Community Clustering: Leveraging an Academic Crowd to Form Coherent Conference Sessions

Paul André,[*] Haoqi Zhang,[†] Juho Kim,[‡] Lydia Chilton,[§] Steven P. Dow,[*] and Robert C. Miller[‡]

[*]Carnegie Mellon
Pittsburgh, PA
{pandre,spdow}@cs.cmu.edu

[†]Northwestern University
Evanston, IL
hq@northwestern.edu

[‡]MIT CSAIL
Cambridge, MA
{juhokim,rcm}@mit.edu

[§]University of Washington
Seattle, WA
hmslydia@cs.washington.edu

Abstract

Creating sessions of related papers for a large conference is a complex and time-consuming task. Traditionally, a few conference organizers group papers into sessions manually. Organizers often fail to capture the affinities between papers beyond created sessions, making incoherent sessions difficult to fix and alternative groupings hard to discover. This paper proposes *committeesourcing* and *authorsourcing* approaches to session creation (a specific instance of clustering and constraint satisfaction) that tap into the expertise and interest of committee members and authors for identifying paper affinities. During the planning of ACM *CHI'13*, a large conference on human-computer interaction, we recruited committee members to group papers using two online distributed clustering methods. To refine these paper affinities—and to evaluate the committeesourcing methods against existing manual and automated approaches—we recruited authors to identify papers that fit well in a session with their own. Results show that authors found papers grouped by the distributed clustering methods to be as relevant as, or more relevant than, papers suggested through the existing in-person meeting. Results also demonstrate that communitysourced results capture affinities beyond sessions and provide flexibility during scheduling.

A core part of conference scheduling is creating sessions of related papers. For large conferences with hundreds of papers, this is a complex and time-consuming task. We observed the session creation process for ACM CHI, the largest conference on human-computer interaction. CHI 2013 received nearly 2000 paper submissions and accepted almost 400. The conference organizers formed 80-minute sessions with 4–5 papers each, with 16 parallel sessions spanning four days. As a typical current practice, a handful of organizers used paper printouts to generate initial sessions at a committee meeting.

In interviews, organizers noted that "*papers fit into sessions in complex ways*" and that "*getting a session together that makes sense is hard*." Organizers at the committee meeting focus on creating good sessions quickly, and do not capture paper affinities beyond the single session in which they are grouped. Since time is limited and available organizers' expertise may be incomplete, this process can lead to sessions with incoherent themes and stray papers forced into existing sessions. Since paper affinities beyond created sessions are not captured, incoherent sessions are hard to fix and alternative groupings hard to discover. Methods based on linguistic or statistical techniques offer automated groupings of related papers. For example, affinity-based methods such as TF-IDF can be used to identify similar papers (Salton and McGill 1983) and topic-modeling methods such as LDA can be used to discover topic-based groupings (Blei, Ng, and Jordan 2003). However, a general lack of human cognition, domain expertise, and natural language ability can lead automated methods to produce poor results that fail to capture fine-grained distinctions among papers.

The session creation problem can be generalized as taking as input a set of items, and clustering based on affinity as well as satisfying certain constraints, e.g., a clustering should be 4–5 papers. In contrast to the existing session creation process, we consider a community-supported process in which: 1) program committee members create preliminary affinity scores between papers, 2) authors refine paper affinities, and 3) conference organizers fix incoherent sessions (Figure 1). By reaching out to the broader conference community, we leverage the interest and efforts of people with expertise, reduce the burden on organizers, and make more coherent sessions.

In Stage 1, we recruit committee members to group papers in their specific area of expertise over the Internet. We study two community clustering methods—Cascade (Chilton et al. 2013) and partial clustering (Gomes et al. 2011; Strehl and Ghosh 2003)—that embody different approaches for grouping papers and generating affinity data. These paper affinities can be used to generate a filtered list of potentially related papers for authors to judge in the second stage. The author judgments allow us to compare the two committeesourcing methods with an existing manual process (physical paper clustering) and an automated approach (TF-IDF).

In Stage 2, we recruit authors to specify which papers fit (and do not fit) in a session with their paper and which papers they would like to see. These responses help refine the knowledge of paper affinities and also inform which papers of interest should not be scheduled in the same timeslot. Authors should be good at the task because they have the topic expertise and the intrinsic motivation to see their paper land

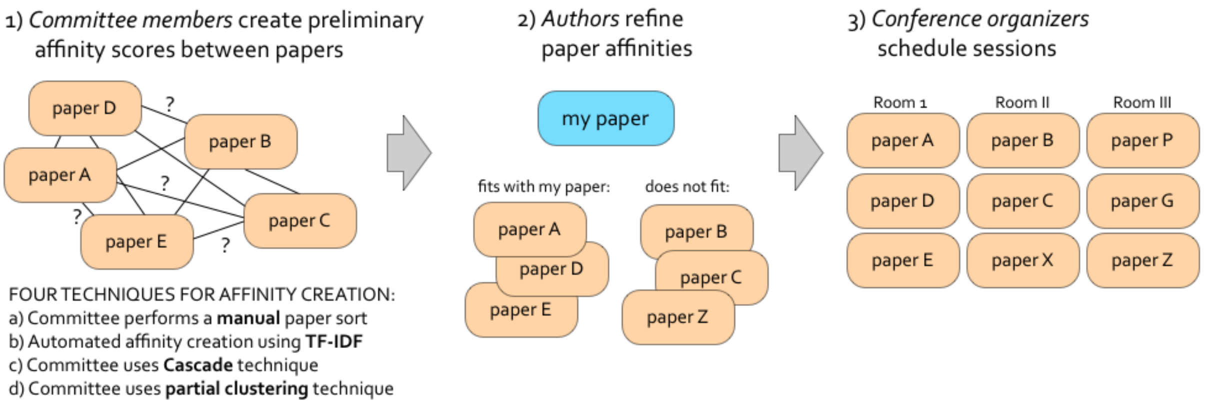

1) Committee members create preliminary affinity scores between papers
paper D
paper B
paper A
paper C
paper E
?
?
?
?
FOUR TECHNIQUES FOR AFFINITY CREATION:
a) Committee performs a manual paper sort
b) Automated affinity creation using TF-IDF
c) Committee uses Cascade technique
d) Committee uses partial clustering technique
2) Authors refine paper affinities
my paper
fits with my paper:
does not fit:
paper A
paper D
paper E
paper B
paper C
paper Z
3) Conference organizers schedule sessions
Room 1
Room II
Room III
paper A
paper B
paper P
paper D
paper C
paper G
paper E
paper X
paper Z

Figure 1: *We introduce a communitysourcing approach to clustering papers and forming sessions for large multi-track conferences.*

in a session with related papers.

In Stage 3, organizers fix incoherent sessions and refine the schedule. Organizers can use affinity data refined by authors to detect papers that do not fit well in a session, and replace them with papers from other sessions that do fit well. The affinity data is used by visualization and intelligent scheduling tools described elsewhere (Kim et al. 2013).

In this paper, we provide a comparison of methods for creating initial affinity scores in Stage 1 and demonstrate the value of authorsourcing in Stage 2 for making coherent sessions in Stage 3. We report on the deployment of our community-supported process for session creation as part of a larger process for planning CHI 2013. Our approach (a) incorporates humans and automation in a structured way; (b) broadens participation beyond prior approaches; and in this paper (c) empirically evaluates existing, automated, and novel crowdsourced approaches. Results show that committeesourcing methods perform at least as well as the manual paper sorting methods, but they take less of each individual's time and provide affinities beyond a single session.

We first discuss related work in clustering and communitysourcing. We describe our communitysourced process, focusing on the design of our clustering methods and the details of our study. We then present results from our deployment at CHI 2013. We address limitations of our study, and conclude with recommendations for conference organizers and thoughts on future work.

Related Work

Clustering papers into sessions is a core part of the scheduling process. Recent machine learning and crowdsourcing work has considered utilizing human cognition for determining object similarities and for clustering. These approaches use adaptive triadic comparisons (Tamuz et al. 2011), aggregate worker annotations from partial clusters of an entire dataset (Gomes et al. 2011), or extend the latter technique using matrix completion to reduce the needed comparisons for partitioning of the entire dataset (Yi et al. 2012). Each process successfully uncovered meaningful categories within the data, although they were tested on images not text. An exception is work by Chilton et al. (2013) on Cascade, which introduces a crowd workflow for creating taxonomies of text datasets such as Quora questions. André et al. (2014) further consider category creation and clustering, focusing on the impact of re-representation techniques prior to a clustering step. Our project extends these by empirically testing partial clustering and Cascade techniques, on a complex qualitative text dataset requiring domain expertise, in which committee members generate affinities between papers.

The traditional, manual process of sorting papers faces limitations in the availability and range of expertise of committee members. This paper considers how online techniques that distribute micro-tasks to an expert crowd might aid in the scheduling process. Communitysourcing approaches have been used successfully in physical spaces for problems such as grading papers (Heimerl et al. 2012) or collecting scientific data (Evans et al. 2005), and online for encouraging contributions and incentivizing participation either through extrinsic or intrinsic means (Kraut and Resnick 2011). In contrast to most other attempts at communitysourcing, we consider a scenario in which community members provide information for solving a specific problem (making conference sessions) whose solution affects themselves and the community at large.

Information retrieval researchers have considered a related problem of assigning papers to reviewers. Papers are automatically assigned, generally based on paper content and reviewer expertise or interest judgments. For example, Dumais & Nielsen (1992) use latent semantic indexing, Hettich & Pazzani (2006) convert NSF proposals to TF-IDF space, and Karimzadehgan et al. (2008) model reviewers and papers with probabilistic latent semantic analysis (PLSA). These expert finding techniques can help inform the problem of grouping papers. Of particular note, TF-IDF may be used to optimize initial matches (Hettich and Pazzani 2006), PLSA is able to extract multiple aspects of a paper (Karimzadehgan, Zhai, and Belford 2008), and papers may contain multiple domains but be best suited to one or the other (Karimzadehgan, Zhai, and Belford 2008; Mimno and McCallum 2007). However, the goals of these techniques are a little different to those of clustering papers and making sessions, where the affinity of an entire group is important (not just pairwise affinity), and other constraints such as session size exist.

A Community-Clustering Process

In order to create coherent sessions of relevance and interest, we seek to capture the affinities among papers. We empirically compare methods for creating affinities in a two-stage process: (1) creating an initial affinity matrix with the committee; and (2) refining those suggestions with author relevance and interest judgments.

Stage 1: Committeesourcing Initial Affinities

We describe four different affinity creation methods: a manual paper clustering method, an automated approach that leverages TF-IDF, and two distributed human computation approaches. We then discuss the advantages and disadvantages of each method.

a) Manual Paper Sort at Committee Meeting. We report on the specifics of the initial session creation process for CHI, although many other large conferences use a similar in-person, manual process. After papers are accepted, a small group of associate chairs help the conference organizers to roughly create categories and suggest sessions. Over two days, the organizers and a few assistants build a rough preliminary schedule. The process is paper-based, collaborative, and time-consuming; its output is highly dependent upon the specific knowledge of the individuals in the room. The CHI committee uses categories or personas to broadly group related papers, this year resulting in 13 personas such as online communities, health, or design. This year, 'groups' of approximately 12 candidate papers were created, and in an ad hoc process, 'sessions' were then created by grouping four to six papers.

b) Automated Affinity Creation using TF-IDF. Conferences have experimented with automatic techniques for tasks requiring knowledge of similarities between papers. For instance, *UIST'12, '13*, and *CHI'13* used TF-IDF (term frequency-inverse document frequency, often used as a measure for scoring search result relevance) to suggest papers for reviewers and to assign people to chair sessions. TF-IDF compares the relative frequency of words in a specific document to the inverse proportion of that word over the entire document corpus. This provides a sense of how relevant the word is in a given document: a term t in document d is given a high weight when the term appears many times in a small number of documents, or a low weight when the term occurs fewer times in a document, or occurs in many documents. Alternative statistical techniques such as topic modeling may be useful, but others have noted that such methods can require significant user input and parameter tweaking (Chuang et al. 2012). For the paper suggestions in this experiment, we computed TF-IDF scores using paper titles, abstracts, and keywords.

c) Committeesourcing with Cascade. Cascade is a crowd workflow that coordinates human labor with automated techniques to create taxonomies (Chilton et al. 2013). The process consists of two human-based steps: generate and categorize. In the generate step, we show contributors a paper title and abstract and ask them to come up with a label (Figure 2). This identifies a set of general and specific concepts that papers can then be grouped into.

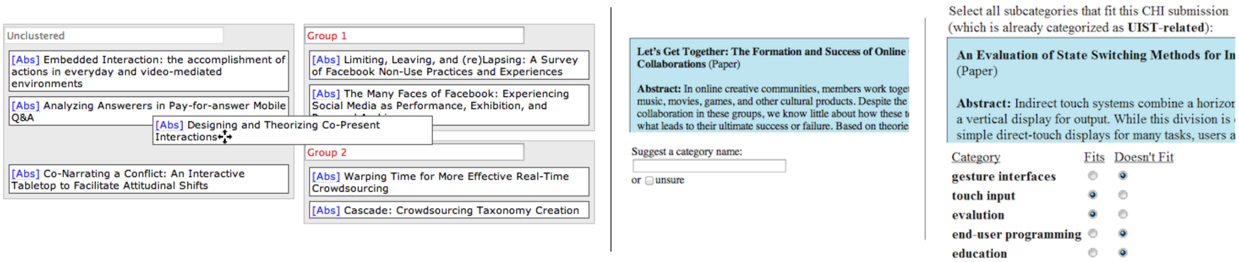
In the second step, contributors categorize papers based on the labels from the first step (Figure 2). For example, if the first step produced a label such as "human computation," all papers that concern human computation will likely get placed in that group. Cascade solves the problem of redundant labels by consolidating any two labels with high overlap in the papers categorized into them. We then eliminate labels that have fewer than three papers. The result is a list of labeled categories with papers where every category is meaningful, sufficiently large, and not redundant with any other category.

d) Committeesourcing with Partial Clustering. Partial clustering is a method of grouping subsets of the entire dataset, with some overlap between subsets in order to infer clustering over the entire dataset. We adapt the method used by Gomes et al. (2011) on images to the CHI text dataset. We use the object distribution algorithm from Strehl and Ghosh (2003) to cluster N items into groups of M, with an overlap of V. Practically, this means items are randomly distributed with some overlap, such that each item appears in V groups. See Figure 2 for an example: papers are grouped into sets of 15, with each paper appearing in 5 other groups. We ask contributors to read 15 paper titles—with abstracts available on hover—and to drag similar items into groups (see Figure 2 left). We construct a similarity matrix by increasing an affinity score each time contributors put papers together in a group. The resulting global similarity matrix can be clustered with any clustering algorithm; we use the hierarchical clustering tool from Fernandez and Gomez (2008).

Comparison of Affinity Creation Methods. The manual technical program committee method (hereafter: TP Meeting), partial clustering, and Cascade all use experts, but differ in cognition, computation, and scale. The main limitation of the manual approach is that not all expertise and viewpoints can be represented during the physical meeting. Anecdotally, organizers describe the entire schedule creation process as "painful" and "painstaking," and conference attendees and paper authors complain of occasional incoherent sessions. Due to the organic nature of how organizers make connections between papers, many sessions have odd papers mixed in, and the process does not capture affinities between papers in different sessions.

TF-IDF is far more scalable than the manual approach and is "free" in terms of time and effort, but lacks human expertise. Since TF-IDF operates on text frequencies, it may need to be augmented with semantic information about papers to produce desirable results. Furthermore, since TF-IDF focuses on pairwise affinities and not on creating sessions, it may be more useful for fixing sessions than used for session making, where the semantic concept behind the grouping may be important.

The committeesourcing approaches seek to leverage the expertise and efforts of community members to scale high-quality affinity creation. Cascade and partial clustering em-



Unclustered

[Abs] Embedded Interaction: the accomplishment of actions in everyday and video-mediated environments

[Abs] Analyzing Answerers in Pay-for-answer Mobile Q&A

[Abs] Designing and Theorizing Co-Present Interactions

[Abs] Co-Narrating a Conflict: An Interactive Tabletop to Facilitate Attitudinal Shifts

Group 1

[Abs] Limiting, Leaving, and (re)Lapsing: A Survey of Facebook Non-Use Practices and Experiences

[Abs] The Many Faces of Facebook: Experiencing Social Media as Performance, Exhibition, and

Group 2

[Abs] Warping Time for More Effective Real-Time Crowdsourcing

[Abs] Cascade: Crowdsourcing Taxonomy Creation

Let's Get Together: The Formation and Success of Online Collaborations (Paper)

Abstract: In online creative communities, members work toge music, movies, games, and other cultural products. Despite the collaboration in these groups, we know little about how these t what leads to their ultimate success or failure. Based on theorie

Suggest a category name:

or □ unsure

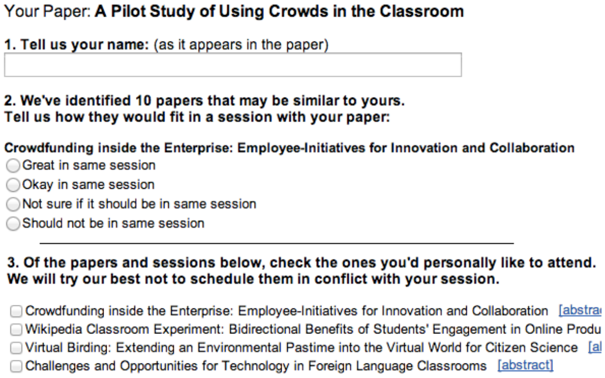
Select all subcategories that fit this CHI submission (which is already categorized as **UIST-related**):

An Evaluation of State Switching Methods for In (Paper)

Abstract: Indirect touch systems combine a horizor a vertical display for output. While this division is simple direct-touch displays for many tasks, users a

Category	Fits	Doesn't Fit
gesture interfaces	○	●
touch input	●	○
evalution	●	○
end-user programming	○	●
education	○	●

Figure 2: *Distributed clustering interfaces, as provided in the committeesourcing stage. Left: partial clustering, grouping a list of items into subgroups. Right: Cascade generate and categorize steps.*



Your Paper: **A Pilot Study of Using Crowds in the Classroom**

1. Tell us your name: (as it appears in the paper)

2. We've identified 10 papers that may be similar to yours. Tell us how they would fit in a session with your paper:

Crowdfunding inside the Enterprise: Employee-Initiatives for Innovation and Collaboration

○ Great in same session
○ Okay in same session
○ Not sure if it should be in same session
○ Should not be in same session

3. Of the papers and sessions below, check the ones you'd personally like to attend. We will try our best not to schedule them in conflict with your session.

□ Crowdfunding inside the Enterprise: Employee-Initiatives for Innovation and Collaboration [abstra
□ Wikipedia Classroom Experiment: Bidirectional Benefits of Students' Engagement in Online Produ
□ Virtual Birding: Extending an Environmental Pastime into the Virtual World for Citizen Science [al
□ Challenges and Opportunities for Technology in Foreign Language Classrooms [abstract]

Figure 3: *Authors are presented with a list of suggestions generated from the four committeesourcing stages, and asked to rate relevance and interest.*

body different approaches. In Cascade, the generate step is a *classification* method of category learning, presenting a single item and extracting limited features to predict an individual category label (Medin and Schaffer 1978). This may be particularly effective if experts can draw on their existing mental model and prior knowledge of a paper's area.

An alternative cognitive approach is to infer a single label for a group of items. The partial clustering technique asks a worker to draw connections between items by creating a single label for a group, thereby explicitly enforcing *mapping and comparison*, techniques which have been shown to facilitate schema induction (Gick and Holyoak 1983). Additionally, making interproperty relationships more salient may result in more nuanced categories (Lassaline and Murphy 1996). However, partial clustering may miss some groupings if items are rarely or never presented together. And while Cascade can decompose tasks down to a single action at the level of an individual paper, partial clustering depends on presenting multiple papers at once.

We compare these four methods in our deployed study. We hypothesize that committeesourcing approaches will provide affinities as accurate as an in person meeting, but with less effort from each individual.

Stage 2: Authorsourcing Refined Paper Relevance

The committeesourcing methods create varying sized groups or lists of potentially related papers. In the authorsourcing stage (see Figure 3), we present a list of papers to each paper author and ask two questions: how *relevant* is each paper (i.e., should it be in the same session as the author's paper); and is the paper *interesting* (i.e., would the author like to see this paper's presentation). The relevance feedback provides fine-grained information about which papers should appear in the same session. By showing papers suggested as relevant by each of the four methods from Stage 1, the judgments can further be used as evaluation of methods for generating initial affinities.

Incentivizing Participation

In order to effectively recruit and engage community members in committeesourcing and authorsourcing tasks, we draw on their particular interests and motivations.

Committeesourcing. A few days after CHI's physical TP meeting, we asked committee members to help group papers in their area of expertise. We constrained tasks from the two methods to be short (approximately 10 minutes long), with the option to do multiple tasks. To encourage repeated participation, we provided extrinsic incentives in the form of global and per persona leaderboards.

Authorsourcing. We hypothesize that authors are motivated to participate so that their paper may end up in a session with relevant papers. The task also provides them with an advance preview of the accepted papers before the program is announced. Additionally, asking authors for their feedback may provide them with a sense of ownership or involvement in the process.

Method

Our study evaluates four methods for creating preliminary affinity scores between papers: in-person clustering at the TP meeting, an automated TF-IDF method, and two distributed human clustering mechanisms. In addition, we evaluate the value of the communitysourcing process on planning the conference program.

Study Design

To support committee members in creating affinities, we allow them to select a topic area in their area of expertise so that they are only presented with papers assigned to that persona (e.g., Health, Input/Output, Security & Privacy). This simplifies the task for committee members: instead of working with a set of 450 mostly unrelated papers, they can work within a subset of $\sim$40 somewhat related papers.

We use as topics the personas generated at the TP meeting. While this helps to reduce the search space and allows participants to perform a task suited to their expertise, one drawback to this approach is that the communitysourcing results are dependent on the accuracy of the initial persona groupings. Later we discuss how this process could be further evaluated and improved.

Since we could not predict the level of participation, we chose to focus on seven of the more popular topic areas so as to concentrate effort. Committee members experienced the following workflow: They click on a URL provided in e-mail from conference chairs and see a landing page showing a description of the task, a current leaderboard, and a choice of topic areas. Once they choose a topic area, they were randomly assigned (and balanced to maintain equal participation) to one of two conditions: partial clustering or Cascade. Once they finish the designated tasks, they see a thank you page, again with a leaderboard and an option to do more tasks. Each task required around 10 minutes of work.

Measures: Evaluation via Authorsourcing

Each of the four techniques for affinity creation provides a list of the most relevant papers for a given seed paper. For each accepted paper, we generated a list of suggested papers and asked authors to indicate how well their paper fits in a session with each of the suggested papers (the options are great, ok, not sure, and not ok). Authors were provided an initial list of ten papers, with an option to view ten more. We balanced the list to include papers suggested by all four methods of clustering. We construct the list of ten papers as follows: 3–5 papers in the set as generated manually in the committee meeting, 3–4 papers from either the partial clustering or Cascade techniques (when data is available), and 2–3 papers suggested by TF-IDF (weighted towards highest ranking).

Results

We first present participation statistics for each stage of the communitysourcing process and then examine the relative differences between the different affinity creation methods.

Participation

From January 6 to 18, 2013, associate chairs performed committeesourcing tasks. Sixty-four of 211 associate chairs participated. Their work created affinities for 1722 pairs of related papers; a breakdown of participation by technique is in Table 1. Since affinities were not captured completely for most topics, we were only able to compute suggestions for certain papers and personas for use in authorsourcing. We revisit this limitation later in the paper.

From January 29 to February 12, 2013, authors participated in refining the affinity data; 654 authors provided 7095 judgments of how well papers fit in a session with theirs. Despite the recruitment email being sent to only contact authors, authors of 87% of the accepted papers contributed data, with an average of 1.3 authors participating per paper.

	Partial Cluster	**Cascade**
Participants	29	35
Avg. time per task	9.0 min	10.5 min
Total time all tasks	4.3 hours	10.0 hours

Table 1: *Participation statistics for committeesourcing.*

Relevance of Suggestions

We map the authors' judgments from "not okay in same session" to "great in same session" onto a 1 to 4 scale. We consider the average relevance of the top 10 suggestions from each method (see Table 2 for full details). First, we restrict to the 7 personas for which we experimentally tested the distributed clustering methods (since not limiting to those 7 may unfairly penalize TF-IDF if the excluded personas had uncharacteristically irrelevant papers). Since the human clustering methods were limited to suggesting papers within a specific persona, we also calculated a TF-IDF score that restricted to only papers within a persona rather than looking globally, we name this TF-IDF-Persona.

A one-way between-subjects ANOVA was conducted to compare the effect of affinity method on relevance of suggested papers. Relevance differed significantly across conditions, $F(4{,}5508)=30.42$, $p<.001$. Bonferroni post-hoc comparisons indicate that partial clustering and TF-IDF-Persona were borderline similar (p=.05), and that both methods differed significantly from TP Meeting, Cascade, and TF-IDF, which were all comparable to each other. This demonstrates that one of the distributed techniques was able to outperform the existing meeting, and limiting TF-IDF to within persona was similarly successful.

We also consider the top 10 suggestions averaged across all tracks. We see similar results, but with a more dramatic reduction in TF-IDF-Persona, indicating that one of the tracks was likely particularly hard to suggest relevant papers for (upon investigation this was the 'Miscellaneous' track).

The prior analyses do not account for the specific rank at which a paper is suggested. It may be informative to penalize a method if a highly relevant paper appears at the bottom of a list of suggestions. Similar to traditional information retrieval evaluation metrics, we use discounted cumulative gain (DCG) to treat each author's paper as if it were a query into the set of all other papers. Results show that when taking order into account, results from TF-IDF-Persona and partial clustering again outperform other methods, and TF-IDF and Cascade are considered to be slightly better than results from the TP-meeting.

It is possible that the two distributed clustering methods suffer when averaging across personas due to incomplete data. To investigate this issue, we look to two personas that had almost full completion: Visualization for Cascade, and Health for partial clustering (see Table 2). We do see an increase in scores for the distributed methods (particularly noticeable for Cascade in Visualization). However, other methods also improve, suggesting that these tracks may have had inherently more relevant candidate papers.

	Avg. Rel. (7 personas)	SE	Avg. Rel. (all personas)	SE	DCG (7 personas)	Visualization persona (SE)	Health persona (SE)
TF-IDF-Persona	2.95	0.04	2.86	0.03	12.88	2.91 (0.12)	3.10 (0.08)
Partial Cluster	2.79	0.04	—	—	12.28	2.82 (0.12)	2.97 (0.07)
TP Meeting	2.57	0.03	2.55	0.02	11.12	2.77 (0.11)	2.62 (0.07)
TF-IDF	2.51	0.03	2.49	0.02	11.35	2.63 (0.09)	2.62 (0.07)
Cascade	2.49	0.05	—	—	11.38	2.68 (0.16)	2.49 (0.11)

Table 2: *Relevance of papers suggested by the different affinity creation methods (rated from 1 to 4, higher is more relevant). We present results for the seven personas included in the committeesourcing experiment and across all personas. TF-IDF constrained to within a persona and partial clustering outperform all other methods (posthoc results in text). We calculate Discounted Cumulative Gain (DCG) to give a sense for graded relevance performance, and show the average relevance scores for the two specific personas in which the distributed methods had more complete data.*

Quantity and Use of Affinities Discovered

An advantage of communitysourcing approaches is in discovering affinities among papers that are missed at the in-person meeting. We found that out of 226 suggestions made by partial clustering that were not present in data from the in-person meeting, 43 papers were judged by authors as great in a session with their paper (and 44 ok fits). Cascade contributed 69 additional great fits (and 113 ok fits), and TF-IDF another 344 great fits (and 420 ok fits). These identified relevant papers provide flexibility during scheduling for resolving conflicts that would otherwise have been difficult using only information from the in-person meeting.

Note that the numbers above are not directly comparable due to different participation levels, and TF-IDF is not limited by participation. The in-person meeting did identify larger groups than just the final 4–5 papers in a session (groups of approximately 12 candidate papers), and while it did also generate other relationships that were not captured, these were generally discarded negative possibilities, i.e., papers that did not fit together and were moved.

Papers judged by authors to be poor fits in a session suggest problems in the initial schedule that require the organizers' attention. We found that 129 pairs of papers within manually created sessions at the in-person meeting were judged by two or more authors as poor fits in the same session (out of a possible 688 pairs, or 19%). While refining the schedule in Stage 3, the organizers used the authorsourcing data to visualize and fix conflicts, resolving 87 of the 129 poor-fit conflicts (Kim et al. 2013).

Discussion

To better understand the performance of each method and the perceptions of relevance provided by an author, we first compare methods aimed at creating clusters of papers and then discuss the performance of TF-IDF, which focused more on pairwise affinity creation than session-creation.

Comparing Clustering Methods

The TP meeting, partial clustering, and Cascade all focus on creating groups of related papers and capturing affinities among papers as a side effect of clustering. As we had hypothesized, we found that reaching out to a broader set of committee members beyond organizers, and providing the ability to place papers in multiple groups during committeesourced clustering, produced relevant groupings not previously discovered. Both distributed methods at least matched the relevance of results from the in-person meeting, while partial clustering outperformed it. We hypothesize that seeing a set of papers in partial clustering may have assisted committee members in producing better clusters, but more experimentation and data is required to understand contributing factors.

Our results suggest that unlike current practice, a distributed approach can both save time and effort and provide high quality data that contains affinities beyond sessions. With even more contributions from the community, the distributed methods have the potential to provide more suggestions of higher quality by using more complete data to infer affinities.

Examples of Relevant and Irrelevant Suggestions

Committeesourcing methods were able to discover matches judged to be highly relevant by authors that were not grouped together in the in-person meeting. For example, partial clustering grouped "*Warping Time for More Effective Real-Time Crowdsourcing*" with "*A Pilot Study of Using Crowds in the Classroom*," and Cascade grouped "*Patina: Dynamic Heatmaps for Visualizing Application Usage*" with "*Quantity Estimation in Visualizations of Tagged Text*."

Conditioned on a suggestion being considered relevant by an author, it is hard to distinguish qualitative differences among the methods. The suggested paper tends to fit in terms of domain or methodology with the source paper. When a suggestion was considered not at all relevant, it was generally because of a potential but spurious connection, e.g., social in "*Revisiting Social Practices Surrounding Music*" & "*Write Here, Write Now!: An Experimental Study of Group Maintenance in Collaborative Writing*"; or a lack of any connection, e.g., "*Shifting Dynamics or Breaking Sacred Traditions? The Role of Technology in Twelve-Step Fellowships*" & "*Understanding the Privacy-Personalization Dilemma for Web Search: A User Perspective*."

Comments from authors reveal some of the more nuanced reasons for disagreeing with suggested papers. Some authors considered their work at the intersection of two topics, but only saw suggestions for one of those two. One author wrote that their "*paper is about the intersection of touch and visual analytics. All the papers listed above were about touch in-*

teraction, but none were about visual analytics which might also be a good fit." Others considered the difference between domain and methodology. "*While our paper does take [domain] as a case study, our central argument is not specific to this context. ...Our talk would be better suited to a session taking a critical or feminist perspective on research and design*." Others seemed to be about the misinterpretation of the focus of a paper, e.g., "*I think our paper would better fit in a session about online security than search*." We found that these issues were present regardless of the method that provided suggestions.

Role and Performance of TF-IDF

Given that TF-IDF lacks any semantics, it was at first surprising to see that relevance scores for TF-IDF were as high as for the current manual method and furthermore, when limited to suggestions with a manually created persona, TF-IDF outperformed existing methods and is on a par with partial clustering. But since TF-IDF focuses only on affinities, it is not limited from a pairwise affinities perspective as TP meeting, partial clustering, and Cascade are. The latter clustering methods seek to create semantically relevant groups of mutually relevant papers, which is needed for session-making and not provided by TF-IDF.

The CHI conference may also be a special case in that it has a wide variety of content for which the personas almost create sub-conferences within the wider conference. Since TF-IDF recommended papers within a persona performed significantly better than TF-IDF computed over the entire corpus, this suggests that a human step to attach semantic information is still necessary. But since TF-IDF does not require human effort and can generate affinity data over the entire set of papers, it can suggest potential matches beyond the persona and present suggestions to authors and organizers even when other methods lack data.

Finally, unlike committee members who may have theoretical or political lenses or biases, TF-IDF is atheoretical. It works directly on the words that authors used, and so may be closer aligned to an authors' mental model. Committee members may also be conceptually reframing a paper, even despite specific language used, or using a more global or nuanced view to attempt to create a particular thematic session while an author may think their paper aligns with a different theme. Neither is incorrect, and it may require organizers to resolve differences in making final decisions on sessions.

Limitations

Our distributed clustering techniques were deployed on personas developed at the committee meeting. While helping to focus committee members' efforts, limiting to a persona grouping may have also removed papers that an "ideal" process would have otherwise suggested. Other methods of coarse grouping or more computationally efficient implementations of community-clustering methods than can consider the entire corpus may potentially provide better results.

Our deployment of distributed techniques came at a time that the committee would traditionally have been done with their conference duties. As mentioned, we did not receive enough participation to complete our clustering and recommendations were based on incomplete data.

Our community clustering process is part of a larger process of session making and scheduling. Our analysis does not consider high-level goals of organizers beyond affinity that may influence or suggest tweaks in how we collect affinity data from the community.

Implications & Future Work

Based on our findings, we make a couple of suggestions for how conference organizers can draw on the community for making coherent sessions:

Engage All Committee Members in Clustering

We have demonstrated that a distributed approach such as partial clustering can outperform manual clustering, reduce time needed to create those clusters, and provide affinities beyond sessions. Even with the relatively low participation of committee members—30% of ACs—our results show value in the techniques and committee participation, though further data would provide more alternatives in scheduling. The participation level suggests that, in order to generate richer affinity data, organizers may wish to integrate some form of distributed clustering during the existing committee meeting, say for an hour at the end of the day.

Authorsourcing Provides Valuable Data

Authorsourcing provides fine-grained affinity data on how papers can be grouped into coherent sessions. The authorsourcing stage saw significant participation, with authors of 87% of all papers represented. Authors are interested in seeing their paper in a session of related papers. Many authors thanked us for the opportunity to engage and take ownership in the process. Some even wished for more control, such as suggesting other potential categories or seeing more papers in their area of interest.

Our findings also point to areas for future work in community-supported processes for session-creation and scheduling:

Authorsourcing and TF-IDF

We have shown that by restricting to a broad group of human-clustered papers, TF-IDF can suggest highly relevant papers. It is possible that combining TF-IDF with authorsourcing may provide a rich enough affinity matrix that can be used for clustering papers into sessions. However, committeesourcing methods capture semantic information that TF-IDF does not, and may serve as an alternative for session-creation while TF-IDF cannot. Testing more sophisticated techniques such as Explicit Semantic Analysis which may better deal with semantic limitations can help to uncover papers that are similar but do not share terms. We are interested in seeing how automated methods may be combined with authorsourcing to produce richer affinity data with less human effort.

Session Creation and Workflow

While we found value in the distributed clustering techniques, future work may also prioritize satisfying constraints such as creating sessions of appropriate size in addition to uncovering relevant papers. The current workflow has the committee and organizers perform the session and schedule creation tasks. Alternative workflows may capitalize on the interest (and potentially greater participation) of the authors to provide as much information as possible prior to a final organizer approval; though authors can most likely only be involved once final accept/reject decisions have been made. We have also considered extending the tools (in Stage 3) to enable community-wide session making and conflict resolution within an entire schedule.

Aligning Community Incentives

Involving community members in planning a conference requires aligning their incentives with methods that elicit useful data. In addition to committee members and authors, we have also experimented with means of collecting data from all attendees. With a web application, users can bookmark papers and receive social recommendations on other papers they may be interested in that they can then add to their personal schedule. By helping attendees decide on where to spend their time during a conference, we are also collecting data about users' interests that can be used to group papers of mutual interest, place related sessions in different time slots, and schedule popular talks in larger rooms.

Acknowledgments

We thank Wendy Mackay, Michel Beaudouin-Lafon, Patrick Baudisch, and the many committee members and authors at CHI 2013 for their assistance with the study. This work is supported in part by Quanta Computer as part of the Qmulus project, by the Ford-MIT Alliance, and by NSF under awards SOCS-1111124 and IIS-1208382. Juho Kim is partly supported by the Samsung Fellowship.

References

André, P.; Kittur, A.; and Dow, S. P. 2014. Crowd synthesis: Extracting categories and clusters from complex data. In *Proc. CSCW 2014*. ACM.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Chilton, L. B.; Little, G.; Edge, D.; Weld, D.; and Landay, J. 2013. Cascade: Crowdsourcing taxonomy creation. In *CHI 2013*.

Chuang, J.; Ramage, D.; Manning, C.; and Heer, J. 2012. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proc. CHI 2012*, 443–452. ACM.

Dumais, S. T., and Nielsen, J. 1992. Automating the assignment of submitted manuscripts to reviewers. In *Proc. SIGIR 1992*, 233–244. ACM.

Evans, C.; Abrams, E.; Reitsma, R.; Roux, K.; Salmonsen, L.; and Marra, P. P. 2005. The neighborhood nestwatch program: Participant outcomes of a citizen-science ecological research project. *Conservation Biology* 19(3):589–594.

Fernández, A., and Gómez, S. 2008. Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *Journal of Classification* 25(1):43–65.

Gick, M., and Holyoak, K. 1983. Schema induction and analogical transfer. *Cognitive psychology* 15(1):1–38.

Gomes, R.; Welinder, P.; Krause, A.; and Perona, P. 2011. Crowdclustering. In *Advances in Neural Information Processing Systems (NIPS 2011)*.

Heimerl, K.; Gawalt, B.; Chen, K.; Parikh, T.; and Hartmann, B. 2012. Communitysourcing: engaging local crowds to perform expert work via physical kiosks. In *Proc. CHI 2012*, 1539–1548. ACM.

Hettich, S., and Pazzani, M. J. 2006. Mining for proposal reviewers: lessons learned at the national science foundation. In *Proc. KDD 2006*, 862–871. ACM.

Karimzadehgan, M.; Zhai, C.; and Belford, G. 2008. Multi-aspect expertise matching for review assignment. In *Proc. CIKM 2008*, 1113–1122. ACM.

Kim, J.; Zhang, H.; André, P.; Chilton, L. B.; Mackay, W.; Beaudouin-Lafon, M.; Miller, R. C.; and Dow, S. P. 2013. Cobi: A community-informed conference scheduling tool. In *Proc. UIST 2013*.

Kraut, R., and Resnick, P. 2011. *Evidence-based social design: Mining the social sciences to build online communities*. MIT Press.

Lassaline, M., and Murphy, G. 1996. Induction and category coherence. *Psychonomic Bulletin & Review* 3(1):95–99.

Medin, D. L., and Schaffer, M. M. 1978. Context theory of classification learning. *Psychological review* 85(3):207.

Mimno, D., and McCallum, A. 2007. Expertise modeling for matching papers with reviewers. In *Proc. KDD 2007*, 500–509. ACM.

Salton, G., and McGill, M. J. 1983. *Introduction to Moderm Information Retrieval*. McGraw-Hill.

Strehl, A., and Ghosh, J. 2003. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research* 3:583–617.

Tamuz, O.; Liu, C.; Belongie, S.; Shamir, O.; and Kalai, A. 2011. Adaptively learning the crowd kernel. In *International Conference on Machine Learning (ICML)*.

Yi, J.; Jin, R.; Jain, A.; and Jain, S. 2012. Crowdclustering with sparse pairwise labels: A matrix completion approach. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.