score (R); Co-author relation (Co); Cited author relation (Ci); Prior user interaction history (I); or Other mechanism (Misc.). How (Relevance Distance): Direct (D) or Indirect (I).**

| | Who | What | When | Where | Why | Relevance Source | Relevance Distance |
|---|---|---|---|---|---|---|---|
| ComLittee | Curated | P,A | S | Push | Exp | R,Co,Ci | I |
| FeedLens [23] | Inferred | P,A | S | Highlight | Rec/Exp | R | D |
| Kang et al. [20] | Inferred | P | B | Highlight | Rec/Exp | Co,Ci,I | D,I |
| Bridger [43] | Inferred | A | – | Push | – | Misc. | D |
| Baseline | Curated | P | S | Highlight | Rec/Exp | R | D |

devoted to developing interactive systems. However, most prior systems have focused on *documents* (i.e., research papers) as the core primitive for both user interactions and discovery, and do not support exploring social signals (i.e., authors and their relations to other authors) around the documents. For example, Papers101 [8] helps scholars search for additional relevant literature by generating unused keywords for query expansion. Kang et al. showed that enabling search for analogous papers based on the purpose-mechanism schema [22] or diverse domains [21] can increase creativity of scientists' ideas. Once a set of papers has been discovered, systems can support subsequent tasks; *e.g.,* PaperQuest [42] with triaging which papers to read next, LitSense [49] with overviews and filtering of a collection of searched papers, CiteSense [57] with appraisal and grouping papers, Threddy [19] with organizing papers into notable threads of research while reading, and Wang et al.'s system [55] with visualizing collected papers in a broader narrative structure. In addition, Passages [14] helps users collect text snippets while reading papers, which can be re-represented into a relational form (*e.g.,* matrix) later. Relatedly [39] helps users discover relevant paragraphs from papers, and CoNotate [38] and Interweave [40] help by expanding queries that promote 'active' searching [37]. Finally, several systems have been developed to help reduce the cognitive cost of reading papers and documents (*e.g.,* ScholarPhi [16], CiteRead [44], Scim [44], Fuse [27], Crystalline [31], and Wigglite [32]). Despite differences in use case scenarios, these systems share a commonality in design that centers *papers* as the mode of interaction. In ComLittee, we also focus on designing an interactive system that can help users explore and discover papers relevant to a topic of interest. In contrast to the above prior work, we treat *authors* as the mode of interaction and discovery, while at the same time allowing participants to actively explore relevant papers in the context of relevant authors.

## 2.2 Expert-finding Systems

A parallel line of research on expert and social recommendation has also been developed (see [13, 51] for a review). Systems such as ReferralWeb [24] and Expertise Recommender [34] contributed understanding of social conditions in which expert recommendations become useful, while Guy et al. [12] found microblogs as a valuable source of data for expertise-matching. A particularly relevant line of research explored authorship networks, such as Liben-Nowell and Kleinberg's work [30] on predicting future collaborations among scholars using co-authorship networks, and Conference Closure [54] which proposed a new form of triadic closure (*i.e.,* scholars who attend mutual conferences are more likely to form connections) and studied its effects on future collaborations. As a specific form of expert recommendation, Bridger [43] proposed algorithms that facilitate *author* recommendations and burst filter bubbles by allowing users to select facets of expertise from an author's prior publications, and matching on the selected factors while diversifying on others in recommendation. Theoretical work such as Burt's 'Structural Hole' theory [5] suggests the importance of author recommendation for brokering knowledge across different fields, and empirical evidence suggests an increasing importance in the face of deepening specialization of knowledge [46, 50]. Studies that examined the practice of literature discovery, such as Sandstorm's [47] and Pirolli's [41], showed how authors play a highly valuable role in the process. These studies also give inspirations to recent work by Kang et al. [20] and Kaur et al. [23] that leveraged authorship graphs to augment encountered papers with highlights of potentially-relevant authors in them, showing improved user engagement and discovery experience. Taken together, author recommendation systems and recent work that leverages authorship graphs suggest a burgeoning design space for future author-augmented literature discovery systems; hence, we use them to form the bases of our design space (Section 3).

## 2.3 Interactive Machine Learning

Though literature discovery can be modeled as users submitting a series of independent search queries and reviewing the retrieved results for each, complex user intent and evolving knowledge may be better served by a system that can interactively learn from user feedback throughout the course of discovery. Here, the field of interactive machine learning offers relevant examples and insights. A particularly relevant example is the Regroup system [2], which proposed a novel probabilistic approach that iteratively updates priors based on a user's feedback on group members as they curate them. A core insight from this work is that search-by-name and system-generated recommendations have complementary strengths, with the former effective for forming small, well-defined groups while the latter helpful for large, varied groups. Another insight comes from Kocielnik et al.'s study [25] that showed how a recall-oriented machine learning system objective improves user perception and

willingness to adopt over a precision-oriented objective at equal performance levels. This is also consistent with how users benefited from diverse author recommendations on Bridger [43]. The question is then, how can users effectively and continuously discover novel authors? Addressing this question requires studies that analyze user and system behavior over time to shed light on effective user navigation strategies in the context of changing alignment between human- and AI-model of relevance. Our study and collection of behavioral data in the course of user interaction was designed to contribute to this gap in the literature.

### 2.4 Relevance Explanation

One key component of interactive recommender systems is supporting users in making sense of the recommended items. For example, SearchLens [6] and FeedLens [23] adopted a *lens* metaphor and provided interactive at-a-glance explanations and relevance filters. Findings in RelevanceTurner [52] also showed the benefits of making recommendation sources more transparent in discovery tasks. One promising approach in prior work is to explain newly recommended items by drawing connections to previously discovered or familiar items. For example, Apolo used a *relative spatial layouts* design where users can iteratively explore parts of the citation graph around familiar papers [7], and Kang et al. showed benefits in providing personalized relevance explanations around emailed paper recommendations, which described connections between the user and the authors of the papers [20]. When designing ComLittee, we were inspired by the high-level ideas in the above prior work. Specifically, we adopted the "lens" metaphor at both the author and paper levels to surface ones that were most relevant. In addition, we also drew connections between a recommended author to the set of authors already familiar and saved by a user as a way to explain their relevance.

## 3 DESIGN SPACE FOR AUTHOR-AUGMENTED LITERATURE DISCOVERY

Recently, three prior works on literature discovery systems that incorporate author entities (described in Section 2.2) [20, 23, 43] have also explored various designs for author-augmented literature discovery. In order to situate our work and inform future work, we surveyed past design decisions and formulated a design space (Table 1) for author-augmented literature discovery. The prior work was contrasted to ComLittee and the baseline implementation, for contextualizing our evaluation and convenience of reference. The seven axes represent dimensions ranging from the scope of user controllability to how author recommendations were sourced. The axes were chosen not only for their coverage of salient features of system designs in the space, but also for their generative potential; though we demarcate categorical values along each axis in the space here, future work may use it to envision novel system designs that either instantiate new values along a specific axis, expand the space by contributing new axes, or by proposing new combinations of values along the axes.

- *Who (User controllability and agency in iterative steering of author recommendations: Curated vs. Inferred).* One of the salient design choices lies in the decision of how much user controllability the system supports in iterative steering of author recommendations, and conversely how much should be automated. On one end of the spectrum is designs without any direct curation support, which sacrifices users' steering capabilities in favor of the convenience of minimum required effort. Previous work [20] instantiated this design choice by inferring reference authors thought to be of interest (Indirect reference authors) from user publication and interaction logs, which were then used as proxies for the user to help identify relevant authors. In contrast, ComLittee supports manual curation by directly adding/removing indirect reference authors to/from an elected committee.
- *What (Explained Entity Types: Paper vs. Author).* The information algorithmically gleaned from how relevant authors were found (e.g., a paper of theirs was cited by a relevant author) may be featured as explanations on an author's papers (*e.g.,* when viewing on author page with publications), or in an aggregate form as an author summary, which was shown to improve user understanding and engagement [20]. FeedLens [23] provides paper- and author-level explanations using recommender scores. ComLittee extends this to incorporate explanations based on citation and coauthor relations, which show the relation of the papers to authors of interest.
- *When (Delay between user feedback and system changes: Sequential vs. Batched).* Another salient design choice lies in how quickly the system responds to user feedback. When steering author recommendations early on, users may need to see effects take place immediately upon their most recent feedback to form a mental model of the system, requiring sequential (immediate) rather than batched (slow) updates. However, enabling low-latency sequential updates may require sacrificing some degree of accuracy [23]. Furthermore, whether and how divergence between the most recent snapshot of user feedback and the overall feedback the system has accumulated over time may manifest in the literature discovery process remains an open empirical question.
- *Where (Whether authors are presented as top-level push recommendations vs. In-situ highlights within an encountered paper context).* A system can either 'push' authors as top-level recommendations, or highlight them *in situ* in the context of papers the user encounters. ComLittee and author recommenders (*e.g.,* Bridger [43]) do the former. Most literature discovery systems do neither, but recently some do the latter by highlighting authors in the context of papers the user encounters during exploratory search [23] or paper recommendations [20].
- *Why (The purpose of author highlights: explanatory vs. to recommend novel authors).* In-situ author highlights can either make you aware of authors you know, or highlight new authors you may wish to explore. The former can be used as an explanation of the relevance of a paper to you [20], whereas the latter is more useful as a launchpad for exploration [23]. Systems may not know which authors the user knows, so highlights may function as either type depending on user knowledge. ComLittee only shows in-situ explanations, so as not to confuse users by recommending authors in-situ and also as top-level entities.

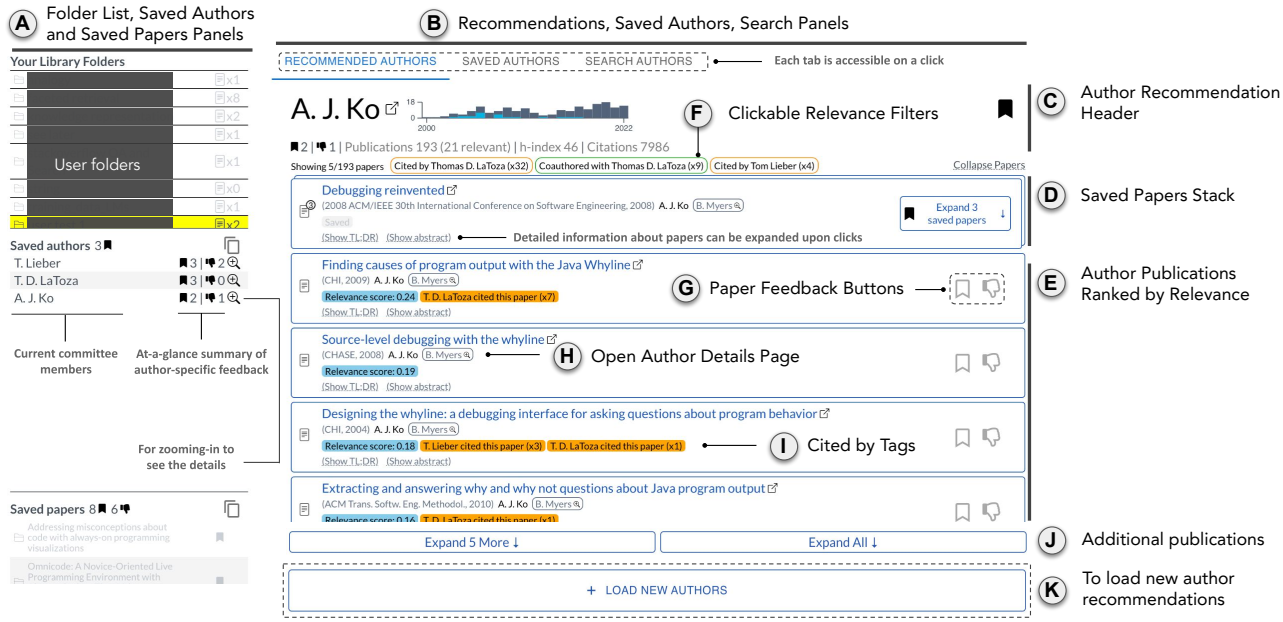



**Figure 2: ComLittee features: Ⓐ topic folders, saved papers and authors (*i.e.*, currently saved committee members) panels, Ⓑ the primary recommendations tab and direct search-by-name of authors (not shown), Ⓒ – Ⓙ various features for interacting with author recommendations (see text), Ⓚ button for generating new recommendations with updated user feedback.**

- *How: Relevance Source via recommender (R) vs. author (Co, Ci) vs. user (I).* Author relevance may be determined by scoring relevance of their papers using a paper recommender algorithm [23], by identifying their citation (Ci) or co-author relations (Co) to other (reference) authors of interest [20], or by identifying authors based on the user's interaction history (I) such as how many times the user has saved their papers to their library [20]. ComLittee determines relevance using all three of these categories, and additionally explores leveraging a new aspect of the user's interaction history—the authors they have saved to their committee. In contrast, Bridger [43] makes author recommendations using faceted content mined from the text of each author's publications (denoted as Misc. in Table 1) rather than traversal on authorship and citation graphs.
- *How: Relevance Distance via Direct vs. Indirect reference author.* The reference authors of interest from which relevant co-authors and cited authors are determined (see Relevance Source above) may either be the user (if they are an author), or a different author of interest. Kang et al. [20] terms the former "Direct" relations and the latter "Indirect" relations (expressing the further, indirect relation to the user in that case, via a proxy author of interest). Most author recommendation systems like Bridger [43] focus on direct relations to the user. Kang et al. [20] explore both; however, indirect authors are inferred, which can be confusing to the user when they are not relevant or the user does not know about them. ComLittee overcomes these issues by allowing users to specify indirect authors to the system. ComLittee does not currently implement direct author relations since our experimental task is focused on finding entities related to a topic rather than the user in general.

## 4 SYSTEM DESCRIPTION

ComLittee uses various sources of signals to identify and provide both author and paper recommendations. To start, we assume users can find a small set (i.e., $\geq 5$) of papers relevant to a topic to initialize the underlying paper recommender (detailed below). Though not strictly required, we also assume users might already know some authors relevant to the topic that may constitute their initial "committee." The system then helps users discover other relevant authors and their papers, while surfacing connections between familiar committee authors and new authors.

### 4.1 Example User Scenario

A research scientist wants to learn more about a new research area that she recently started exploring. She uses a scholarly search engine and an online paper recommender service to discover recent papers based on a few papers that she had saved for the topic. However, she was also interested in learning more about who the relevant authors in the field were, which would allow her to reach out to them for potential collaborations. From reading papers, she notices a few with relevant recent publications that she liked, but neither a paper recommender nor search engines can support her to proactively explore more relevant authors.

Feeling frustrated she switches to ComLittee and imports the relevant papers in her library folder to the system. She can immediately see a list of authors relevant to the topic, some of whom were familiar to her from recent reading. Recognizing them, she quickly clicks the bookmark buttons next to their names (Fig. 2c) to save them to her "committee" for the topic. In addition, she also learns that some of the familiar authors had written relevant papers she did not know about, which were ranked to the top of their publications page based on the relevance to the folder (Fig. 2e).

Using the Paper Feedback Buttons (Fig. 2g) she can save additional relevant papers to her folder, or downvote irrelevant papers to help ComLittee better understand her interests.

Later she clicks on the Load New Authors button (Fig. 2k) to get a new batch of recommendations. Now she notices ComLittee starting to recommend authors that she does not recognize. In their author cards, she can see their connections to authors she already knew. For example, one unfamiliar author was frequently cited by several familiar authors, and she could see multiple interesting and relevant publications in the paper list. Feeling confident, she saves this new author to her folder. As she saves more, ComLittee also continues to find more relevant authors, improves its explanations about them based on committee authors, and ranks their papers with better accuracy.

## 4.2 Author-Centric Recommendation Strategies

To support the features described above, ComLittee introduces the following four strategies that *expands* from papers and authors already familiar to the users to recommend unknown authors and papers for discovery.

*4.2.1 Library-extracted.* The user may start using ComLittee with only few saved papers in a library folder. The user may be familiar with some of the papers, their concepts, and/or authors, which prior work showed could be effective for boosting user engagement in the email paper recommendation alert context [20].

**Procedure.** Given a few saved papers, authors of the papers are tallied and sorted in a descending order of frequency.

*4.2.2 Authored multiple relevant new papers.* Based on papers saved in the user's folder or papers they had downvoted (Fig. 2g), ComLittee searches an index of recent publications for 100 papers that were the most topically relevant (within the last 180 days, as commonly used in public paper recommenders). These 100 papers then cast 100 'votes' on their authors. The votes on authors are tallied and sorted in a descending order of counts.

**Implementation of the relevance prediction model.** Similar to [23], we use an ensemble regression model that scores each paper on a scale of [-1, 1], where a negative score represents predicted irrelevance and a positive score represents predicted relevance (stronger towards both ends of the scale). The model averages the outputs of two linear Support Vector Machines (SVMs) that differ in terms of its training procedure and specifically how each paper is represented as a feature vector: the first SVM uses textual features (unigrams and bigrams) with Tf-Idf normalization, similar to the public arxiv-sanity recommender (https://arxiv-sanity-lite.com), and returns high-precision direct matching on terms within search queries. The second SVM uses SPECTER embeddings, with a focus on matching on the multifacted semantic relatedness beyond term-matching of the first SVM, and showed good performance on paper recommendation tasks [10]. The ensemble of the two therefore balances the strength of each matching model and is iteratively re-trained in real-time, based on each user's feedback on paper recommendations. We used the Semantic Scholar Academic Graph API to access extracted author names from their publications. We refer to [48] for documentation on the quality of the name extraction pipeline for interested readers.

*4.2.3 Coauthorship-based expansion.* Users may also find new authors who co-authored with familiar authors they trust for a topic. In such cases, the familiar authors can be viewed as mediating wedge-shape paths between the user and each new author on the publication graph. Recent work [20] also showed highlighting author names in paper recommendations based on the same mechanism of triadic closure [11, 26] on citation graphs increases user engagement.

**Procedure.** Starting with user-saved authors, each of the saved author's list of publications is searched and assigned a relevance score, using the same prediction model described above. When the user has already provided feedback on a paper (*e.g.,* when she received one of the coauthors of the paper as a recommended author earlier and encountered the papers), we overwrite the relevance score with either 1 (*i.e.,* the user has saved this paper earlier) or -1 (*i.e.,* downvoted), in order to treat user feedback as ground truth. Then relevant papers are filtered (*i.e.,* have scores $> 0$). Finally, using a similar voting procedure from papers to their authors as before, authors with the highest votes are collected.

*4.2.4 Citation-based expansion.* Another way trust can be propagated between familiar and unfamiliar authors is through citations in their papers. The assumption here is scholars cite papers they trust in their own papers, so that users may find value in discovering unfamiliar authors through papers frequently cited by trusted and familiar authors.

**Procedure.** Step 1) Using each user-saved author's publications, we collect up to 100 most relevant papers based on their scores. Step 2) For the collective referenced papers from the papers in Step 1, we design a voting procedure in which each saved author casts 1 vote on a reference if the author has at least one paper that is a) included in the sampled relevant papers in Step 1, and b) cites the reference. We exclude self citations for diversification. We assign votes at the author level rather than the paper level to prevent authors with many publications from dominating the votes. Step 3) Sample the top 100 references with highest votes (higher vote counts means more of the saved authors have previously cited that work). Step 4) Using the references from Step 3, repeat a similar paper-to-authors voting procedure as earlier, and finally collect the authors with highest votes over the references. See Appendix B for pseudo-code implementation.

*4.2.5 Batch generation.* Each of the four strategies above generates a ranked list of author recommendations. Initially, the top two recommendations from each list are selected and interleaved (Appendix A) to create a *batch* of eight author recommendations. A cursor is then moved to point to the next top ranked recommendation in each list. When users clicks d on the "Load New Authors" button (Fig. 2k), a new batch of recommendations is generated using the cursors. Whenever the user saves new authors or papers, or down-votes a paper, the lists replenish and cursors are reset to the top. ComLittee shows all available explanations for each author, regardless of the strategy it was selected from. For example, if an author was selected because it was frequently cited by familiar authors, but also coauthored with a committee author before, both

'cited by' and 'coauthored' explanations are shown. Relevance explanations of current author recommendations within each batch are updated as the user interacts with ComLittee.

### 4.3 Relevance explanation features

The author-level explanations above are translated into interactive relevance explanation filters (Fig. 2f) in each author recommendation header. An exception is the information about the number of predicted relevant papers the recommended author published which is displayed simply as a static text tag next to the number of total publications (a filter is not needed because papers are sorted by relevance scores by default). Relevance filters also interact with a small publication year-count histogram next to each author's name; clicked filter adds overlays that correspond to the count of papers included in the relevance relation (*e.g.,* A "Coauthored with... (x9)" filter will bin the 9 papers into published years and add corresponding visual marks – bars – to the vis upon a click). ComLittee also features paper-level relevance explanations which show the predicted relevance score for each paper at the time of recommendation, and up to three authors who have cited the paper most often, with the number of their papers that cited it, while excluding self citations.

### 4.4 Design Iterations

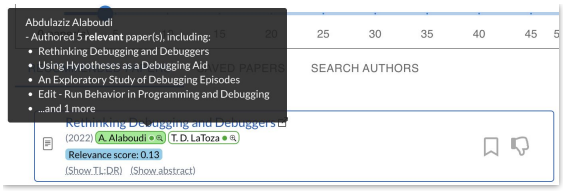
Our design team involved a senior UI designer familiar with search interfaces and literature support tools who provided feedback on the usability and clarity of our system design through the iterations. We also ran three rounds of pilots to seek design feedback and iterate before running the evaluation study. The iterations sought to improve clarity around the main confusion points discovered from pilots (described below) and usability (*e.g.,* adding a central state and in-line action indicators for loading latencies). See Appendix E for description descriptions and design rationales.

### 4.5 Final System Interface

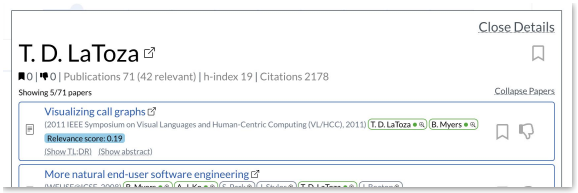
The final interface is shown in Fig. 2. Ⓐ & Ⓑ: Saved authors and papers are shown in the corresponding panels on the left-hand side, to increase user's context awareness and allow them to track their progress over time. Ⓒ In each author recommendation header, pertinent information about the author such as their name[1], the number of their papers the user has saved or downvoted, the number of total publications and estimated relevant ones, their h-index, and the number of citations. Next to the author's name is a small histogram visualization of the author's publication records over time, with defeault blue overlay bars in the vis showing the number of papers predicted relevant by the system at the time of recommendation over the years. This visualization updates when the user clicks on a relevance explanation 'pill' available under the author name (Ⓕ), by adding the corresponding counts of papers for the filter as an additional overlay using the same color. Clicking on a filter also filters the corresponding papers in the author's publications list (Ⓔ), sorted in a descending order of the predicted relevance score (default). Ⓓ A stack of papers the user has already provided feedback on appears as a stack at the top. The user can provide feedback on each paper to save or downvote it, which updates the backend while adding the title of the corresponding paper to the saved papers panel (Ⓖ). In each paper recommendation, users can click on any author name to open an author details modal (not shown) that looks exactly like the author recommendation cards in the main tab (Ⓗ). Users can view who among the saved authors cited each author's papers (Ⓘ). In particular, this information 'bubbles up' to the recommendation header (if the citing author is not featured in author-level 'cited by' filters). By default only 5 papers for an author are shown to prevent overload, but is expandable (Ⓙ). Finally, users can click on the 'Load More Authors' button at the bottom to receive a new batch of recommendations (Ⓚ). Saved author names are highlighted in green in a paper recommendation context (*i.e.,* they serve as *explanation* author highlights in contrast to *recommendation* highlights in the baseline).

[1]We collapse the first name to reduce subconscious focus on presumed gender of the author. However, users are instructed to mouse-over to see the full name of the author or click on the author's name to see additional details of that author on the corresponding author details page on Semantic Scholar.

### 4.6 Baseline




**(a) Mousing over an author name reveals a list of paper titles published by the author, predicted as relevant by the system.**




**(b) The author details page modal is accessible by clicking on an author name in any paper recommendation.**

**Figure 3: Salient Baseline interface features.**

The baseline and ComLittee interfaces differed in the organization of recommendations. The top-level recommendations in the baseline featured a list of paper recommendations, with an additional mouseover interaction for author names in each paper to reveal predicted relevant papers published by that author in a tooltip when clicked (Fig. 3a). Authors with a high number of relevant papers (the threshold for highlighting was adjustable via a slider at the top) featured a green dot (FeedLens [23]) next to their names with a highlighted border. The baseline system instantiated a FeedLens mechanism in which it memorized the paper lists the user encountered over time (*e.g.,* an author's publications when her detail's page is opened or the author is directly searched by name; when new paper recommendations are added to the main tab), while also scoring each paper in a list **using the same relevance prediction model implemented for ComLittee**, to update the author highlights in real-time. Both interfaces supported click interaction on author names for opening a modal view of author details, including a ranked list of author publications using predicted relevance

scores. In ComLittee author details pages featured more relevance explanation features (Fig. 3b).

# 5 EXPERIMENTAL DESIGN

## 5.1 Objective & Research Questions

Our goal in the evaluation was to study how ComLitteeand the author-centric interactions it instantiates benefit scholars wanting to discover relevant and interesting authors and papers in a personalized domain. Our research questions focused in part on the efficiency and quality aspects of scholars' literature discovery experience, for two modalities of discovery (*i.e.,* authors and papers). We operationalized the *discovery efficiency* construct as the aggregated quantity of saved authors (or papers) for a fixed amount of time, *quality* as the average post-task ratings of either relevance or interestingness of discovered items, and *average discovered author novelty* as the ratio between the number of unfamiliar-yet-relevant authors to known-and-relevant authors. Concretely our research questions were:

- RQ1) Does ComLittee improve the efficiency and quality of scholars' author discovery over the baseline?
- RQ2) Can ComLittee users save known-and-relevant authors and discover unfamiliar-yet-relevant authors?
- RQ3) Comparing to a paper-centric baseline, does ComLittee inhibit paper discovery? and
- RQ4) How do ComLittee users engage in paper discovery, and specifically does effort of discovery increase?

## 5.2 Participants

We recruited 16 participants (8 female) for the study. The mean age of participants was 28.3 (SD: 4.32) and all actively conducted research at the time of the study (1 Post-doc, 13 PhD students, 2 Pre-doctoral Investigators). Participants' fields of studies included (multiple choices): HCI (9), NLP (6), Information Retrieval (2), Neuroscience (1), Oncology (1).

## 5.3 Procedure

*5.3.1 Structure.* We employed a within-subjects study to compare ComLittee to a baseline system (see above for details of implementation). We asked each participant to choose two different research topics they wished to explore, and randomly assigned systems to the topics for timed exploration tasks. We counterbalanced the order of presentation using 8 Latin Square blocks and randomized rows. Participants followed the following procedure in the study, which took place remotely using Zoom (Fig. 4): Introduction, Consent, Demographics survey and curation of topic folders for the main tasks; Tutorial of the first system, Main task for the first system, Post-task rating, and Survey; Repeat for the second system; Debrief. In the topic folder curation, the interviewer guided each participant to navigate to a popular online scholarly search engine to create two topic folders, one per participant's research topic. At this stage, participants were given time to freely search for two research papers that they thought represented each topic, and instructed to save them into the folders. They were asked to share their screen and think-aloud the main timed tasks. The study lasted for 1.5 hours and participants were compensated $45 USD. The study received Internal Review Board approval.

*5.3.2 Tutorials.* Before participants start with each of the two main task with different conditions, they were given a tutorial of the assigned systems via screen sharing. The interviewer demonstrated the main features of each system based on a prepared script that took around 5 minutes for the baseline condition and 10 minutes for the system condition that had more features. Participants were then instructed to save as many relevant and interesting authors and papers as possible during each task, and were recommended to save at least 5 in each category. Aside from it, they were also told to downvote 3 or more irrelevant papers early on to calibrate the recommender system.

*5.3.3 Timed Main Tasks (15 mins each).* The main tasks used the two different topics that participants chose as personally motivating for discovering new papers and authors in. We randomly assigned each topic to a condition. Each system used the two seed papers participants curated for each topic to generate the initial set of recommendations.

*5.3.4 Post-task Ratings and Surveys.* After each task, participants clicked on a button in the interface to copy a random subsample of their saved authors and papers (up to 15, respectively, so that rating did not take overly long for any participant) and pasted this copied content onto a Google Spreadsheet that the interviewer shared with them. In the spreadsheet were three questions for each saved author and three questions for each saved paper. The first question for each author was a binary yes/no question ("*Were you familiar with the author before the experience?*") and the last two questions were 7-point Likert scale questions ("*I found this author to be relevant.*" and "*I found this author to be interesting.*"). For each paper, similar questions of familiarity, relevance, and interestingness were followed.

In the survey administered after each task, participants were asked about their subjective feelings related to the experience. For demand (including physical and cognitive) and overall performance we adopted the validated 6-item NASA-TLX scale [15], with the original 21-point scale mapped to a compact 7-point scale [45]. For technological compatibility with participants' existing discovery workflows and the ease of learning we adapted the Technology Acceptance Model survey from [56] (4 items). We also included additional questions to measure participants' subjective feelings of the system's effectiveness in supporting author (4 items) and paper discovery (3 items). Finally, we included additional questions for common (4 items) and condition-specific features (2 items in the baseline condition, 6 items in the treatment condition) of each system to measure their effectiveness (See Appendix F for details of the questionnaire).

*5.3.5 Data Collection.* During each participant's interaction with each system, we collected their behavioral traces *i.e.,* timestamped actions and their details. When a participant provided feedback on a paper, the unique paper identifier, its estimated relevance score from the recommender system at the time of feedback, and the context in which it appeared (*i.e.,* whether on an author detail's page) was stored. For each participant's feedback on an author, the unique author identifier, existing relations to saved authors at the time of feedback (treatment only), and in case of an author recommended in the main recommendations tab, which sourcing mechanism was


**Figure 4: The entire procedure of our study. The order of the middle section of the procedure was swapped based on the assignment (A/B). This order assignment was randomized and counterbalanced across participants (see text).**

used (treatment only), and whether the author was directly searched was stored. We filtered the behavioral traces based on modality (*i.e.,* authors or papers) and transformed values (*e.g.,* average counts of saved authors; ratio between saved-to-downvoted papers) for analysis. Participants' think-alouds during the tasks, open-ended questions, and debrief were recorded and later transcribed.

*5.3.6 Data Analysis.* The mappings between analyses of collected data and research questions are as follows.

- RQ1) We analyzed the efficiency and quality measures of discovered authors between conditions using the paired Student's t-test. We analyzed Likert items using non-parametric tests such as the paired-samples Wilcoxon's signed rank test (for paired-samples data such as participants' responses to survey questions) and the Mann-Whitney U test (for independent data such as judgment on saved authors or papers).
- RQ2) We analyzed the ratios between the number of discovered authors who were unfamiliar-yet-relevant ('unfamiliar') to the number of known-and-relevant ('known') authors. A mean novelty value $\in [-1, 1]$ was computed for each participant in each combination of experimental factors by averaging 'unfamiliar' $\mapsto 1$ and 'known' $\mapsto -1$ over saved authors, such that a value closer to 1 meant more unfamiliar authors were discovered for a unit number of known authors, and vice versa. We ran a one-way Repeated Measures (RM) ANOVA test with the experimental condition as a two-level factor (*i.e.,* ComLittee vs. baseline). RM ANOVA was chosen over regular ANOVA for its advantage in controlling for the random effect from subjects in the within-subjects experimental design. We tested the assumption of sphericity using Mauchly's test [33] and ran post-hoc Tukey's HSD comparisons to identify significant pairwise differences.
- RQ3) We analyzed the efficiency and quality measures of discovered papers, similarly with RQ1.
- RQ4) We analyzed both quantitative and qualitative data. The quantitative analyses included the average estimated model relevance scores of papers at the time of saving (the predicted relevance score on each paper ranged between $[-1, 1]$, where a positive score corresponded to relevance and vice versa, with higher significance towards both ends. This score represented how the recommender predicted the paper to be relevant, given all of the user's feedback on papers up to that point. The model for calculating the scores was held constant between the two conditions to control for analysis of the trends in user steering), the average number of papers saved for each discovered author, the balance of two steering operations performed on paper recommendations (*i.e.,* the mean ratio between the number of saving-to-downvoting was similarly calculated with the mean author novelty described above, by mapping 'save' $\mapsto 1$; 'downvote' $\mapsto -1$, where a value closer to 1 represented a low net steering effort by a user for each saved paper and vice versa). We analyzed the average number of papers saved or downvoted for each author in each condition, using a two-way ANOVA (two two-level factors as experimental conditions and feedback type) followed by post-hoc Tukey's HSD tests; for analyses involving time progression, we ran RM ANOVAs as before. We checked the suitability of ANOVA by examining the homogeneity of variances in factor groups using Levene's test [29]. For qualitative analysis two authors analyzed transcripts through open coding, then discussed and merged main themes appeared from it.

# 6 FINDINGS

## 6.1 RQ1. ComLittee increased author discovery efficiency without decreasing quality

*6.1.1 Users saved more authors in ComLittee and found saved familiar authors interesting.* Users saved significantly more authors overall in ComLittee (M=10.6, SD=3.88) than in the baseline condition (at the $\alpha = .001$ level after correcting for multiple tests using the Bonferroni procedure, M=6.6, SD=2.71, $t_{\text{paired}}(26.80)$=-4.72, $p$=0.0003); Fig. 5a). Between the familiar vs. unfamiliar authors who were saved, the distribution skewed towards familiar authors in the baseline condition, while a similar skew was not observed for the treatment condition ($\chi^2(1)$=10.86, $p$=0.001). This difference in distribution was reflected in the results of paired t-tests between the two conditions, with a significantly higher number of unfamiliar authors being saved in the treatment condition (M=4.6, SD=2.66; Baseline: M=1.7, SD=2.33, $t_{\text{paired}}(29.50)$=-2.98, $p$=0.009, Fig. 5b), while the number of familiar authors saved in each condition did not differ significantly (Treatment: M=5.4, SD=2.71; Baseline: M=5.1, SD=3.23, $t_{\text{paired}}(29.10)$=0.33, $p$=0.75, Fig. 5c), using a random subsample of authors whose familiarity was rated by users. Directed search (*i.e.,* users typed in known author names) was significantly more common in the baseline than the treatment condition, and this was consistent with how users saved significantly more unfamiliar authors in ComLittee where browsing and serendipitously discovering authors by clicking on their names in recommended papers was common (Fig. 5e).

In terms of the quality of discovery, measured as saved authors' interestingness and relevance, both treatment (76%=117/154) and baseline (84%=79/94) conditions resulted in majority High interestingness (See Fig. 11 in Appendix D. for the aggregate response count distribution). In addition, the overall distribution between High vs. Low interestingness authors did not differ significantly between the two conditions ($\chi^2(1)$=1.83, $p$=0.176). However, we saw a marginally significant difference in interestingness between the authors whom users were familiar with prior to the task in the treatment condition (M=6.3, SD=0.93) and the baseline condition

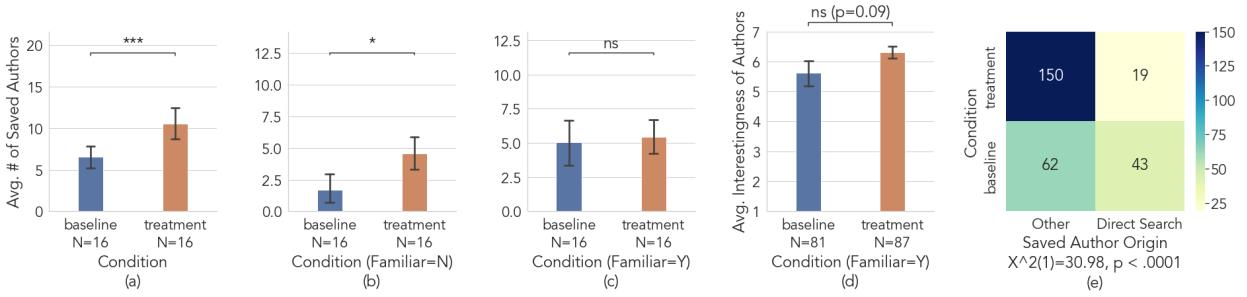



**Figure 5: Users' author discovery outcomes differed significantly between the conditions. (a) Users saved significantly more authors in ComLittee than the baseline (the three tests in a–c were Bonferroni-corrected for multiple testing). (b & c) The average number of unfamiliar authors saved was significantly higher in ComLittee, whereas the average number of familiar authors saved did not differ between the two conditions. (d) The difference of average interestingness was marginally significantly higher for authors that users were familiar with prior to the task but not for unfamiliar authors. (e) Users in the baseline used direct search with author names significantly more, reflecting the primary means for finding familiar authors.**

(M=5.6, SD=1.97. Mann-Whitney $U$=3135, $p$=0.089, Fig. 5d), while no such difference was observed among the unfamiliar authors (Mann-Whitney $U$=1088, $p$=0.80). We return to this difference in Section 6.4.4, RQ4. The average relevance of saved authors did not differ significantly between the two conditions (ComLittee: M=6.4; baseline: M=6.4, $t_{\text{two-tailed}}(161.93)$=0.19, $p$=0.85).

*6.1.2 Users found ComLittee features helpful for discovery.* The survey results corroborated these performance gains in ComLittee. The workload required to complete the task (measured via NASA-TLX) was significantly reduced in ComLittee (for ComLittee, M=14.1, SD=5.56; for baseline M=17.0, SD=6.01, Wilcoxon $W$=13.5, $p$=0.01 ). Users also responded that ComLittee better supported (a) author discovery: 'helped me find relevant authors', M=6.1 (ComLittee) vs. M=3.9 (baseline), Wilcoxon $W$=2.5, $p$=0.002 ; 'helped me make sense of author's research', M=4.8 (ComLittee) vs. M=3.5 (baseline), Wilcoxon $W$=0.0, $p$=0.008 ; 'made me curious about author's research', M=6.1 (ComLittee) vs. M=4.3 (baseline), Wilcoxon $W$=3.5, $p$=0.001 ; 'explanations of relevant authors became more helpful the more I used the system', M=4.9 (ComLittee) vs. M=3.9 (baseline), Wilcoxon $W$=11.0, $p$=0.02 and (b) paper discovery: 'helped me find relevant papers', M=6.1 (ComLittee) vs. M=3.9 (baseline), Wilcoxon $W$=6.0, $p$=0.002 ; 'made me curious about the papers I found', M=6.2 (ComLittee) vs. M=4.8 (baseline), Wilcoxon $W$=3.5, $p$=0.01 ; 'explanations of relevant papers became more helpful the more I used the system', M=4.6 (ComLittee) vs. M=3.6 (baseline), Wilcoxon $W$=21.5, $p$=0.05 (see Table 3 for details). Consistent with the perception of helpfulness, users favored ComLittee in terms of the overall technology compatibility with their existing scholarly discovery workflows (for ComLittee, M=22.6, SD=2.99; for baseline, M=19.4, SD=5.08, Wilcoxon $W$=24.0, $p$=0.02 ) and the plausibility of future adoption (for ComLittee, M=6.2, SD=0.75; for baseline M=4.9, SD=1.50, Wilcoxon $W$=4.0, $p$=0.003 , see Table 2 for details).

## 6.2 RQ2. ComLittee users saved familiar authors to scaffold subsequent discovery of unfamiliar authors

Users' *a priori* familiarity judgment on a random subsample of the saved authors showed that users in both conditions started with twice as many familiar authors as unfamiliar authors in the first half

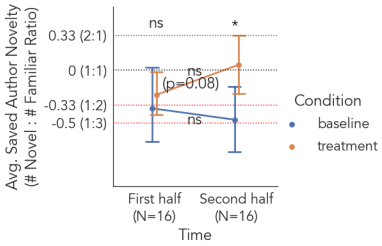



**Figure 6: Though users in both conditions started with ~2× more familiar authors to novel authors in the 1st half of the task, users in the treatment condition saved significantly more novel authors in the 2nd half of the task, reaching familiar:novel parity.**

of the task, but in ComLittee users were finding more unfamiliar authors later on, nearing the parity between the number of familiar to unfamiliar authors in the second half (Fig. 6; As a measure of robustness, we examined possible variations in the number of saved authors, and found that they did not change significantly between the first and second half of the session; $\chi^2(1)$=0.03, $p$=0.86).The result of RM ANOVA showed a marginally significant overall effect $F(1, 15) = 4.23, p = 0.057, \eta^2 = .22$. Additional Tukey's pairwise HSD comparisons revealed that in the second half of the experiment users in ComLittee had a significantly higher ratio of novel authors saved than the baseline ($T = -2.66$, $p = 0.01$, Cohen's $d = -.94$, partial $\eta^2 = .18$). Time progression in ComLittee had a marginally significant ($T = -1.06$, $p = 0.08$) positive effect (Cohen's $d = .64$, partial $\eta^2 = .09$), but not in the baseline condition ($T = .17$, $p = 0.87$). In terms of the sources of saved authors, all 3 mechanisms of recommendation seemed equally represented in the origins of these saved authors (a two-way ANOVA analysis showed no significant main effect from the recommendation mechanism type $F(2, 90) = 1.25, p = .29$, nor a significant interaction effect with the familiarity of saved authors $F(2, 90) = 1.29, p = .28$).

## 6.3 RQ3. ComLittee users saved more papers and rated saved papers as more interesting

*6.3.1 Efficiency.* On average, users saved significantly more papers in ComLittee (M=25.5, SD=13.55) than in the baseline condition

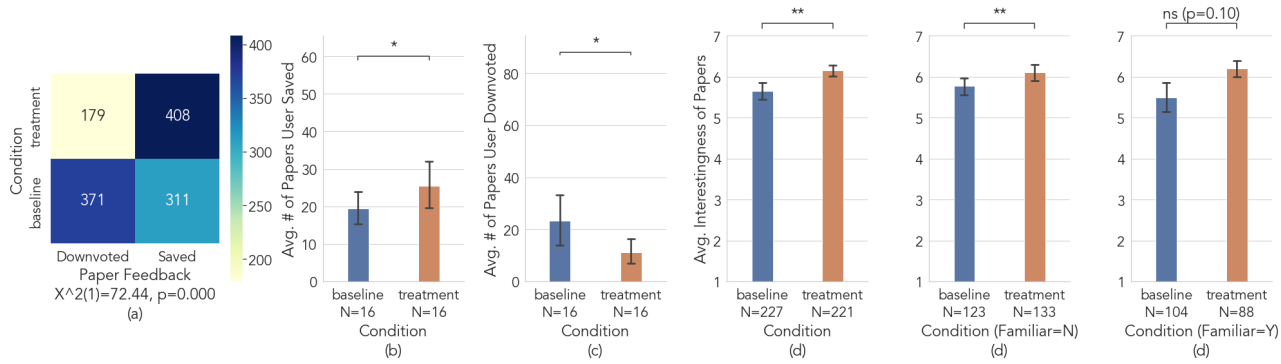



**Figure 7: Comparisons of user actions on and perceptions of paper recommendations between conditions. (a) The distribution of the number of saved vs. downvoted papers differed significantly, (b) with a skew towards saved papers in ComLittee (c) while more downvoted papers in the baseline. (d) The average post-task interestingness response on saved papers was significantly higher in ComLittee, (f) but specifically for papers that werenfamiliar, while the average familiar paper interestingness did not differ significantly. Analyses in (d–f) were Bonferroni-corrected for multiple testing (*i.e.,* three tests).**

(M=19.4, SD=9.59, $t_{\text{paired}}(27.01)$=-2.62, $p$=0.02; Fig. 7b). Users in the baseline condition downvoted more papers (Treatment: M=11.2, SD=10.05; Baseline: M=23.2, SD=0.56, $t_{\text{paired}}(21.78) = -2.17, p = 0.05$, Fig. 7c), indicating an improved efficiency. The distribution of saved-to-downvoted papers differed significantly, $\chi^2(1)$=72.44, $p$=0.00, Fig. 7a). Furthermore, among the random sample of these papers rated by users post-task, participants in both conditions saved a similar number of familiar papers (Treatment: M=5.5, SD=3.43; Baseline: M=6.5, SD=3.97) and unfamiliar papers (Treatment: M=8.3, SD=3.40; Baseline: M=7.7, SD=3.70).

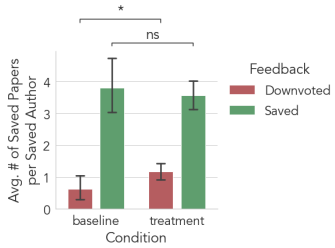
*6.3.2 Quality.* The average (7-point Likert) interestingness response on saved papers was significantly higher in ComLittee (M=6.1, SD=1.07) than in the baseline condition (M=5.6, SD=1.54, $t_{\text{two-tailed}}$ (403.00)= 4.02, Fig. 7d). To further investigate the distributional differences in interestingness judgment, we coded the response options 6 and 7 on the Likert scale as 'High' interestingness, and the response options 4 and 5 as 'Low' interestingness (they corresponded to *moderate-to-strong* and *neutral-to-slight* agreement levels, respectively. See Fig. 10 in Appendix D for count distribution). The resulting 2 (Interestingness) × 2 (Condition) matrix showed a significant skew towards High interestingness in both conditions, but with a higher degree in ComLittee (83% of rated papers in ComLittee were judged as High vs. 73% in baseline, $\chi^2(1)$=5.03, $p$=0.025). On average, familiar papers were judged marginally significantly more interesting in ComLittee (M=6.2, SD=0.91) than Baseline (M=5.5, SD=1.88, Mann-Whitney $U$=3954, $p$=0.10, Fig. 7f), while for unfamiliar papers the difference was significant (Treatment: M=6.1; Baseline: M=5.8, Mann-Whitney $U$=6622, $p$=0.009, Fig. 7e) which users judged as similar relevance (Treatment: M=5.7; Baseline: M=5.6, $p = 0.41$).

## 6.4 RQ4. ComLittee helped 'shortcutting' to more relevant papers, leading to efficiency gains and better human-AI alignment on relevance

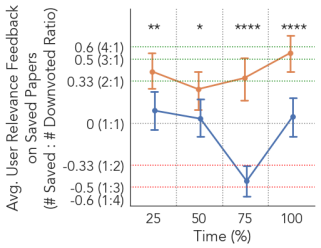
*6.4.1 Users saved multiple papers from each discovered author at once.* Users in both conditions visited an author details page to find over 3 relevant papers at once and save them (Treatment: M=3.6, SD=2.96; Baseline M=3.8, SD=4.44, Fig. 8a). The number of downvoted papers was significantly lower than the number of saved papers in both conditions (two-way ANOVA with Condition and Feedback Type as factors and the number of papers receiving feedback as a DV showed a significant main effect from Feedback Type: $F(1, 542) = 121.88, p < .0001$ but not from Condition ($p = .53$), nor from their interaction ($p = .12$)). The effect size of Feedback Type was Cohen's $d = -.98$, partial $\eta^2 = .20$ (in ComLittee; $T = -9.05$, Tukey's $p = 0.001$) and $d = -.93$, $\eta^2 = .18$ (Baseline; $T = -6.69, p = 0.001$). The result indicated the primary driver for navigating to a specific author's page to be finding relevant papers. However, the average number of papers downvoted in the treatment condition (M=1.2, SD=1.77) was significantly higher than in the baseline condition (M=0.6, SD=1.96, Tukey's $p = 0.02$), suggesting that navigation to each author's publications was more motivated and contextualized in ComLittee to also recognize which threads of research were not relevant to the topic.

*6.4.2 ComLittee users expended significantly less effort during paper discovery.* How much effort users expended during their discovery was evident in how much negative feedback they had to provide for the same number of saved papers, in order to steer the system according to their changing notion of relevance. Compared to ComLittee, baseline users provided much more negative feedback for a given number of saved papers (Fig. 8b). The overall effect of experimental condition was significant, RM ANOVA $F(1, 15) = 21.51, p = .0003, \eta^2 = .59$. Post-hoc pairwise Tukey HSD comparisons between the two conditions were significant in all four quarters, with the highest difference being the 3rd quarter: $T = -7.81$, $p = 0.001$, Cohen's $d = -.89$, partial $\eta^2 = .16$).

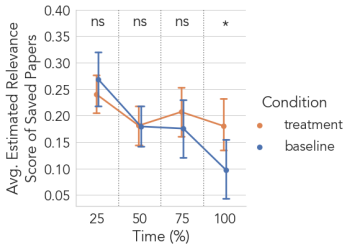
*6.4.3 Human-AI alignment on relevance was improved in ComLittee.* The relevance model used for sourcing and sorting paper recommendations was held at constant between the two systems, hence examining its relevance scores (a score closer to 1 indicates higher relevance) calculated for *saved* papers at the time of saving represents the degree of alignment between the human user's and AI's notions of relevance. We found that system's predicted scores of relevance on papers saved by users exhibited a widening gap on




**(a) Users saved and downvoted multiple papers from each saved author's publications in both conditions, suggesting how they recognized interrelated bodies of work.**




**(b) While in author-centric exploration, more than half of feedback was positive throughout the task, in paper-centric exploration users provided significantly more negative feedback in the third quarter of the task (see text).**




**(c) Predicted paper relevance scores decreased on user-saved papers in the 1st half of the task. However, in the 2nd half the average model relevance scores decreased significantly more in the baseline.**

**Figure 8: (a & b) Users actions and steering efforts; (c) Changes in machine-predicted relevance of saved papers.**

alignment in the first half of the task as shown in the decreasing average estimated relevance scores for both conditions. However, in the second half of the task, the average predicted relevance scores of saved papers decreased more in the baseline condition, leading to a significant difference between the two conditions in the last quarter of the task (Treatment: M=0.18, SD=0.246; Baseline: M=0.10, SD=0.247, post-hoc Tukey's HSD: $T = -2.14$, $p = .03$, Cohen's $d = -.33$, partial $\eta^2 = .03$, Fig. 8c).

*6.4.4 ComLittee users felt authors represented contextualized 'patches' of relevant research.* In a qualitative analysis we found themes that contextualize the results thus far. Users felt that ComLittee helped them find unfamiliar-yet-relevant authors (*e.g.,* "*find a bunch of interesting authors that I didn't know about*" – P16 ). In particular, P5 connected their experience to an analogy of foraging where "*[I would] drill down as looking at a specific author and the papers they publish... [it] helps me go from one world to another... like jumping from patch to patch.*" Furthermore, ComLittee was perceived as "*providing more context*" (P14) to exploration, helping users realize connections between two authors ("*Now that's very interesting, because I didn't know these people were connected*" – P2) or discover earlier, or less familiar, work that they had not known for a familiar author: "*I know [Author] does other work that's not relevant to my interest but the signals (explanation features) really helped me tease out which of his work is relevant... and a lot of them I haven't read before. So I think this is great in terms of helping me discover some of his earlier work that I can find helpful.*" (P10).

# 7 DISCUSSION AND FUTURE WORK

## 7.1 Study Design Limitations

The study design makes several important trade-offs for practical considerations. First, post-task ratings is efficient to collect retrospective data about participants' experience during the study. However, they may also be subject to confirmation biases towards their own earlier judgements. Second, conducting lab studies allowed us to control for unintended factors such as amount of time engaged with each systems. However, the relatively short duration of lab studies opens up the possibility that unobserved effects of time pressure may have led participants to engage with recommended items in a shallow manner. Furthermore, longer term effects of using the system requires a prolonged field deployment study with significantly more resource demands. Both of these limitation should apply to both conditions equally. Finally, to keep our study as realistic, we allowed participants to freely choose the two topics they wished to explore during the study to ensure their engagement and prior knowledge. The trade-off here is that the two topics may have differed qualitatively along the dimensions of topical familiarity or the level of abstraction. To mitigate this, participants were instructed to think of topics at a similar level of abstraction (*e.g.,* headings in the related work section of their own papers), and the topics participants chose were randomly assigned to the conditions in the experiment. For these reasons we do not expect to see a significant confounding effect from differences between the two topics.

## 7.2 Technical and human factors design implications for future author-centric discovery systems

*7.2.1 Latency of recommendations.* The efficiency of discovery in ComLittee was observed in spite of its significantly longer latency for retrieving recommendations. On average a recommendation request took significantly longer in ComLittee (M=12.9s, SD=8.69s) than Baseline (M=4.1s, SD=2.77s, $t_{\text{two-tailed}}(89.58)=8.99$, $p=3.73\times10^{-14}$), leading to a sizable difference in the numbers of user requests in each condition (N=83 in ComLittee vs. N=185 in baseline). Optimization could shorten this latency to be conducive to scaling and longer term use (*e.g.,* scoring and ranking authors' papers efficiently using pre-computed summary embeddings [23]).

*7.2.2 Combatting early cold-start phenomena.* Users in ComLittee attended new authors' relations to existing saved authors when saving them. As expected, initially authors users saved featured few relations to saved authors (cold start), as shown in the low % of authors saved that had *any* relation to a saved author (Fig. 9a, M=19%, SD=3.9%) in the 1st quarter. However, in the 2nd quarter this quickly increased to 87% (SD=3.4%, $t_{\text{two-tailed}}(82.24)=-8.56$, $p=5.0\times10^{-13}$), and plateaued for the remaining. The average number of relevance relations featured for a saved author, an indication of how strongly an author is related to the set of saved authors via coauthorship- and citation-based relations at the time of saving,

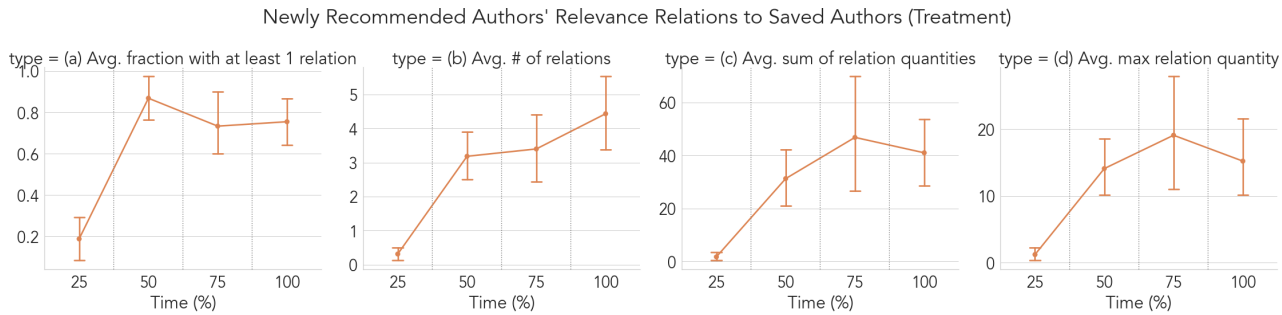



**Figure 9: (a, b) Initially author recommendations users saved did not have relations to saved authors (cold start), but ComLitteecould quickly recommend authors with relations to existing saved authors, to help users rapidly form a committee from the 2nd quarter and on as shown in the significant increase in % of authors with a relation. (c, d) However, the utility of relation strength (*i.e.,* the number of papers that relate a recommended author and a saved author) became marginal over time, suggesting that above a certain threshold participants did not need to differentiate the number of papers involved in a relation.**

also increased significantly from the 1st to 2nd quarter (from M=0.3, SD=0.72 to M=3.2, SD=2.28; $t_{\text{two-tailed}}(42.87)$=-7.49, $p$=3.0 × $10^{-9}$, Fig. 9b). The quantity of each relation (*i.e.,* how many papers did the two authors coauthor?; or cite from one another?) for a saved author increased from the 1st to 2nd quarter, and remained high (Fig. 9c, d). Taken together, these results show that while cold-start may be a challenge for author-centric discovery systems, users can combat this by recognizing relevant authors through recommendations sourced via triadic closure on citation networks, augmented with interactive relevance explanations, and iteratively curating them.

*7.2.3 Getting stuck in a particular 'school of thought' vs. steering the system to diversify discovery in later stages.* While users perceived ComLittee as helpful for making sense of scholarly relations among authors, on the flip side they paid more attention to and noticed more whether a new author recommendation belonged to a group of authors. Despite having saved more authors and especially unfamiliar ones compared to the baseline system (RQ2), users felt they had difficulty steering ComLittee to recommend authors with more diverse research backgrounds. P5 reasoned this as "*Only a small group – twenty to thirty – authors that, like, really work in that space a lot, and... co-author a lot of stuff together so it's kind of easy to stay in that insular community of authors.*" (P5). Users recognized potential dangers of "*falling into an echo chamber... where people that have the most papers are the ones given the most attention*" (P15) and "*it could be kind of hard to get out of a close-knit group of authors because they are interconnected*" (P7). P10 described that:

> "*[The system] is too good in recommending to a point where I had to fight with it a little in order to step out of the immediate circle... so I was a bit hesitant to add more authors from this school of thought because that will give the system even stronger signals... even if I wanted to step out of the immediate circle.*" – P10

Users pointed out an interesting dimension of authors, seniority, and how it may be considered differently during the exploration: "*I liked seeing these (familiar) authors early on... but it makes you wonder, if I have a perfect search engine... would it recommend new or up-and-coming authors who published at less known venues later?* – P5; "*So I blocked a bunch of (well recognized) authors towards the end who have broad interests, cited by everyone... compared to students who have more niche interests... and it (the system) was recommending less familiar names (after blocking them).*" – P11.

Interestingly, users also pointed out a similar steering challenge while interacting with the baseline, but with "*getting stuck in a bubble*" (P16) at a paper, not author, level. This in turn may have negative downstream consequences on author discovery: "*Maybe I steered the model too much to what I already know... the papers I'm seeing are the ones that I've already read before... and therefore discovering new authors now is a little bit hard to do.*" – P16; "*It's interesting because I think going around in circles with the same group of authors or papers... it happened more in the (baseline system).*" – P12; "*It felt like the system kind of ran out of things to recommend.*" – P5. Taken together the challenges around diversifying discovery with continued use, and especially for branching out of a particular school of thought point to an interesting design implications for future systems aimed at supporting scholarly discovery of both authors and papers (Section 7).

## 7.3 Beyond triadic closure on citation networks

Our work also contributes to the research in network analysis and social psychology on homophily [28] and triadic closure (*cf.* [3]). Recent work by Abebe et al. [1] suggests that the forces of triadic closure on networks can have positive, desegregating potential to increase overall network integration. Here, we explored whether this insight also generalizes to scholarly discovery, and in doing so, instrumented an interactive system to study user behaviors and showed that they rapidly saved familiar authors early on, as a form of 'navigational springboards' (Section 6.2), to then discover significantly more novel authors. Despite this gain in novel author discovery, users had concerns regarding branching out of a closely related group of authors in a later stage of using the system, and in some cases avoided saving more authors who were strongly connected to the committee (Section 7.2.3). These findings demonstrate the feasibility of leveraging triadic closure on authorship graphs for its desegregating potential, by way of recommending authors connected via authors familiar to the user to then navigate them

to save more unfamiliar authors. Fruitful avenue for future work lies in understanding the criteria for effective system designs for surfacing core navigational author nodes.

## 7.4 Finding non-interacting bodies of literature

A potential limitation of methods relying on citation networks is its reduced discoverability potential of work outside frequently co-cited bodies of literature, leading to a form of filter bubbles [36]. Yet, bursting filter bubbles can have outsized potential for catalyzing significant innovations [46] especially for domains less likely to interact with each other [9, 50]. Recent work on analogical scientific inspirations [22] and product innovation [17] show early evidence on how analogical relations between papers may be computationally extracted, thereby filling in a 'discovery hole' from relying only on conventional approaches or citation-based mechanisms to scholarly search. In addition, alternative complementary approaches may use differences between knowledge domains more explicitly to induce cross-domain retrieval in recommendations (*e.g.,* [21]). This line of prior work therefore points to fruitful avenues for future extensions.

## 7.5 Diverse intentions of user feedback

Our findings suggest that potential misalignment between the user's and the system's model of relevance need to be carefully handled. Our quantitative (Fig. 8c), behavioral (Fig. 8a,8b), and qualitative analyses suggest that this may be an important issue for user adoption and the utility of the system. In particular, users perceived that the system was converging to a specific type of similarity and hypothesized that this was due to its optimization objective in both conditions, which was described as "*going around in circles with a close-knit group of authors*" in ComLittee and "*papers topically too similar in nature*" in the baseline system. Potential implications for future systems that aim to support the changing notion of user relevance and their alternating desires – sometimes seeking broader and other times more focused results – are around how various user intentions of feedback may be supported, such as 1) steering the recommendations with varying degrees of expected changes in the outcomes (*i.e.,* small amount of feedback for tuning vs. large amount for jumping); 2) saving interesting (but not necessarily relevant to the task) intermediate results; and 3) subselecting from the user's own feedback to experiment with model behaviors, form a mental model, and re-use the subselection as a query or a sub-folder of the topic in case good results are yielded.

## 7.6 Flow of time

Our findings also uncover how time is an important dimension of literature discovery, and especially for systems that aim to leverage users' prior knowledge early in the interaction. The users of ComLittee exhibited a distinctive behavioral pattern that consisted of rapidly leveraging known, familiar authors early on (Fig. 9), and seeking novel authors afterwards (Fig. 6). Successfully adapting the system's objective in relation to this temporal context could improve user perception of its alignment and subsequently adoption. Some users explicitly commented on how they later blocked several of the more senior and well-known authors they found relevant earlier, in hopes of targeting authors with "*more specified, niche research interests.*" Therefore future systems may design for the ability to filter outcomes based on certain attributes in a post-hoc manner, or prioritize items that meet certain criteria at the recommendation time. Other users also hypothesized that accumulation of their feedback over time made effecting desired changes in outcomes require commensurable amounts of feedback, which felt laborious and potentially frustrating in repeated use. They expressed the need to "*tell the machine to forget about my early feedback, without having to go back and redo it myself.*" While a long line of research exists for modeling user's cognitive state, here we emphasize how simple interaction affordances such as being able to pull up the history of user feedback, locate a time range of their feedback, and specify the types of feedback no longer relevant may prove to be effective. Finally, yet others commented on how they thought they were saving authors who are disproportionately more well-known or familiar using ComLittee, even though the results showed that users overall saved more, not less, unfamiliar authors. This implies users may want to actively reflect on their progress over time, and making it explicit may help to minimize the potential perception biases that may increase due to sensitivity to certain attributes of the saved results.

## 7.7 Supporting application designs using artifacts of exploration

While our evaluation showed the feasibility of integrating author-centric organization with paper recommendations to support users in rapidly building mental models of the literature, this also opens up many possibilities for future application designs that build on top of the artifact users curated during the exploration. One such example is interactive, affinity-based grouping of saved authors. Our users alluded to this possibility by commenting on how "*different schools of thought*" co-exist in the literature which was made visible from looking at explanations of authors' citation and coauthorship relations, and how they might want to break out of one school to another for their own exploration. Incorporating additional attributes of authors may help with identifying different affinity signals related to this. For example users described how disciplinary background of an author may represent a specific kind of frequently used epistemological approaches, and how being able to group them based on this axis would contribute to understanding how different approaches in a topic evolved over time and interact with each other. Another example application design is to extend the publication time visualization featured in ComLittee to groups of authors as a means to communicating which areas of research and epistemological approaches have recently been popular, and thereby empowering users to see what may be yet-to-be discovered or overlooked areas of future research opportunities. Our users also alluded to this possibility by commenting on how certain literature is 'sparser' or 'more densely published' than others, and how interesting spaces for interdisciplinary approaches can be found from looking at the adjacent areas of literature that they may pull from, which suggests a fruitful avenue for future application designs that employ interactive visualization techniques to support effective user exploration.

## 8 CONCLUSION

An important aspect in the domain of supporting scholarly discovery via interactive systems, which has not received significant attention from researchers, is how scientists' evolving knowledge of others may be captured and utilized to enhance their experience of discovering new authors and papers. Here we reflect on the design space of such tools and introduce ComLittee, a literature discovery system that supports enhanced author-centric exploration by first enabling user curation of relevant authors, second using the curated authors to compute relevance signals on other authors and their work, and lastly using these signals to recommend further relevant authors and enhance understanding of the recommendations. We demonstrate the feasibility and value of ComLittee in a controlled study, and specifically show how ComLittee leads to discovering a larger number of relevant papers and authors in a given amount of time, and how participants also rated discovered authors as more novel, while discovered papers as more interesting. Furthermore, we show how initially forming a group of authors familiar to the user can lay down an enriched path for subsequent discovery.

## ACKNOWLEDGMENTS

This project is supported by NSF Grant OIA-2033558. The authors thank Luca Soldaini and Chris Wilhelm for their advice on and help with engineering the system; Daniel S. Weld, Doug Downey, and the researchers in the Semantic Scholar team for discussions and thoughtful feedback on the project. We also thank the anonymous reviewers for their constructive feedback. Finally, this work would not have been possible without our pilot test and user study participants.

## REFERENCES

[1] Rediet Abebe, Nicole Immorlica, Jon Kleinberg, Brendan Lucier, and Ali Shirali. 2022. On the Effect of Triadic Closure on Network Segregation. In *Proceedings of the 23rd ACM Conference on Economics and Computation* (Boulder, CO, USA) *(EC '22)*. Association for Computing Machinery, New York, NY, USA, 249–284. https://doi.org/10.1145/3490486.3538322

[2] Saleema Amershi, James Fogarty, and Daniel Weld. 2012. Regroup: Interactive Machine Learning for on-Demand Group Creation in Social Networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 21–30. https://doi.org/10.1145/2207676.2207680

[3] Aili Asikainen, Gerardo Iñiguez, Javier Ureña-Carrión, Kimmo Kaski, and Mikko Kivelä. 2020. Cumulative effects of triadic closure and homophily in social networks. *Science Advances* 6, 19 (2020), eaax7310.

[4] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications* 8, 1 (2021), 1–15.

[5] Ronald S Burt. 2004. Structural holes and good ideas. *American journal of sociology* 110, 2 (2004), 349–399.

[6] Joseph Chee Chang, Nathan Hahn, Adam Perer, and Aniket Kittur. 2019. SearchLens: Composing and capturing complex user interests for exploratory search. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 498–509.

[7] Duen Horng Chau, Aniket Kittur, Jason I. Hong, and Christos Faloutsos. 2011. Apolo: Interactive Large Graph Sensemaking by Combining Machine Learning and Visualization. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Diego, California, USA) *(KDD '11)*. Association for Computing Machinery, New York, NY, USA, 739–742. https://doi.org/10.1145/2020408.2020524

[8] Kiroong Choe, Seokweon Jung, Seokhyeon Park, Hwajung Hong, and Jinwook Seo. 2021. Papers101: Supporting the discovery process in the literature review workflow for novice researchers. In *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*. IEEE, 176–180.

[9] Johan SG Chu and James A Evans. 2021. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences* 118, 41 (2021), e2021636118.

[10] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2270–2282. https://doi.org/10.18653/v1/2020.acl-main.207

[11] Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology* 78, 6 (1973), 1360–1380.

[12] Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, and Inbal Ronen. 2013. Mining expertise and interests from social media. In *Proceedings of the 22nd international conference on World Wide Web*. 515–526.

[13] Ido Guy and David Carmel. 2011. Social recommender systems. In *Proceedings of the 20th international conference companion on World wide web*. 283–284.

[14] Han L Han, Junhang Yu, Raphael Bournet, Alexandre Ciorascu, Wendy E Mackay, and Michel Beaudouin-Lafon. 2022. Passages: Interacting with Text Across Documents. In *CHI Conference on Human Factors in Computing Systems*. 1–17.

[15] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[16] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 413.

[17] Tom Hope, Ronen Tamari, Daniel Hershcovich, Hyeonsu B Kang, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2022. Scaling Creative Inspiration with Fine-Grained Functional Aspects of Ideas. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 12, 15 pages. https://doi.org/10.1145/3491102.3517434

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

[19] Hyeonsu B. Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An Interactive System for Personalized Thread-based Exploration and Organization of Scientific Literature. In *The 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA.

[20] 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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 302, 23 pages. https://doi.org/10.1145/3491102.3517470

[21] Hyeonsu B Kang, Sheshera Mysore, Kevin J Huang, Haw-Shiuan Chang, Thorben Prein, Andrew McCallum, Aniket Kittur, and Elsa Olivetti. 2022. Augmenting Scientific Creativity with Retrieval across Knowledge Domains. In *Second Workshop on Bridging Human-Computer Interaction and Natural Language Processing at NAACL 2022*. arXiv. https://doi.org/10.48550/ARXIV.2206.01328

[22] Hyeonsu B. Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting Scientific Creativity with an Analogical Search Engine. *ACM Trans. Comput.-Hum. Interact.* (mar 2022). https://doi.org/10.1145/3530013 Just Accepted.

[23] Harmanpreet Kaur, Doug Downey, Amanpreet Singh, Evie Yu-Yen Cheng, Daniel S. Weld, and Jonathan Bragg. 2022. FeedLens: Polymorphic Lenses for Personalizing Exploratory Search over Knowledge Graphs *(UIST '22)*.

[24] Henry Kautz, Bart Selman, and Mehul Shah. 1997. Referral Web: combining social networks and collaborative filtering. *Commun. ACM* 40, 3 (1997), 63–65.

[25] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

[26] Gueorgi Kossinets and Duncan J Watts. 2006. Empirical analysis of an evolving social network. *science* 311, 5757 (2006), 88–90.

[27] Andrew Kuznetsov, Joseph Chee Chang, Nathan Hahn, Napol Rachatasumrit, Bradley Breneisen, Julina Coupland, and Aniket Kittur. 2022. Fuse: In-Situ Sensemaking Support in the Browser. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, Article 34. https://doi.org/10.1145/3526113.3545693

[28] Paul F Lazarsfeld, Robert K Merton, et al. 1954. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society* 18, 1 (1954), 18–66.

[29] Howard Levene. 1961. Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling* (1961), 279–292.

[30] David Liben-Nowell and Jon Kleinberg. 2003. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*. 556–559.

[31] Michael Xieyang Liu, Aniket Kittur, and Brad A. Myers. 2022. Crystalline: Lowering the Cost for Developers to Collect and Organize Information for Decision Making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 68, 16 pages. https://doi.org/10.1145/3491102.3501968

[32] Michael Xieyang Liu, Andrew Kuznetsov, Yongsung Kim, Joseph Chee Chang, Aniket Kittur, and Brad A. Myers. 2022. Wigglite: Low-Cost Information Collection and Triage. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 32, 16 pages. https://doi.org/10.1145/3526113.3545661

[33] John W Mauchly. 1940. Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics* 11, 2 (1940), 204–209.

[34] David W. McDonald and Mark S. Ackerman. 2000. Expertise Recommender: A Flexible Recommendation System and Architecture. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (Philadelphia, Pennsylvania, USA) *(CSCW '00)*. Association for Computing Machinery, New York, NY, USA, 231–240. https://doi.org/10.1145/358916.358994

[35] Sheshera Mysore, Mahmood Jasim, Haoru Song, Sarah Akbar, Andre Kenneth Chase Randall, and Narges Mahyar. 2023. How Data Scientists Review the Scholarly Literature. *arXiv preprint arXiv:2301.03774* (2023).

[36] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity. In *Proceedings of the 23rd International Conference on World Wide Web* (Seoul, Korea) *(WWW '14)*. Association for Computing Machinery, New York, NY, USA, 677–686. https://doi.org/10.1145/2566486.2568012

[37] Srishti Palani, Zijian Ding, Stephen MacNeil, and Steven P. Dow. 2021. The "Active Search" Hypothesis: How Search Strategies Relate to Creative Learning. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 325–329. https://doi.org/10.1145/3406522.3446046

[38] Srishti Palani, Zijian Ding, Austin Nguyen, Andrew Chuang, Stephen MacNeil, and Steven P. Dow. 2021. CoNotate: Suggesting Queries Based on Notes Promotes Knowledge Discovery. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 726, 14 pages.

[39] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, Hamburg, Germany.

[40] Srishti Palani, Yingyi Zhou, Sheldon Zhu, and Steven P. Dow. 2022. InterWeave: Presenting Search Suggestions in Context Scaffolds Information Search and Synthesis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 93, 16 pages. https://doi.org/10.1145/3526113.3545696

[41] Peter Pirolli. 2009. An elementary social information foraging model. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 605–614.

[42] Antoine Ponsard, Francisco Escalona, and Tamara Munzner. 2016. PaperQuest: A visualization tool to support literature review. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2264–2271.

[43] Jason Portenoy, Marissa Radensky, Jevin D West, Eric Horvitz, Daniel S Weld, and Tom Hope. 2022. Bursting Scientific Filter Bubbles: Boosting Innovation via Novel Author Discovery. Association for Computing Machinery, New York, NY, USA.

[44] Napol Rachatasumrit, Jonathan Bragg, Amy X. Zhang, and Daniel S Weld. 2022. CiteRead: Integrating Localized Citation Contexts into Scientific Paper Reading. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 707–719. https://doi.org/10.1145/3490099.3511162

[45] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018. Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. 1, 4 (2018).

[46] Andrey Rzhetsky, Jacob G Foster, Ian T Foster, and James A Evans. 2015. Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences* 112, 47 (2015), 14569–14574.

[47] Pamela Sandstrom. 2001. Scholarly communication as a socioecological system. *Scientometrics* 50, 3 (2001), 573–605.

[48] Shivashankar Subramanian, Daniel King, Doug Downey, and Sergey Feldman. 2021. S2AND: A Benchmark and Evaluation System for Author Name Disambiguation. *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (2021), 170–179.

[49] Nicole Sultanum, Christine Murad, and Daniel Wigdor. 2020. Understanding and supporting academic literature review workflows with litsense. In *Proceedings of the International Conference on Advanced Visual Interfaces*. 1–5.

[50] Don R. Swanson. 1986. Undiscovered Public Knowledge. *The Library Quarterly* 56, 2 (1986), 103–118.

[51] Jiliang Tang, Xia Hu, and Huan Liu. 2013. Social recommendation: a review. *Social Network Analysis and Mining* 3, 4 (2013), 1113–1133.

[52] Chun-Hua Tsai and Peter Brusilovsky. 2021. The effects of controllability and explainability in a social recommender system. *User Modeling and User-Adapted Interaction* 31, 3 (2021), 591–627.

[53] R Van Noorden. 2014. Global scientific output doubles every nine years [blog post]. *Retrieved from nature. com at http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html* (2014).

[54] Wei Wang, Xiaomei Bai, Feng Xia, Teshome Megersa Bekele, Xiaoyan Su, and Amr Tolba. 2017. From triadic closure to conference closure: The role of academic conferences in promoting scientific collaborations. *Scientometrics* 113, 1 (2017), 177–193.

[55] Yun Wang, Dongyu Liu, Huamin Qu, Qiong Luo, and Xiaojuan Ma. 2016. A guided tour of literature review: Facilitating academic paper reading with narrative visualization. In *Proceedings of the 9th International Symposium on Visual Information Communication and Interaction*. 17–24.

[56] Jen-Her Wu and Shu-Ching Wang. 2005. What drives mobile commerce?: An empirical evaluation of the revised technology acceptance model. *Information & management* 42, 5 (2005), 719–729.

[57] Xiaolong Zhang, Yan Qu, C. Lee Giles, and Piyou Song. 2008. CiteSense: Supporting Sensemaking of Research Literature. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1357054.1357161

# A HOW AUTHOR RECOMMENDATIONS WERE SORTED IN A BATCH FOR PRESENTATION

The position of an item on a list may determine whether it receives user's meaningful attention. While the batchified exploration in our system did not serve a large quantity of recommendations at once (up to 8 items), it nonetheless needed an order of presentation among the items in each batch. We iterated on its design through pilot studies. Our initial ordering simply interleaved the different sourcing mechanisms, which led to stratified sampling. Participants commented that this was confusing due in part to how author recommendations with several relevance explanation filters were featured lower than those that had none, because they expected a higher relational strength (for example, this happened when the author recommendations from the citation-based expansion mechanism were assigned to appear later than those from the relevant paper recommendations-based mechanism). To prevent this, we defined a relevance ratio as follows. For each author $a_i$, and her author-level relevance tags $\tau_j(a_i)$, we count the number of unique papers $C_p$ that appeared in them: $\sum_j C_p(\tau_j(a_i))$. To further enrich the strength of the relevance signal while also making comparison between authors fairer (*i.e.,* authors who have *less* total publications but *more* related publications out of the total may be perceived to have a higher density of relevant work, and therefore more interesting to the user), we normalized this quantity by the number of publications: $\sum_j C_p(\tau_j(a_i))/C_p(a_i)$.

# B PSEUDOCODE OF AUTHOR RECOMMENDATION

Pseudocode for author recommendation is shown in Algorithm 1.

# C USERS' ACTIONS REPRESENTED AUTHENTIC RELEVANCE AND INTERESTINGNESS

Importantly, few of the randomly sampled saved papers were judged as irrelevant (on average users in both conditions rated around 13

**Algorithm 1** Pseudocode Descriptions of Co-authorship- and Citation-based Recommendation Algorithms

**procedure** Sort-Sample($S$, $k$, $N$) ▷ Approximating the most topically relevant papers for efficiency
  Sort $s \in S$ in a descending order of Accessor($s$, $k$) ▷ Accessor($s$, $k$) returns value of the field $k$ from $s$
  Sample $p_1, \cdots, p_N$ from the top of the sorted list
  **return** $\{p_1, \cdots, p_N\}$
**end procedure**
**procedure** Vote-Multi($P$) ▷ Each paper adds 1 vote to each of its authors
  $\Omega \leftarrow$ Empty Dictionary $\omega_\emptyset$ ▷ $\Omega$ := (key: Author ID, value: Frequency $\geq 0$) store
  **for** $\forall p \in P$ **do**
    **for** $\forall a \in A_p$ **do** ▷ $A_p$ := Authors of Paper $p$
      $\Omega[a_{\text{ID}}]$ += 1 ▷ $a_{\text{ID}}$ := ID of author $a$
  **return** $\Omega$
**end procedure**
**procedure** Get-Relevant-Papers($A$, $P_{\text{feedback}}$)
  $P_{\text{filtered}} \leftarrow \emptyset$
  **for** $\forall a \in A$ **do**
    **for** $\forall p \in P_a$ **do** ▷ $P_a$ := Publications of Author a
      **if** $p \notin P_{\text{feedback}}$ **then**
        $p_{\text{score}} \leftarrow$ Score($p$, $P_{\text{feedback}}$) ▷ Score($p$, $P_{\text{feedback}}$) returns score of $p$ with ensemble SVMs
      **else**
        $p_{\text{score}} \leftarrow$ Retrieve($p$, $P_{\text{feedback}}$) ▷ Retrieve($p$, $P_{\text{feedback}}$) returns user-feedback on $p \in \{-1, 1\}$
      **if** $p_{\text{score}} > 0$ **then**
        $P_{\text{filtered}} \leftarrow P_{\text{filtered}} \cup \{p\}$
  $P_{\text{sampled}} \leftarrow$ Sort-Sample($P_{\text{filtered}}$, score, 100)
  **return** $P_{\text{sampled}}$
**end procedure**
**procedure** Co-authorship-based Recommendation($A$, $P_{\text{feedback}}$) ▷ $A$ := Committee authors
  $P_{\text{sampled}} \leftarrow$ Get-Relevant-Papers($A$, $P_{\text{feedback}}$) ▷ $P_{\text{feedback}}$ := Papers with user feedback
  $\Omega \leftarrow$ Vote-Multi($P_{\text{sampled}}$)
  $\Omega \leftarrow \Omega \setminus \{\text{self}, \forall a \in A\}$ ▷ Remove the user herself and committee authors
  **return** $\Omega$
**end procedure**
**procedure** Vote-Author($P$, $A$) ▷ Each paper receives up to 1 vote (when cited) from each author
  $\Omega \leftarrow$ Empty Dictionary $\omega_\emptyset$ ▷ $\Omega$ := (key: Paper ID, value: Frequency $\geq 0$) store
  **for** $\forall a \in A$ **do**
    **for** $\forall p \in P$ **do**
      **if** $a$ cites $p$ **then**
        $\Omega[p_{\text{ID}}]$ += 1 ▷ $p_{\text{ID}}$ := ID of paper $p$
  **return** Sort-Sample($\Omega$, ID, 100)
**end procedure**
**procedure** Citation-based Recommendation($A$, $P_{\text{feedback}}$) ▷ $A$ := Committee authors
  $P_{\text{sampled}} \leftarrow$ Get-Relevant-Papers($A$, $P_{\text{feedback}}$) ▷ $P_{\text{feedback}}$ := Papers with user feedback
  $P_{\text{cited}} \leftarrow$ Get-References($P_{\text{sampled}}$) ▷ Returns a set of referenced papers from $\forall p \in P_{\text{sampled}}$
  $P_{\text{voted}} \leftarrow$ Vote-Author($P_{\text{cited}}$, $A$)
  $\Omega \leftarrow$ Vote-Multi($P_{\text{voted}}$)
  $\Omega \leftarrow \Omega \setminus \{\text{self}, \forall a \in A\}$ ▷ Remove the user herself and committee authors
  **return** $\Omega$
**end procedure**

out of 15 randomly sampled papers as 4 or higher on the relevance scale), suggesting that users' save-paper actions represented their authentic judgment of relevance. We also validated the significant difference in the average number of saved authors between the conditions by using user-vetted relevance ratings on saved authors. For save-author actions, we ran an additional validity check by removing authors rated as 4 or lower on the relevance scale[2],

[2] *I.e.,* 4 corresponded to neutral agreement on '*I found this author to be relevant.*' This presents a possibility that the author may become relevant had there been more time to explore. For this reason, we include response 4 as an indication of relevance. However we still observe a statistically significant difference between the conditions when authors scored 4 are excluded in the analysis ($t_{\text{paired}}(28.00)$=-2.88, $p$=0.01).

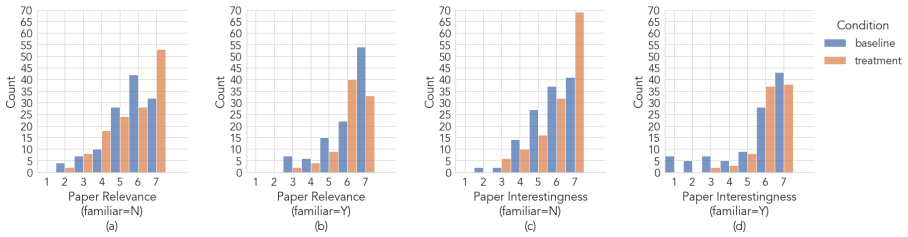



**Figure 10: Histogram of (a & b) post-task paper relevance ratings for either familiarity type and (c & d) paper interestingness.**

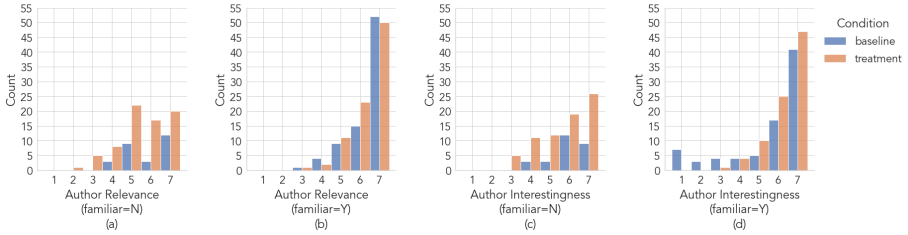



**Figure 11: Histogram of (a & b) post-task author relevance ratings for either familiarity type and (c & d) author interestingness.**


**(a) An example paper recommendation featuring a "cited by [a saved author]" label. Self-citations were excluded (see text).**


**(b) An example 'stack of (judged) papers' UI for balancing the need for seeing familiar vs. new papers for an author (see text).**

**Figure 12: UI designs in the baseline interface**

and re-analyzing trends in the number of saved authors. We see a consistent trend where users saved significantly more authors in ComLittee (M=9.6, SD=2.63) than Baseline (M=6.7, SD=2.75, $t_{\text{paired}}(29.94)=3.71$, $p=0.002$), suggesting that users' decisions to save an author similarly represented their authentic judgment of relevance of the author. Taken together, we conclude that user save actions likely represented a level of authentic user interest and relevance in the saved items, beyond merely as a means to steering the recommender system.

## D DISTRIBUTION OF POST-TASK RATINGS ON SAVED PAPERS AND AUTHORS, BY FAMILIARITY

Distribution of the counts on paper and author ratings are shown in Fig. 10 and Fig. 11, respectively.

## E DESCRIPTION OF DESIGN ITERATIONS

**Removing self-citations from relevance explanations.** Pilot users expressed their intent for clicking on a relation explanation filter of an author (*i.e.,* "cited by [a saved author]") as to see other papers by the saved author that cited the recommended author's papers. When these papers included self-citations, however, users did not feel it matched their intent and question the usefulness of relations due to the self-promoting nature of self-citations. To align with the user intent, we excluded self-citations from the data for featuring author-level explanations, and also from paper-level citation explanation labels for consistency (Fig. 12a).

**Increasing the information density on an author details page by adaptively minimizing judged papers.** Pilot users also expressed wanting to see other papers by an author that they had not seen before first and foremost, rather than seeing papers that they had already provided feedback on. This makes sense especially in cases when the user is receiving new publication recommendations from authors whose earlier work they are already familiar with. However, in short-term use scenarios, we anticipated that there may be tension with users wanting to see familiar papers to build a mental model of and increase their confidence in judgment for a new author recommendation, especially when the user is trying to make a decision to save or downvote a paper. Therefore, we approached this trade-off by designing a mechanism for collapsing the familiar papers – sorted from most to least recently interacted with – that the user has provided feedback on into a stack of papers UI at the top of the author's publications list (Fig. 2d), while also featuring an 'expand' button next to the stack in case the user wanted to see individual papers in the stack (Fig. 12b).

**Presentation order of recommendations.** Pilot users also pointed out the prominence of the 'predicted number of relevant papers' tag featured for each author recommendation and how it could be misleading when author recommendations higher on the rank did not feature a higher quantity. Because this number was perceived useful for pilot users, but was not the main determinant of the presentation order of author recommendations within a batch (see Section 4.2.5), we moved this information and made it less prominent in the final interface design (Fig. 2c).

## F ADDITIONAL SURVEY RESULTS

Descriptions of survey items and participants' responses grouped by condition are presented in Table 2 and 3. Two-sided paired samples t-tests were performed to compute the $p$-values between conditions. See Section 6.1.2 for discussions of the results.

| | **Description** | Baseline | ComLittee | $p$-val. |
|---|---|---|---|---|
| 1. NASA-TLX | Sum of the participants' responses to the five NASA-TLX's [15] Likert-scale questionnaire items below. The original 21-point scale was mapped to a 7-point scale, similarly with [45]. | 17.0 (SD=6.01) | 14.1 (SD=5.56) | .01* |
| 1a. Mental | "How mentally demanding was the task?" | 3.6 (SD=1.55) | 3.4 (SD=1.55) | .79 |
| 1b. Physical | "How physically demanding was the task?" | 3.9 (SD=1.54) | 2.4 (SD=1.21) | .002** |
| 1c. Temporal | "How hurried or rushed was the pace of the task?" | 3.1 (SD=1.69) | 2.7 (SD=1.40) | .38 |
| 1d. Effort | "How hard did you have to work to accomplish your level of performance?" | 3.5 (SD=1.26) | 3.1 (SD=1.54) | .33 |
| 1e. Frustration | "How insecure, discouraged, irritated, stressed, and annoyed were you?" | 3.0 (SD=1.59) | 2.4 (SD=1.50) | .20 |
| 2. TAM | Sum of the participants' responses to the 4 questionnaire items below adopted from [56] measuring the technological compatibility with participants' existing scholarly discovery workflows and the easiness of learning. | 19.4 (SD=5.09) | 22.6 (SD=2.99) | .02* |
| 2a. Compatibility | "*Using the system is compatible with most aspects of how I search for scholars and their papers.*" (The response Likert scales for this question and below are 1: *Strongly disagree*, 7: *Strongly agree*) | 4.4 (SD=5.08) | 5.1 (SD=1.24) | .15 |
| 2b. Compatibility | "*The system fits well with the way I like to search for scholars and their papers.*" | 4.3 (SD=1.65) | 5.1 (SD=1.02) | .14 |
| 2c. Easy-to-Learn | "*I think learning to use the system is easy.*" | 5.8 (SD=1.05) | 6.2 (SD=1.02) | .12 |
| 2d. Adoption | "*Given that I had access to the system, I predict that I would use it.*" | 4.9 (SD=1.50) | 6.2 (SD=0.75) | .003** |
| 3. Author Discovery | Sum of participants' responses to the 4 questionnaire items below. | 15.6 (SD=5.76) | 21.9 (SD=4.33) | .001** |
| 3a. Finding | "*The system helped me find relevant authors.*" | 3.9 (SD=1.82) | 6.1 (SD=1.34) | .002** |
| 3b. Curiosity | "*The system made me curious about authors' research.*" | 4.3 (SD=1.58) | 6.1 (SD=1.34) | .001** |
| 3c. Sensemaking | "*The system helped me make sense of authors' research.*" | 3.5 (SD=1.46) | 4.8 (SD=1.53) | .008** |
| 3d. Explanation Helpfulness | "*The system's explanations of relevant authors became more helpful the more I used the system.*" | 3.9 (SD=1.77) | 4.9 (SD=1.59) | .02* |

**Table 2: Descriptions of additional questionnaire items and responses grouped by condition. $p-$values are from two-sided paired samples Wilcoxon's signed rank tests. The results show that the overall workload was significantly lower in the ComLittee condition than the Baseline condition. While the adoption plausibility was higher in the ComLittee condition, the overall TAM responses did not differ significantly between the two conditions. ComLittee responses were significantly more favorable towards all author discovery helpfulness questions than those of the Baseline condition.**

| | **Description** | Baseline | ComLittee | $p$-val. |
|---|---|---|---|---|
| 4. Paper Discovery | Sum of participants' responses to the 3 questionnaire items below. | 13.6 (SD=4.21) | 17.0 (SD=2.85) | .002** |
| 4a. Finding | "*The system helped me find relevant papers.*" | 3.9 (SD=1.82) | 6.1 (SD=1.34) | .005** |
| 4b. Curiosity | "*The system made me curious about the papers I found.*" | 4.8 (SD=1.64) | 6.2 (SD=0.83) | .01* |
| 4c. Explanation Helpfulness | "*The system's explanations of relevant papers became more helpful the more I used the system.*" | 3.6 (SD=1.63) | 4.6 (SD=1.54) | .052 |
| 5. Common Features | Avg. of participants' responses to the 4 questionnaire items below. | 3.3 (SD=1.26) | 3.5 (SD=1.27) | .25 |
| 5a. # of Papers | "*I found the authors' total number of publications useful.*" (Example provided) | 3.1 (SD=1.48) | 3.8 (SD=1.81) | .07 |
| 5b. # of Relevant Papers | "*I found the number of relevant papers estimated by the system useful.*" (Example provided) | 3.6 (SD=1.46) | 3.8 (SD=1.57) | .65 |
| 5c. h-index | "*I found authors' h-index useful.*" | 3.3 (SD=1.39) | 2.9 (SD=1.39) | .15 |
| 5d. # of Citations | "*I found authors' citation counts useful.*" | 3.5 (SD=1.32) | 3.7 (SD=1.82) | .72 |
| 5e. Relevance Score | "*I found the "relevance score" explanation for each paper useful.*" (Example provided) | 4.5 (SD=1.90) | 4.5 (SD=1.90) | .97 |
| Specific Features | Avg. of participants' responses to the condition-specific feature questionnaire items below. | 4.8 (SD=1.33) | 4.8 (SD=0.96) | .74 |
| Most Favored Feature | Avg. of the highest-rated condition specific feature for each participant. | 5.5 (SD=1.51) | 6.2 (SD=0.83) | .05* |
| Coauthor Filter | "*I found the "co-authored with [a saved author]" filter buttons useful.*" (Example provided) | NA | 5.3 (SD=1.45) | NA |
| Cited-by Filter | "*I found the "cited by [a saved author]" filter buttons useful.*" (Example provided) | NA | 5.3 (SD=1.54) | NA |
| Histogram | "*I found the histogram useful.*" | NA | 3.8 (SD=1.47) | NA |
| Histogram Filter | "*I found being able to see the selected filter counts on the histogram useful.*" (Example provided) | NA | 4.2 (SD=1.68) | NA |
| Saved Coauthor Highlight | "*I found the "co-authored with [a saved author]" explanation for each paper useful.*" (Example provided) | NA | 4.9 (SD=1.41) | NA |
| Cited-by Paper Exp. | "*I found the "[a saved author] cited this paper" explanation for each paper useful.*" (Example provided) | NA | 5.4 (SD=1.41) | NA |
| FeedLens [23] Dots | "*I found the green dots next to author names useful.*" (Example provided) | 4.8 (SD=1.76) | NA | NA |
| Tooltip | "*I found the author tooltip explanation useful.*" (Example provided) | 4.9 (SD=1.45) | NA | NA |

**Table 3: Descriptions of additional questionnaire items and responses grouped by condition, continued. $p$−values are from two-sided paired samples Wilcoxon's signed rank tests. ComLittee responses were significantly more favorable towards all paper discovery helpfulness questions than those of the Baseline condition. For common features responses did not show any significant difference between the two conditions. Furthermore, the average responses to system-specific feature questions showed no significant advantage of one system over the other. However, aggregating over the most favored feature from each user, ComLittee showed a significantly higher average.**