Crowdlines: Supporting Synthesis of Diverse Information Sources through Crowdsourced Outlines

Kurt Luther[1,2], Nathan Hahn[2], Steven P. Dow[2], Aniket Kittur[2]

[1]Center for Human-Computer Interaction, Virginia Tech
kluther@vt.edu

[2]Human-Computer Interaction Institute, Carnegie Mellon University
{nhahn, spdow, nkittur}@cs.cmu.edu

Abstract

Learning about a new area of knowledge is challenging for novices partly because they are not yet aware of which topics are most important. The Internet contains a wealth of information for learning the underlying structure of a domain, but relevant sources often have diverse structures and emphases, making it hard to discern what is widely considered essential knowledge vs. what is idiosyncratic. Crowdsourcing offers a potential solution because humans are skilled at evaluating high-level structure, but most crowd micro-tasks provide limited context and time. To address these challenges, we present Crowdlines, a system that uses crowdsourcing to help people synthesize diverse online information. Crowdworkers make connections across sources to produce a rich outline that surfaces diverse perspectives within important topics. We evaluate Crowdlines with two experiments. The first experiment shows that a high context, low structure interface helps crowdworkers perform faster, higher quality synthesis, while the second experiment shows that a tournament-style (parallelized) crowd workflow produces faster, higher quality, more diverse outlines than a linear (serial/iterative) workflow.

Introduction

Learning the structure of an information space, including the key factors, dimensions, and organizational schemas, is fundamental to information seeking problems ranging from shopping for a new camera, to deciding whether to invest in Bitcoin, to developing a course curriculum. Understanding the deep structure and important factors defining a space—e.g., learning that megapixels matter less today for choosing a camera than lens quality or ergonomics; or that soil acidity and temperature zones are key factors in choosing plants for a garden; or that memory is a more core topic than personality for an introductory psychology course—may be more valuable than specific facts and more generalizable across people with differing goals and expertise (Fisher, Counts, and Kittur 2012).

However, learning the structure of an information space has become a daunting proposition for a novice as the information available online continues to grow exponentially (Abbott 1999). Some information will overlap across sources, some will contradict, and some will be unique. This diversity represents one of the great strengths of the Internet, giving people access to an incredibly rich wealth of perspectives and experiences. Yet, taking advantage of this diversity presents significant challenges, as people must draw connections between this information in order to learn from it and make decisions. Doing such synthesis requires deep knowledge of the domain, and can be difficult even for experts (Abbott 1999).

One potential solution to this problem is to take advantage of expert-generated sources of structure that already exist. Such sources can range from the table of contents in digitized books, to sections of review articles, to online course syllabi. These sources already encode the structure of the information space as perceived by an expert, and being able to leverage them could dramatically bootstrap novices' learning and sensemaking.

However, there are a number of challenges that make it difficult to directly use structured sources such as syllabi or review articles. Even expert-authored sources rarely share a single canonical view of what the structure should be. There may be conflicting views in the field that give rise to differing perspectives; there may be new discoveries over time that shift the prevailing views; or there may be differing goals for sources that alter the focus and framing of the structure, e.g. the incentive for a review article to contribute a different framing from others that have come before.

In this paper, we present a system, Crowdlines, that leverages crowdsourcing methods to synthesize structures that experts have already generated, in order to develop a common structure that better describes the aggregate view of an information space. Crowdsourcing undergirds our approach because human intelligence is often more effective at synthesizing diverse information sources than automated approaches (André et al. 2013). Our design goals include keeping the provenance of sources as they are integrated, surfacing the distribution of sources that agreed on particular structures, and doing so in a way that can be made sense of by novices. Prior studies of crowd synthesis in several domains have separately emphasized the tradeoffs between structure and context (e.g. André et al. 2013; Chilton et al. 2014) and serial and parallel workflows (e.g. André, Kraut, and Kittur 2014; Little et al. 2010). We systematically explore the effectiveness of different amounts of context and structure, and different workflows, within one domain to distill broader principles.

We report on two experiments justifying the design of Crowdlines and demonstrating its usefulness to searchers. Experiment 1 (N=153) compared four synthesis interfaces for a single crowdworker, representing different combinations of context and structure. We found that a High context, low structure (HCLS) interface led to significantly higher quality, faster completion times, and higher completion rates. We next investigated two workflows for synthesizing information across multiple crowdworkers, linear and tournament, and found that the tournament workflow was 2–3 times faster. Experiment 2 (N=115) compared Crowdlines to a typical web search for a complex information synthesis task: designing a course syllabus. We found that Crowdlines helped participants develop syllabi significantly more similar to experts and with more sources, compared to web search alone. These experiments informed the design of the Crowdlines system and produced generalizable insights about effective interface and workflow mechanisms for crowd synthesis.

Related Work

Crowdsourced Synthesis and Sensemaking

Researchers have explored the value of using crowdsourcing, either alone or combined with automated techniques, to synthesize information with diverse or unknown schemas. One fruitful approach has been to blend crowdsourcing with ML algorithms. Partial clustering (Gomes et al. 2011; Yi et al. 2012) and crowd kernel (Tamuz et al. 2011) are two such examples, but their application domain is imagery rather than documents, and they focus on low context merges between pairs or triplets of items. We too evaluate pairwise merges in Experiment 1 but we also compare higher context interfaces.

Other crowdsourcing research explores higher-context clustering. Cascade (Chilton et al. 2013) produces crowdsourced taxonomies of hierarchical data sets by letting workers generate, and later select, multiple categories per item. Frenzy (Chilton et al. 2014) is a web-based collaborative session organizer that elicits paper metadata by letting crowdworkers group papers into sessions using a synchronous clustering tool. We draw design inspiration from these projects, particularly the notion of integrating microtasks into more collaborative, unstructured interfaces like Frenzy and other forms of crowdware (Zhang et al. 2012). We build on this earlier work by evaluating the benefits of these clustering-style interfaces compared to other interfaces and workflows, and in a new domain.

Researchers have also studied how much context to provide crowdworkers during clustering tasks. Willett et al. (2013) developed color clustering with representative sampling for reducing redundancy and capturing provenance during crowdsourced data analysis, comparing this to a pairwise "distributed clustering" approach. André et al. (2013) compared automated clustering (TF-IDF), Cascade (Chilton et al. 2013), and crowdsourced partial clustering adapted from Gomes et al. (2011), finding that all three methods could outperform collocated experts in developing conference paper sessions. In other work, André, Kittur, and Dow (2014) experimented with giving crowdworkers different amounts of context prior to clustering Wikipedia barnstars. We expand on these studies by investigating a higher upper bound for context, its interaction with task structure, and synthesis across multiple documents.

Crowdlines also differs from many of the above approaches in its focus on leveraging the existing schemas embedded in many online materials, such as article headings or class topics in a syllabus, rather than generating and clustering bottom-up schemas. Crowdlines signals the importance of a topic through its representation across diverse sources while also allowing searchers to drill down on how those sources address the same topic in different ways.

Prior research has considered the tradeoffs of iterative vs. parallelized crowd workflows, but results are mixed. Some studies (André, Kraut, and Kittur 2014; Little et al. 2010) found iterative workflows superior to parallelized or simultaneous ones, while others (Chilton et al. 2013) ruled out iteration due to negative preliminary data. Building on this work, we ran Experiment 2 to quantify the benefits of iterative vs. parallelized workflows for crowd synthesis.

Searchers engaged in sensemaking tasks develop better mental models when they have access to previous searchers' schemas (Fisher, Counts, and Kittur 2012; Kittur et al. 2014). We report on several experiments to identify the most effective ways to develop these schemas using crowdsourcing.

Ontology Alignment

Computer scientists in the databases research community have long wrestled with the challenge of making connections between data with diverse schemas. One productive thread of research has built tools for humans to reconcile differences in database structures, known as *ontology alignment* or *ontology mapping* tools (Choi, Song, and Han 2006; Falconer and Noy 2011). While these tools are effective and widely used, they are designed for expert database administrators. The Crowdlines synthesis interface draws inspiration from the two-column layout used by many ontology alignment tools, but unlike these, it is designed for non-expert crowdworkers.

Another thread of databases research investigates the value of crowdsourcing for performing ontology alignment. For example, CrowdMap (Sarasua, Simperl, and Noy 2012) generates pairwise microtasks to compare two database schemas, focusing on relationships between table fields. Unlike CrowdMap and most ontology alignment tools, Crowdlines leverages additional context to map deeper, more abstract relationships (e.g. not "ArticleTitle" vs. "ArticlePublisher" but the subject matter of articles). We took inspiration from CrowdMap's pairwise comparison interface when designing Crowdlines' synthesis interface, but as we will see, our evaluations showed higher-context approaches to be more effective.

Automated Clustering

A large body of work in the machine learning (ML) and natural language processing (NLP) research communities examines how documents can be automatically clustered based on semantic similarity (Salton and McGill 1983). For example, Rathod and Cassel (2013) use an ML classifier to identify computer science course syllabi from the web. However, classifiers like this require thousands of domain-specific examples for training data and their granularity is generally coarse (document- or primary topic-level). Crowdlines creates numerous deep subtopic-level connections across sources and does not need training data.

Metro Maps (Shahaf, Guestrin, and Horvitz 2012) link documents, such as news articles or scientific papers, by relevance and time, generating visual timelines of information that balance coherence and coverage. However, like many NLP-based clustering approaches, this technique is not able to generate meaningful labels for the clusters that characterize the relationships between documents. Crowdlines not only makes connections across documents, but also meaningful labels in the form of topics and subtopics.

The Crowdlines System

Our vision of Crowdlines is a web-based system that uses paid crowd workers to generate a rich outline of a user-defined knowledge area drawn from online information sources. The user begins by specifying a high-level topic he or she wants to learn more about, such as mountain biking, smart watches, or Shakespeare's plays. Dozens of crowdworkers are then dispatched to find relevant online sources and merge them using a *synthesis interface* (Figure 1) This merge process leverages human intelligence coordinated through a crowdsourcing workflow (Figure 2) to reconcile differing schemas, allowing the most important topics to emerge while preserving nuances and divergent perspectives. The user then uses a *search interface* (Figure 3) to explore the crowd-generated outline. He or she can quickly identify key topics and examine how each source engages with them; less common topics can also be found.

Our development process for Crowdlines was informed by experimentation at each major step. First, we prototyped interfaces for a single worker to merge two sources (Experiment 1). Next, we developed two workflows, linear and tournament, for multiple crowdworkers to merge many sources. Finally, we compared how participants synthesize information using Crowdlines against typical baselines like web search (Experiment 2). In the following sections, we begin by describing our design goals and prototype development, and then report on our evaluation of that component of the Crowdlines system.

This paper presents both an empirical study of the contextual and structural factors affecting quality and efficiency of the synthesis process, and a set of interfaces demonstrating the value of that process to end users. We propose that examining and evaluating the entire information synthesis pipeline from generation to consumption provides more value than a single piece.

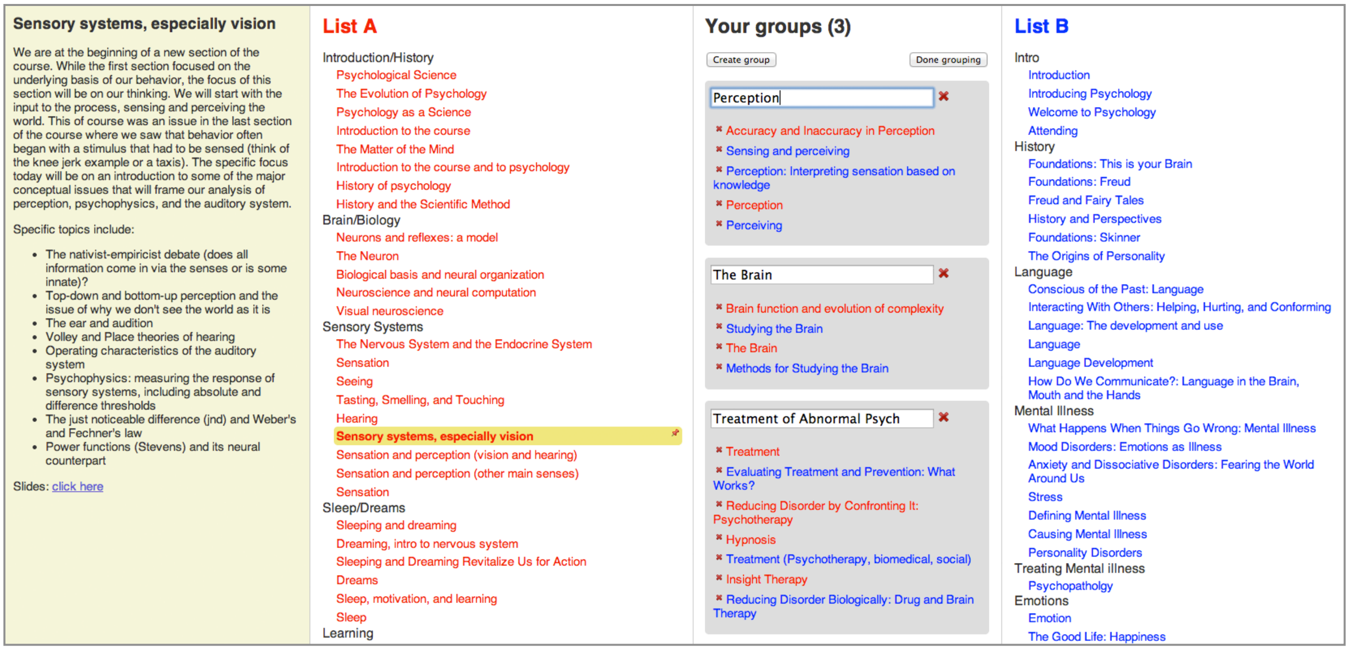
Experiment 1: Evaluating Crowd Synthesis

Designing the Crowd Synthesis Interface

The research on crowd synthesis and clustering, ontology alignment, and crowd-augmented databases provided rich inspiration for our system design. As we considered previous work and new possibilities, two dimensions emerged as especially salient: context and structure. *Context* refers to how much information workers should be exposed to when performing merges (e.g. merging 10 pairs of topics vs. 20 topics all at once). On the one hand, greater context allows workers to consider broader possibilities for connections and relatedness. On the other hand, too much context can become overwhelming, especially for novices.

The second dimension, *structure*, refers to how much guidance the system provides to workers performing merges (e.g. requiring workers to review topics in a particular sequence). A sufficient amount of structure helps direct workers towards their goals and can help them consider possible relationships in a more systematic relationship. Too little structure can leave workers confused or encourage low effort contributions, while too much can leave workers feeling bogged down.

Our goal in Experiment 1 was to identify the most effective combination of context and structure. We designed four variations of a merging interface for crowd workers. All four interfaces are designed to help workers merge two



Sensory systems, especially vision

We are at the beginning of a new section of the course. While the first section focused on the underlying basis of our behavior, the focus of this section will be on our thinking. We will start with the input to the process, sensing and perceiving the world. This of course was an issue in the last section of the course where we saw that behavior often began with a stimulus that had to be sensed (think of the knee jerk example or a taxis). The specific focus today will be on an introduction to some of the major conceptual issues that will frame our analysis of perception, psychophysics, and the auditory system.

Specific topics include:

- The nativist-empiricist debate (does all information come in via the senses or is some innate)?
- Top-down and bottom-up perception and the issue of why we don't see the world as it is
- The ear and audition
- Volley and Place theories of hearing
- Operating characteristics of the auditory system
- Psychophysics: measuring the response of sensory systems, including absolute and difference thresholds
- The just noticeable difference (jnd) and Weber's and Fechner's law
- Power functions (Stevens) and its neural counterpart

Slides: click here

List A

Introduction/History
- Psychological Science
- The Evolution of Psychology
- Psychology as a Science
- Introduction to the course
- The Matter of the Mind
- Introduction to the course and to psychology
- History of psychology
- History and the Scientific Method

Brain/Biology
- Neurons and reflexes: a model
- The Neuron
- Biological basis and neural organization
- Neuroscience and neural computation
- Visual neuroscience

Sensory Systems
- The Nervous System and the Endocrine System
- Sensation
- Seeing
- Tasting, Smelling, and Touching
- Hearing
- **Sensory systems, especially vision**
- Sensation and perception (vision and hearing)
- Sensation and perception (other main senses)
- Sensation

Sleep/Dreams
- Sleeping and dreaming
- Dreaming, intro to nervous system
- Sleeping and Dreaming Revitalize Us for Action
- Dreams
- Sleep, motivation, and learning
- Sleep

Learning

Your groups (3)

Perception ✖
- ✖ Accuracy and Inaccuracy in Perception
- ✖ Sensing and perceiving
- ✖ Perception: Interpreting sensation based on knowledge
- ✖ Perception
- ✖ Perceiving

The Brain ✖
- ✖ Brain function and evolution of complexity
- ✖ Studying the Brain
- ✖ The Brain
- ✖ Methods for Studying the Brain

Treatment of Abnormal Psych ✖
- ✖ Treatment
- ✖ Evaluating Treatment and Prevention: What Works?
- ✖ Reducing Disorder by Confronting It: Psychotherapy
- ✖ Hypnosis
- ✖ Treatment (Psychotherapy, biomedical, social)
- ✖ Insight Therapy
- ✖ Reducing Disorder Biologically: Drug and Brain Therapy

List B

Intro
- Introduction
- Introducing Psychology
- Welcome to Psychology
- Attending

History
- Foundations: This is your Brain
- Foundations: Freud
- Freud and Fairy Tales
- History and Perspectives
- Foundations: Skinner
- The Origins of Personality

Language
- Conscious of the Past: Language
- Interacting With Others: Helping, Hurting, and Conforming
- Language: The development and use
- Language
- Language Development
- How Do We Communicate?: Language in the Brain, Mouth and the Hands

Mental Illness
- What Happens When Things Go Wrong: Mental Illness
- Mood Disorders: Emotions as Illness
- Anxiety and Dissociative Disorders: Fearing the World Around Us
- Stress
- Defining Mental Illness
- Causing Mental Illness
- Personality Disorders

Treating Mental illness
- Psychopatholgy

Emotions
- Emotion
- The Good Life: Happiness

Figure 1. The Crowdlines crowd synthesis interface, slightly modified from the HCLS interface in Experiment 1, and used by crowdworkers for generating the outlines for Experiment 2.

lists of 25 topics each, A (left side) and B (right side), by creating and naming groups of related topics. The worker can mouse over any topic to reveal more detail.

Each of the four interfaces represents a different balance of context and structure:

- *High context, low structure (HCLS)*: Workers see all topics in both lists, and can group or exclude topics in an arbitrary order. This interface most closely resembles the family of ontology alignment tools designed for database administrators (Falconer and Noy 2011).
- *High context, medium structure (HCMS)*: Workers see all topics in both lists. The interface directs them sequentially through each topic in List A, and the worker can choose which topics (if any) to group from List B.
- *Medium context, medium structure (MCMS)*: Workers see only one topic at a time in List A, and all topics in List B. When they finish grouping a List A topic, the interface directs them to the next one.
- *Low context, high structure (LCHS)*: Workers see only a pair of topics at a time, one from each list. They indicate whether the topics are related and, if so, choose a name for the group. The interface then directs the worker to the next topic pair. This condition most closely resembles pairwise crowd clustering (Gomes et al. 2011; Tamuz et al. 2011; Yi et al. 2012).

Other combinations of context and structure were considered, but ruled out during the design and early testing phases. These four represent the most promising options.

Research Questions

Using the interfaces described above, we propose the following research questions and hypotheses:

- **How do different combinations of context and structure affect *group quality* in crowdsourced merging tasks? *Group quality* refers to the quality of a group of topics formed by a crowd worker merging information from multiple sources.** We hypothesize that the medium context, medium structure (MCMS) interface will strike the right balance for crowd workers.
- **How do different combinations of context and structure affect *efficiency* in crowdsourced merging tasks?** We hypothesize that the high context, low structure (HCLS) interface will be fastest, since workers have the fewest constraints on their workflow. The low context, high structure (LCHS) interface will be slowest, as workers are required to consider 625 unique pairs.

Method

To provide content for the merges, we chose two "Introduction to Psychology" course syllabi, collected from public websites. Both syllabi, or *lists*, encompassed 25 class meetings, and each meeting covered one topic, which we refer to as *topic names*. *Topic details* were the descriptions provided by the instructor for each class, including supplementary resources such as links to slide decks or videos.

We recruited crowd workers from Amazon Mechanical Turk (MTurk), and paid 153 workers US$2 each to merge one list. This rate was based on the minimum wage in our location and 15-20 minute average task times observed in our pilots. In total, Experiment 1 cost $306 plus fees.

Each of the four interfaces was an experimental condition. When workers accepted our task, they were randomly assigned one of the four interfaces. We logged each assignment and whether it was completed, in order to analyze attrition rates for each condition. Consequently, when we closed the task a few days later, we had more trials for some conditions than others.

To measure efficiency, we collected the elapsed time for each merge, as well as the aforementioned attrition data.

To measure group quality, we chose to compare crowd merges against a baseline created by experts. We recruited a tenured psychology professor (Expert 1) and a postdoc with a psychology PhD (Expert 2) and asked them to perform the same merge task as crowd workers.

We computed similarity by calculating the f-score (harmonic mean of precision and recall) for each crowd merge, using one of the experts as the baseline. Since all interfaces required users to either group each List A topic or choose "No related topics," we can use List A topics as an anchor point to compare topic groups. For example, if a crowd worker formed a group with topics A1, B2, and B3, we can find the group containing the expert's A1 topic and see if he or she also included B2 and B3.

This measure of similarity to experts serves only as a proxy for group quality, and it's possible for crowd merges to offer a different, yet equally valuable, grouping of topics. We present the results for each expert baseline separately to acknowledge different notions of quality.

Results

HCLS interface produces highest quality merges

Table 1 displays the mean f-scores, compared against Experts 1 and 2, across the four conditions. There is a consistent pattern, where higher context and less structure lead to better f-scores across both experts. The best-scoring condition across both experts is High context, low structure (HCLS), with mean f-scores of 0.46 and 0.48 across Experts 1 and 2, respectively. Low context, high structure performed worst, scoring an f-score of 0.39 compared to either expert. One-way ANOVAs showed a significant effect of condition on f-score for Expert 1 (F(3,3821) = 9.85, p<0.05) and Expert 2 (F(3,3821) = 15.71, p<0.05). Post-hoc Tukey tests for Expert 1 showed that HCLS had significantly higher f-scores than any of the other conditions (p<0.05), and no other differences were significant. Tukey tests for Expert 2 showed that both high context conditions (HCLS and HCMS) scored significantly higher f-scores than the other conditions (p<0.05).

Condition	N	Completion	Time (min)	f (E1)	f (E2)
HCLS	45	**63.4%**	**19.8***	**0.46***	**0.48***
HCMS	39	50.0%	21.6	0.41	0.44*
MCMS	36	25.7%	21.8	0.38	0.39
LCHS	33	16.3%	26.6	0.39	0.39

Table 1. Experiment 1 results. Highest values in bold (p<0.05).*

HCLS interface yields fastest merges, lowest attrition

Table 1 also displays the mean time (minutes) for the four conditions. Conditions with higher context and lower structure have shorter task times. The fastest completion time is High context, low structure (HCLS) with 19.8 minutes on average; the lowest is Low context, high structure (LCHS) which took 26.6 minutes on average. A one-way ANOVA showed that condition had a significant effect on elapsed time (F(3,149)=2.71, p<0.05). Post-hoc Tukey tests showed that HCLS is significantly faster than LCHS, but none of the other differences are significant.

We also compared completion rates. As with task time, completion rates are higher for conditions with higher context and lower structure. HCLS had the highest completion rate (63.4%) while LCHS had the lowest (16.3%).

Discussion

Our evaluation of group quality showed that high context conditions yielded better merge quality. Specifically, HCLS provided significantly higher similarity to both expert gold standards, and HCMS also provided significantly higher similarity to one of the experts. These findings disconfirm our hypothesis that MCMS's middle balance of context and structure would yield the best group quality.

We also hypothesized that the HCLS interface would provide the greatest efficiency, while the LCHS interface would be slowest. Our results support this hypothesis, as mean completion time for HCLS was significantly faster (mean=19.8 min/merge) than LCHS (26.6 min/merge). Additionally, mean completion rates for HCLS were more than 3x higher (mean=63.4%) compared to LCHS (16.3%). This latter attrition rate for the LCHS condition was worryingly high, yet this interface is essentially identical to state-of-the-art of crowd applications in databases research (e.g. Sarasua, Simperl, and Noy 2012). Our data indicates that this approach is unacceptably tedious. The HCLS approach leads to both better (faster, higher quality) results and much lower attrition.

Experiment 2: Evaluating Crowdlines Workflows and Search

Experiment 1 established the HCLS interface as the best option for a single worker. However, our vision for Crowdlines involves using many crowdworkers to integrate many information sources, and producing an outline that searchers can use to learn about an unfamiliar domain. For the second part of our study, Experiment 2, we developed two promising crowd workflows for merging multiple lists across multiple workers, and built and evaluated a search-oriented interface for users to explore and learn from the crowd's output.

Designing the Crowdsourced Workflows

The results of Experiment 1 led us to move forward with the HCLS interface. We made several improvements based on feedback from crowd workers, such as letting users edit and delete topic groups, and drag and drop topics into groups using a direct manipulation interface (Figure 1).

We also made several design changes in anticipation of supporting multiple merges. Most of these changes centered around the need to reconsider the constraints placed on workers, as multiple merges exposed them to lists of varying sizes. Without constraints, "lazy turkers" might group only a few topics, while "eager beavers" might produce dozens of questionable groups. After some experimentation, we settled on the following global constraints:

- *Workers must create at least 15 groups*. When coupled with the topic constraint below, this allows for a resulting merged list of between 30 and 50 topics. This range fit our intuition for how many topics could reasonably be passed on and re-merged.
- *Each group must have at least two topics, one from each list*. Two topics is the minimum needed to form a group, while the requirement to include topics from both lists ensures that workers are actually merging.
- *It's not necessary to group every topic*. This was added when we discovered during pilots that workers tried to group every topic, even when it was a poor fit.

We wanted to identify an effective way to merge a potentially unlimited number of sources. Previous research on crowdsourcing workflows suggested two promising candidates: linear (serial/iterative) and tournament (parallel). In the *linear* workflow (Figure 2, right), merges occur serially. In each round, a new list is merged into the result of all previous merges. For example, in Round 1, a worker merges lists A and B. In Round 2, another worker merges list C into the result of the previous round (A+B). This process repeats until all lists have been merged.

In the *tournament* workflow (Figure 2, left), some merges occur in parallel, and those results are fed into a new round of merges, similar to sports teams competing in a tournament. For example, in Round 1, one worker merges lists A and B, and another worker merges lists C and D. In Round 2, a third worker merges list A+B with list C+D. Again, this process repeats until all lists are merged.

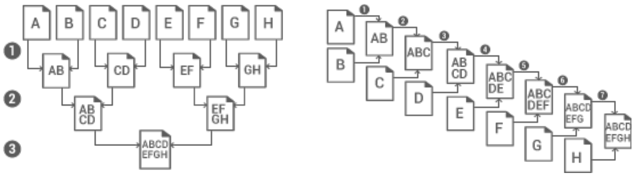


Figure 2. Tournament (left) and linear (right) crowd workflows investigated in Experiment 2.

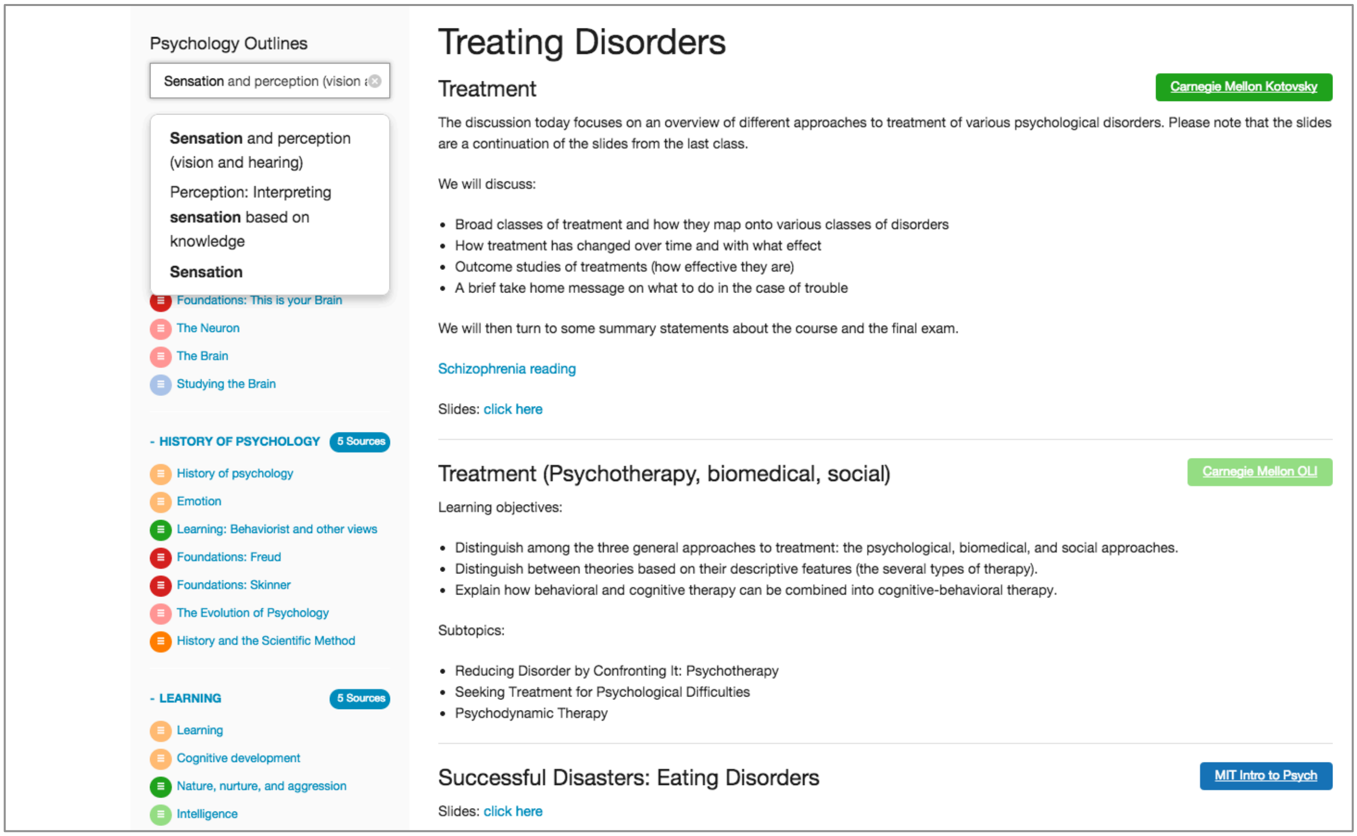
These workflows are well suited to comparison. One reason, mentioned above, is that the literature on iterative vs. parallelized workflows is still inconclusive. Another reason is that these workflows share some key features, while also differing in important ways. Both assume one worker merging two lists as the atomic unit of work, and both require *n*-1 workers to merge *n* lists. Yet, the tournament workflow is parallelized, while the linear workflow is serialized. This suggests that the tournament workflow may be more efficient. However, while the linear workflow merges only one new list at a time (25 topics), the tournament workflow merges multiple lists after the first round (50+ groups and topics). This suggests that the linear workflow may be easier for crowd workers because they have fewer topics to integrate. Both workflows have apparent strengths and weaknesses, but neither is obviously superior for our purposes.

Generating the Crowdsourced Outlines

To prepare for Experiment 2, we wanted to compare a multiple merge involving eight unique lists. This required us to collect six additional Intro Psych syllabi from other universities to supplement the two from Experiment 1. Each list had exactly 25 topics with corresponding details.

We again recruited crowd workers (N=140) from MTurk, paying them $2 per merge. To control for worker pool variation and time of day, we launched tasks for Rounds 1–3 for both workflows at the same time. A unique worker performed each merge.

Workers were randomly assigned to either the linear or tournament version of the synthesis interface. Merging eight lists using the linear workflow requires seven workers over seven rounds (one per round). The tournament workflow also requires seven workers, but over three rounds (four in Round 1, two in Round 2, one in Round 3). Merging all eight lists (200 topics) using either workflow cost $14 plus fees. Because the selection of lists and workers for the initial rounds impacts subsequent rounds, we ran 10 trials for each workflow, randomizing which lists would be merged in each round.


Psychology Outlines
Sensation and perception (vision a
Sensation and perception (vision and hearing)
Perception: Interpreting sensation based on knowledge
Sensation
Foundations: This is your Brain
The Neuron
The Brain
Studying the Brain
- HISTORY OF PSYCHOLOGY
5 Sources
History of psychology
Emotion
Learning: Behaviorist and other views
Foundations: Freud
Foundations: Skinner
The Evolution of Psychology
History and the Scientific Method
- LEARNING
5 Sources
Learning
Cognitive development
Nature, nurture, and aggression
Intelligence
Carnegie Mellon Kotovsky
Carnegie Mellon OLI
MIT Intro to Psych

Treating Disorders

Treatment

The discussion today focuses on an overview of different approaches to treatment of various psychological disorders. Please note that the slides are a continuation of the slides from the last class.

We will discuss:

- Broad classes of treatment and how they map onto various classes of disorders
- How treatment has changed over time and with what effect
- Outcome studies of treatments (how effective they are)
- A brief take home message on what to do in the case of trouble

We will then turn to some summary statements about the course and the final exam.

Schizophrenia reading

Slides: click here

Treatment (Psychotherapy, biomedical, social)

Learning objectives:

- Distinguish among the three general approaches to treatment: the psychological, biomedical, and social approaches.
- Distinguish between theories based on their descriptive features (the several types of therapy).
- Explain how behavioral and cognitive therapy can be combined into cognitive-behavioral therapy.

Subtopics:

- Reducing Disorder by Confronting It: Psychotherapy
- Seeking Treatment for Psychological Difficulties
- Psychodynamic Therapy

Successful Disasters: Eating Disorders

Slides: click here

Figure 3. The Crowdlines search interface evaluated in Experiment 2.

Tournament workflow is 2–3 times faster than linear

Since both workflows use equal numbers of crowd workers, we wanted to establish which workflow was faster. To calculate average speed for the linear workflow, we summed the elapsed time for the 7 rounds of each trial, and averaged the sums. Since multiple merges occur in parallel in the tournament condition, we considered two measures: the mean of (1) the slowest merges and (2) the fastest merges at each round per trial. (We waited on the slowest merge to complete before advancing rounds, but an optimized workflow could reduce this waiting time.)

We found that tournament is much faster than linear, even without optimization. The mean elapsed time per trial (all 7 rounds) for the linear workflow was 182.0 minutes, compared to 88.0 minutes for tournament trials (3 parallel rounds) using the slowest merges, and 60.0 minutes for tournament trials using the fastest merges. A one-way ANOVA found a significant effect of workflow on elapsed time per trial ($F(2,27)=117.2$, $p<0.05$). Post-hoc Tukey tests showed that both the slowest-merge and fastest-merge measures of tournament are significantly faster than linear ($p<0.05$). We further found that elapsed time *per round* increases gradually, from approximately 20 minutes in the first round to 30 minutes in the last round, and an independent samples t-test comparing final round times showed no significant difference between workflows. This gradual increase suggests that either approach can reasonably scale to large numbers of lists.

Designing the Search Interface

We built a search interface for aggregating, displaying and exploring the results of crowd-powered merges (Figure 3). Crowd-generated topic groups are presented in outline format in the left column. The groups are ordered by source diversity, i.e., the number of unique psychology syllabi included in each group, to prioritize the most comprehensive groups. Color codes indicate the different sources for the topics. A search box at the top allows the user to quickly find topics of interest, and an auto-complete feature suggests potentially relevant group and topic names.

Clicking a group or topic on the left column shows a detail view of that information in the right column. The group name appears at the top, followed by each topic, the source syllabus (again color-coded), and finally relevant details and links extracted from the source.

For Experiment 2, we wanted to understand (generally) how crowdsourced merges affect the behavior and subjective experience of people seeking to synthesize diverse information, and (in particular) how the Crowdlines search interface might support their efforts.

Research Question

We propose the following research question and hypothesis: **How does Crowdlines affect the task performance of people synthesizing diverse information?** We hypothesize that Crowdlines will help people perform better than those using only web search or only a list of syllabi.

Method

We recruited 115 participants from MTurk, who were each compensated $4 for their time. We asked them to imagine they were teaching a short Intro Psych course, and their task was to create a syllabus with six class meetings, with topics, subtopics, and links to relevant online sources for each meeting. Participants had 20 minutes to complete the task and could not end the task early.

The experiment had four conditions, assigned randomly. Participants in the first, control condition (N=24) used their preferred search engine to complete the task. Participants in the other three, experimental conditions were likewise encouraged to use a search engine, but were also given one additional resource. Participants in the lists condition (N=36) were given a PDF with all eight syllabi. Participants in the Crowdlines linear condition (N=28) were given the search interface loaded with one of the 10 linear crowd merges. Finally, participants in the Crowdlines tournament condition (N=27) were given the search interface with one of the 10 tournament merges. All 20 possible crowd merges were seen by at least two participants.

To evaluate task performance, we measured the number of class topics (maximum of 6), subtopics, and sources submitted by each participant. We also sought a measure of topic quality by comparing participant class topics to a gold standard created by experienced psychologists. To generate this gold standard, two authors of this paper performed a bottom-up analysis of the eight Intro Psych syllabi described earlier, grouping topics based on relatedness. The resulting gold standard list contained 18 topics comprised of 173 subtopics across the eight syllabi. To determine the most important topics from this list of 18, we chose the eight topics most broadly represented across the eight syllabi. Memory, Disorders, Emotion, and Brain were included in all eight syllabi, while Development, Learning, Sensation and Perception, and Sleep and Dreams were each represented in seven syllabi.

To compare participant performance to the experts, we had to map participants' class topics onto this gold standard. First, two authors of this paper independently mapped 68 topics from pilot data to the 18 topics on the gold standard list. They achieved good initial agreement and discussed points of disagreement. Next, they independently mapped all 653 topics generated from the 115 participants in Experiment 2, blind to condition. The Cohen's kappa for this mapping was 0.82, indicating excellent agreement. If both authors agreed on a mapping for a topic, it was considered a valid topic; if neither author could map the topic to the gold standard or they disagreed on the mapping, it was considered invalid.

Results

Crowdlines outperforms web search for all measures

Task performance results are shown in Table 2. For the quality analysis, we calculated the f-score for all participants, comparing their class topics to those from the gold standard list. The Crowdlines tournament condition was the clear winner, yielding an average f-score 37% to 50% higher than all other conditions. A one-way ANOVA showed that condition had a significant effect on f-score (F(3,111)=3.75, p<0.05). Post-hoc Tukey tests showed that Crowdlines tournament's f-score was significantly higher than any of the other conditions (p<0.05), and no other differences were significant.

Measure	**Web**	**Lists**	**CL Linear**	**CL Tourn**
Total participants	24	36	28	27
Mean f-score	0.32	0.35	0.28	**0.41***
Mean topics	5.5	5.7	5.6	**5.8**
Mean subtopics	14.9	16.3	16.4	**17.0**
Mean sources	10.2	10.2	**15.2***	14.5*
Mean unique sources	8.6	8.6	**10.4**	8.9
Mean syllabi sources	0.71	0.81	6.9*	**9.4***

Table 2. Experiment 2 results. Highest values in bold (p<0.05).*

Participants in the Crowdlines tournament condition generated the most class topics (mean=5.8) and subtopics (mean=17), but the differences across conditions weren't significant. Crowdlines linear participants provided the most sources (mean=15.2), compared to 14.5 for Crowdlines tournament and just 10.2 for either web or lists. A one-way ANOVA showed condition had a significant effect on number of sources (F(3,11)=8.06, p<0.05). Post-hoc Tukey tests showed both Crowdlines conditions produced significantly more sources than web or lists (p<0.05). Crowdlines participants used more sources *from the eight syllabi* (mean=9.4 for linear and 6.9 for tournament) compared to web or lists (both <1). A one-way ANOVA with post-hoc Tukey tests showed condition as a significant effect on syllabi sources (F(3,111)=17.34, p<0.05), and both Crowdlines conditions yielding significantly more syllabi sources than non-Crowdlines conditions (p<0.05). Finally, Crowdlines linear participants produced the most *unique* sources (mean=10.4), while the other three conditions hovered around means of 8.6–8.9, but the differences were not statistically significant. In summary, participants in the Crowdlines conditions outperformed the non-Crowdlines conditions across all measures, with the average number of total sources and syllabi sources being significantly higher for Crowdlines.

Tournament workflow leads to greater source diversity, topic survival

To better understand why Crowdlines tournament outperformed Crowdlines linear, we conducted a follow-up analysis looking at source diversity and topic survival. Source diversity refers to how well represented topics from all eight syllabi are in a final merged list. Using Simpson's (1949) diversity index (0=no diversity, 1=infinite diversity), we found that tournament lists have greater diversity (mean=0.87) than linear lists (mean=0.79), and an independent samples t-test showed the difference is significant (t(18)=-5.09, p<0.05). We also examined topic survival, i.e. what percentage of topics were merged vs. omitted at each round, finding greater topic survival per round for tournament (mean=70.0%) compared to linear (mean=54.8%). An independent samples t-test showed this difference to be significant (t(88)=-3.22, p<0.05).

Discussion

We hypothesized that participants who used Crowdlines would perform better than those who didn't. Our results show that Crowdlines led participants to gather significantly more sources in general, and sources from syllabi in particular. We also found that participants in the Crowdlines tournament condition produced significantly higher quality topics than any of the other conditions. Further, Crowdlines conditions scored higher across all other areas that were measured, including number of topics, subtopics, and unique sources, though these differences weren't statistically significant. Taken together, the evidence suggests that Crowdlines caused participants to produce synthesized information that is more similar to experts and includes more sources, partially supporting our hypothesis.

Implications and Conclusions

Crowd Synthesis Interfaces: Context vs. Structure

Balancing context and structure is a fundamental challenge for designing crowdsourcing interfaces and workflows. Experiment 1 found that, for crowdsourced merges of two schemas, the interface that provided High context, low structure (HCLS) yielded higher quality, faster completion time, and higher completion rates than other combinations of context and structure. The high context provided by this interface may have given crowdworkers a more complete sense of the breadth and depth of the information in each interface, making it easier to make comparisons and connections. The minimal structure of this interface's workflow may have contributed to the faster completion times and higher completion rates, giving workers more freedom to complete the task using a process that fit their particular working style.

While high context and low structure produced the best results for this experiment, notions of "high" and low" are relative. "High" context here meant showing all topics for both schemas being merged, numbering between 25 and 100 topics per schema in most cases, and hiding topic details behind a hover interaction. Our experience has been that most documents on the web have fewer than 25 major topics, but very long or dense material may need to be divided up. "Low" structure here meant allowing workers to merge topics in any sequence, but the synthesis interface still enforced several constraints, including number of groups and topics per group. A completely unstructured interface is feasible, but would need to allow for emergent social norms (if not technical constraints) to encourage high quality contributions, as with Wikipedia (Butler, Joyce, and Pike 2008). Recent work exploring peer assessment in crowdsourcing may offer one promising direction (Zhu et al. 2014).

Crowd Synthesis Workflows: Linear vs. Tournament

Having established an effective synthesis interface for a single-worker merge task in Experiment 1, we next compared two potential workflows for merges across multiple workers—linear and tournament—inspired by prior research in databases and crowdsourcing. Both workflows use the same number of workers, but tournament is 2–3 times faster due to parallelization. Experiment 2 compared performance in an information synthesis (syllabi creation) task using the Crowdlines search interface (with either linear or tournament crowd merges), web search, or the raw information sources. Once again, tournament emerged as the winner, with participants creating significantly more expert-like, diverse, integrated syllabi with more sources. Thus, tournament produced better results using the same number of workers in half the time. These results suggest that a parallelized workflow for crowd synthesis, where material is merged hierarchically, will be most effective.

The only dimension across any experiment where linear proved significantly better than other conditions was total number of sources per participant-generated syllabus. In some areas, like syllabus quality, linear was actually worse than the control condition. Why was tournament so much more effective? One possibility is that tournament's parallelized model offers a more useful "big picture" view of the material to be synthesized. In the first round, each worker merges two entirely different sources. In subsequent rounds, those results, which may offer very different schemas, are merged by a new set of workers. Every tournament worker is exposed to strong contrasts in schemas, topics, and sources at every round, which may suggest broader and deeper kinds of connections. It may also contribute to a more motivating and interesting task. Linear is a more gradual merging workflow where workers see only half as much new material as tournament. The similarity between old and new groups may feel tedious or redundant, causing workers lose focus.

Limitations and Future Work

We conducted our experiments in the domain of psychology and specifically introductory course syllabi. Like many online sensemaking tasks, psychology combines common knowledge with technical information, and generating syllabi involves behaviors like summarizing, prioritizing, and connecting information. We believe our results will generalize well to other domains, but follow-up studies are needed. Additionally, Experiment 2 evaluated the usefulness of Crowdlines for generating short Intro Psych course syllabi; while this is a realistic online sensemaking task for psychology instructors, our participants were non-experts whose motivations, support needs, and practices likely differ from experts. Despite this, we found that Crowdlines helped even these novices create syllabi significantly more similar to experts than other tools.

This paper focuses on crowdsourcing the synthesis of diverse information that has already been collected from the web. Previous work has demonstrated that crowds can be highly effective at gathering online sources for sensemaking tasks (Kittur et al. 2014), so we feel confident that Crowdlines can build on these earlier advances.

Finally, crowd workflow design impacts many types of sensemaking tasks. This paper focused on synthesis of diverse sources, but the tradeoffs of linear vs. tournament workflows may apply more generally, e.g. to voting tasks.

Acknowledgments

We thank Bhawna Agarwal for assistance with the design of experiments and interfaces. This work was supported by NSF awards 1208382, 1122206, 1149797, 1217559, 0968484 and a Google Faculty Research Award.

References

Abbott, R.D. 1999. *The World as Information: Overload and Personal Design*. Intellect Books.

André, P.; Kittur, A.; and Dow, S.P. 2014. Crowd Synthesis: Extracting Categories and Clusters from Complex Data. In *Proc. CSCW 2014*.

André, P.; Kraut, R.E.; and Kittur, A. 2014. Effects of Simultaneous and Sequential Work Structures on Distributed Collaborative Interdependent Tasks. In *Proc. CHI 2014*.

André, P.; Zhang, H.; Kim, J.; Chilton, L.B.; Dow, S.P.; and Miller, R.C. 2013. Community Clustering: Leveraging an Academic Crowd to Form Coherent Conference Sessions. In *Proc. HCOMP 2013*.

Butler, B.; Joyce, E.; and Pike, J. 2008. Don't Look Now, but We've Created a Bureaucracy: The Nature and Roles of Policies and Rules in Wikipedia. In *Proc. CHI 2008*.

Chilton, L.B.; Kim, J.; André, P.; Cordeiro, F.; Landay, J.A.; Weld, D.S.; Dow, S.P.; Miller, R.C.; and Zhang, H. 2014. Frenzy: Collaborative Data Organization for Creating Conference Sessions. In *Proc. CHI 2014*.

Chilton, L.B.; Little, G.; Edge, D.; Weld, D.S.; and Landay, J.A. 2013. Cascade: Crowdsourcing Taxonomy Creation. In *Proc. CHI 2013*.

Choi, N.; Song, I.; and Han, H. 2006. A Survey on Ontology Mapping. *SIGMOD Rec.* 35 (3): 34–41.

Falconer, S. M.; and Noy, N.F. 2011. Interactive Techniques to Support Ontology Matching. In *Schema Matching and Mapping*, 29–51. Springer Berlin Heidelberg.

Fisher, K.; Counts, S.; and Kittur, A. 2012. Distributed Sensemaking: Improving Sensemaking by Leveraging the Efforts of Previous Users. In *Proc. CHI 2012*.

Gomes, R.; Welinder, P.; Krause, A.; and Perona, P. 2011. Crowdclustering. In *Proc. NIPS 2011*.

Kittur, A.; Peters, A.M.; Diriye, A.; and Bove, M. 2014. Standing on the Schemas of Giants: Socially Augmented Information Foraging. In *Proc. CSCW 2014*.

Little, G.; Chilton, L.B.; Goldman, M.; and Miller, R.C. 2010. Exploring Iterative and Parallel Human Computation Processes. In *Proc. HCOMP 2010*.

Rathod, N.; and Cassel, L. 2013. Building a Search Engine for Computer Science Course Syllabi. In *Proc. JCDL 2013*.

Salton, G.; and McGill, M.J. 1983. *Introduction to Modern Information Retrieval*. Mcgraw-Hill College.

Sarasua, C.; Simperl, E.; and Noy, N.F. 2012. CrowdMap: Crowdsourcing Ontology Alignment with Microtasks. In *Proc. ISWC 2012*.

Shahaf, D.; Guestrin, C.; and Horvitz, E. 2012. Metro Maps of Science. In *Proc. KDD 2012*.

Simpson, E.H. 1949. Measurement of Diversity. *Nature* 163: 688.

Tamuz, O.; Liu, C.; Belongie, S.; Shamir, O.; and Kalai, A.T. 2011. Adaptively Learning the Crowd Kernel. In *Proc. ICML 2011*.

Willett, W.; Ginosar, S.; Steinitz, A.; Hartmann, B.; and Agrawala, M. 2013. Identifying Redundancy and Exposing Provenance in Crowdsourced Data Analysis. In *Proc. IEEE VIS 2013*.

Yi, J.; Jin, R.; Jain, A.; and Jain, S. 2012. Crowdclustering with Sparse Pairwise Labels: A Matrix Completion Approach. In *Proc. HCOMP 2012*.

Zhang, H.; Law, E.; Miller, R.; Gajos, K.; Parkes, D.; and Horvitz, E. 2012. Human Computation Tasks with Global Constraints. In *Proc. CHI 2012*.

Zhu, H.; Dow, S.P.; Kraut, R.E.; and Kittur, A. 2014. Reviewing Versus Doing: Learning and Performance in Crowd Assessment. In *Proc. CSCW 2014*.