Efficient Crowd Exploration of Large Networks: The Case of Causal Attribution

DANIEL BERENBERG, University of Vermont, USA
JAMES P. BAGROW, University of Vermont, USA

Accurately and efficiently crowdsourcing complex, open-ended tasks can be difficult, as crowd participants tend to favor short, repetitive "microtasks". We study the crowdsourcing of large networks where the crowd provides the network topology via microtasks. Crowds can explore many types of social and information networks, but we focus on the network of causal attributions, an important network that signifies cause-and-effect relationships. We conduct experiments on Amazon Mechanical Turk (AMT) testing how workers propose and validate individual causal relationships and introduce a method for independent crowd workers to explore large networks. The core of the method, Iterative Pathway Refinement, is a theoretically-principled mechanism for efficient exploration via microtasks. We evaluate the method using synthetic networks and apply it on AMT to extract a large-scale causal attribution network, then investigate the structure of this network as well as the activity patterns and efficiency of the workers who constructed this network. Worker interactions reveal important characteristics of causal perception and the network data they generate can improve our understanding of causality and causal inference.

CCS Concepts: • **Information systems → Crowdsourcing**; *Answer ranking*; • **Human-centered computing → Computer supported cooperative work**; *Collaborative content creation*; • **Computing methodologies →** *Network science*;

Additional Key Words and Phrases: Crowdsourcing; crowdwork; causal attribution; causality; networks; network motifs; Amazon Mechanical Turk; microtasks; self-avoiding walks

ACM Reference Format:
Daniel Berenberg and James P. Bagrow. 2018. Efficient Crowd Exploration of Large Networks: The Case of Causal Attribution. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 24 (November 2018), 25 pages. https://doi.org/10.1145/3274293

1 INTRODUCTION

Crowdsourcing has emerged as a powerful technique for gathering data that are otherwise inaccessible, either computationally or logistically [9, 27]. These data may be training data for machine learning algorithms, survey data, or the results of behavioral experiments [37, 59]. While data gathering is a key use of crowdsourcing, crowd participants are also uniquely capable of ingenuity and creativity, and the most powerful applications of crowdsourcing exploit this to provide crowdsourcers with novel ideas and out-of-the-box thinking [2, 36, 58].

When applying crowdsourcing, there is often an antagonism between the complexity of the task the crowdsourcer wishes to accomplish and the preference of crowd workers in favor of short *microtasks* [38]. This has led to considerable research on decomposing various large-scale tasks

Authors' addresses: Daniel Berenberg, University of Vermont, Computer Science, Burlington, VT, 05405, USA, daniel.berenberg@uvm.edu; James P. Bagrow, University of Vermont, Mathematics & Statistics and Vermont Complex Systems Center, Burlington, VT, 05405, USA, james.bagrow@uvm.edu.

2573-0142/2018/11-ART24 $15.00
https://doi.org/10.1145/3274293

into microtasks more suitable for the crowd [13]. In our case, we are interested in enabling the crowd to efficiently explore large, unknown networks by asking workers to propose some of the nodes and links within a particular network of interest.

The objective of this work is to study how independent crowd workers can efficiently explore a large causal attribution network. We contribute a new algorithm for crowdsourcing large networks, and perform simulations, conduct experiments and perform surveys to assess the activity patterns, efficiency, and efficacy of workers using this algorithm. We focus on causal attribution networks, where workers are asked to provide directed links between causes and effects, and so we also conduct experiments on Amazon Mechanical Turk to better understand how workers attribute causes and effects. Causal reasoning, while affected by cognitive biases, remains one of the biggest differentiators between human intelligence and machine learning methods, making this problem domain an ideal venue for crowdsourcing. Other types of networks, such as knowledge graphs, concept maps, or social networks, may also be explored with crowds.

A simple way to decompose the larger network exploration task into microtasks is to ask workers to validate a single link for each task, but this provides minimal information per task and may not be efficient. Motivated by theoretical studies of network search strategies, we propose and evaluate a network exploration microtask—**Iterative Pathway Refinement**—where workers create and modify *pathways*, short linear paths of nodes, within the larger network. These pathways are easy for workers to modify, provide more information than can be gathered from a single link microtask, and the union of these pathways provides inference of the larger network being explored.

The rest of this work is organized as follows. Section 2 surveys previous research on several aspects of the problem we study, including crowdsourcing, causal attribution, and network search. Section 3 describes our first experiment conducted on Amazon Mechanical Turk (AMT). The goal of this experiment is to estimate the efficacy of crowd workers tasked with proposing and validating individual cause-and-effect relationships. Motivated by this experiment and with the goal of maximizing crowd efficiency, in Sec. 4 we introduce a set of algorithms enabling a crowd of workers to collectively explore a large network while individually participating only in microtasks. Then, Sec. 5 presents our second experiment implementing these algorithms on AMT to derive a large-scale causal attribution network, while Sec. 6 performs a followup analysis on the quality of responses generated with our new algorithms. Section 7 proposes a mathematical model for the exploration algorithm and analyzes it on known test networks, to better understand how a sampled network derived via our exploration algorithm differs from a true, underlying network. We conclude with a discussion in Sec. 8, including how our work here can be generalized and can inform the crowd exploration of other types of networks.

2 BACKGROUND

The focus of this work can best be understood in the context of three research areas: crowdsourcing, causal attribution theory, and search in networks.

2.1 Crowdsourcing

Crowdsourcing is the recruitment and application of large groups of individuals towards the generation of work [9, 20, 27]. The crowd may be voluntary participants or paid workers, and they may or may not need to be subject-matter experts in areas relevant to the particular crowdsourced tasks. Common crowdsourcing tasks include labeling images, disambiguating written records, and participating in surveys and behavioral experiments. Yet, the experience and creativity of crowd workers is one of the most unique aspects of crowdsourcing, and leveraging these assets can achieve results far beyond the confines of "artificial artificial intelligence" [36].

The most popular online platform for recruiting paid workers is Amazon Mechanical Turk (AMT). Large crowdsourcing tasks are generally broken down into "microtasks" referred to on AMT as Human Intelligence Tasks (HITs). Dividing a large task or set of tasks into many small microtasks, either manually or algorithmically, is one of the most effective ways to distribute complex work over a crowd [38, 43]. Many methods use a propose-and-vote/fix-verify/select-validate mechanism for decomposing complex tasks without harming reliability or quality [3, 5, 14, 56]. Monitoring and improving upon the completion times of microtasks and batches of microtasks is an important aspect of crowdsourcing, as it can improve overall efficiency and quality [17, 29, 37]. Efficiency and quality are also enhanced by using statistical aggregation strategies that combine multiple worker responses to microtasks [15, 32, 40].

Many crowdsourcing projects study the problem of assigning workers to a fixed set of predetermined tasks [42], for example annotating a collection of images, but a growing body of work is considering areas where the crowd contributes new tasks to the crowdsourcer [6, 8, 56, 58, 61, 63].

Crowdsourcing the collection of network data is an under-explored area, particularly in the context of non-expert participants. One study considers crowdsourcing a network of synonymous terms—essentially a thesaurus—as a testbed for algorithms to efficiently distribute workers across a growing a set of tasks [45]. Another interesting study is the DREAM predictive signaling network challenge, where research teams were challenged to determine protein signaling networks from experimental data [25, 44, 51]. This challenge is specific to a single research area, and only experts in that area can reasonably contribute work. In general, it remains an open question how best to use crowdsourcing to explore large networks.

2.2 Causal attribution

Identifying and understanding causal relationships is a crucial way humans comprehend the world around them. Causal inference has been a major focus of philosophy, psychology, mathematics, and statistics for centuries [21, 22, 28, 31, 35, 50, 55]. While much progress has been made developing statistical tools, establishing causal relationships remains an outstanding scientific challenge.

Human understanding and perception of cause and effect relationships is complicated and influenced by language structure [11, 26, 34, 60] and sentiment [7]. The famous perception experiments of Michotte *et al.* have led to a long thread of experiments exploring how and why cognitive biases affect causal attribution [30, 54, 57]. Accounting for such biases is crucial to better understand causal attribution at scale.

2.3 Network exploration and search

A key question within Network Science is the problem of network exploration: how can an agent with only local or partial information understand the global structure of a large network? Likewise, how can an agent moving within a network efficiently find a predetermined search target? Search strategies are useful both for finding a given target and for efficiently mapping out the underlying structure of the network topology.

The ability to efficiently identify a target node in a network is a problem that has been studied since the seminal "small-world" work of Travers and Milgram [62]. Many successful strategies exploiting only local information are known for spatial networks and power-law networks [1, 39]. One such strategy to explore a network is to preferentially seek out the highest-degree nodes or hubs, those nodes with the most connections [1]. The more quickly a searcher arrives at a hub, the more avenues it has to explore the rest of the network, although this may lead to a biased view of the hubs if the network is not fully explored. Another well-supported local strategy is to perform a *self-avoiding walk* (SAW), moving randomly over the nodes of the network without returning to any previously visited nodes [1, 65]. The established success of SAW search provides a motivation

and theoretical underpinning for the key phase of the crowdsourcing algorithm we introduce in this work.

3 EXPERIMENT 1 — SINGLE LINK LEARNING

This experiment tests the ability of crowd workers to provide causal attribution information. Each crowdsourcing task focuses on asking workers to validate a single cause-effect pair by asking, for example, "*Does 'intelligence' cause 'foresight'?*". These candidate cause-effect term pairs (A, B) ('intelligent', and 'foresight' in this case) were tested by workers, who could respond with one of several multiple choice answers: (i) "A causes B", (ii) "B causes A", (iii) "A and B are unrelated", (iv) "something else causes both".

Aggregating multiple choice responses from multiple workers as they examine different term pairs allows this approach to study a larger network of causes and effects. However, the focus of Experiment 1 is estimating the efficacy of worker's causal attribution using various benchmark datasets. In Experiment 2 (Sec. 5) we return to the problem of network exploration.

3.1 Materials and methods

To study the efficacy of worker attributions of causal relationships, we extracted candidate cause-effect term pairs from two datasets:

Ground Truth data A benchmark database of 85 established cause-effect pairs [47]. These data are intended for validating causal detection algorithms. Many pairs, though not all, tend to cover scientific domains. Updated versions of this dataset currently provide 108 total term pairs.

Word Association data A set of 90 term pairs sampled randomly from the University of South Florida Free Association Norms dataset [49]. These data were collected over several decades from more than 6,000 participants. Each participant was shown a stimulus word and asked to produce the first word that comes to mind that was meaningfully related or associated with the stimulus. These associations provide a baseline for us to test; the terms in a pair may be causally related or the association may be due to other factors.

Further, we introduced **randomized versions** of both datasets by shuffling the terms between pairs. These randomized term pairs provide a base set of approximate known negatives to see what rate workers may attribute causal connections when a relationship is unlikely. Randomizing the same terms controls for the overall domains and contexts of the data while breaking any preexisting connections, either causal or associational, between terms in a pair. It is possible, of course, that a strong relationship between two random terms may exist, but this is less likely for the randomized data than for the original terms. A total of 350 distinct term-pairs were developed across the four datasets.

Data were collected on Amazon Mechanical Turk. Workers were shown a multiple choice web form for each HIT, asking them to select one of four options relating cause-effect pair (A, B): A causes B, B causes A, A and B are unrelated, something else causes both. An example of this form is shown in Fig. 1. As instructions, workers were shown the HIT web form populated with an example cause-effect term pair ("Population growth", "Food consumption growth") before their first HIT. Workers were rewarded \$0.04 per response. We aimed to collect $n = 50$ responses for each of the 350 distinct term-pairs.

This research procedure has been approved by our IRB (determination number CHRBSS: 15-039).

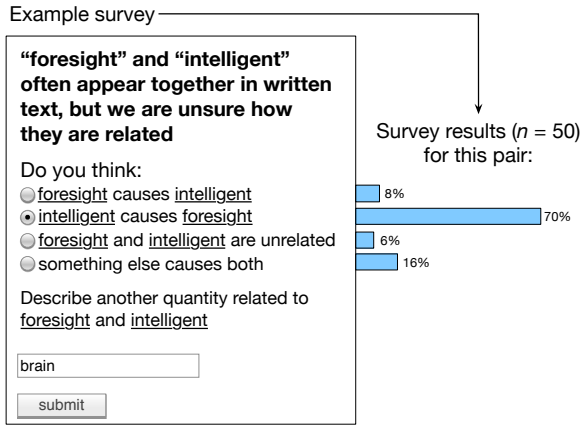


Fig. 1. Example survey conducted as part of Experiment 1, in this case for the terms "foresight" and "intelligent". The crowd reached strong consensus that intelligence causes foresight.

3.2 Results

We gathered 17,556 responses from 726 Mechanical Turk workers. Workers could complete as many HITs as available, but could not respond to the same term pair more than once. The mean and median numbers of responses per worker was 24.18 and 3, respectively.

We aggregated the total crowd responses to each term-pair and classified the link between the terms as either 'causal', 'confounded', or 'unrelated' based on the majority response from workers shown that pair. For example, the pair ("intelligent", "foresight") shown in Fig. 1 had 78% causal responses (attributing causality in either direction) so we classify it as a causal link. In this particular term pair, the direction of causality is clearly supported by the crowd, but other pairs with a majority 'causal' classification may be mixed, with workers split on whether A causes B or B causes A. Such a split is a useful signal that more information may be needed to better understand that particular term pair and perhaps other causes and effects related to those terms.

Figure 2 compares the majority crowd classification for terms across the test datasets. In both non-randomized datasets, the most common classification is 'causal', whereas in the two randomized variants, the most common classification is 'unrelated'. Comparing the original and randomized ground truth dataset (Fig. 2A), there is a clear difference in the proportion of cause-effect pairs labeled as 'causal' versus 'unrelated': 80% of the ground truth pairs have a crowd-majority label of 'causal' compared with 40% for the randomized ground truth data. At the same time, approximately 10% of ground truth pairs are labeled as 'unrelated' by the crowd majority, compared with 60% for the randomized ground truth. A similar classification difference holds between the word association and randomized word association data (also Fig. 2A).

However, workers appear to be over-reporting the incidence of causal relationships: 40% causal relationships for the randomized ground truth data and 80% causal relationships for the word association data both seem quite high. Workers may be biased in favor of explaining causal relationships where none exist [7, 60]. At the same time, however, other factors may be at play: (i) the domain of terms within the ground truth dataset is relatively narrow, making it more likely for random pairs to be related; (ii) the word association terms are generated by individuals who could very well be using cause-effect reasoning when they ideate an associated term. Likely both biased over-reporting and true causal relationships are leading to the causal attribution rates for the shuffled data.

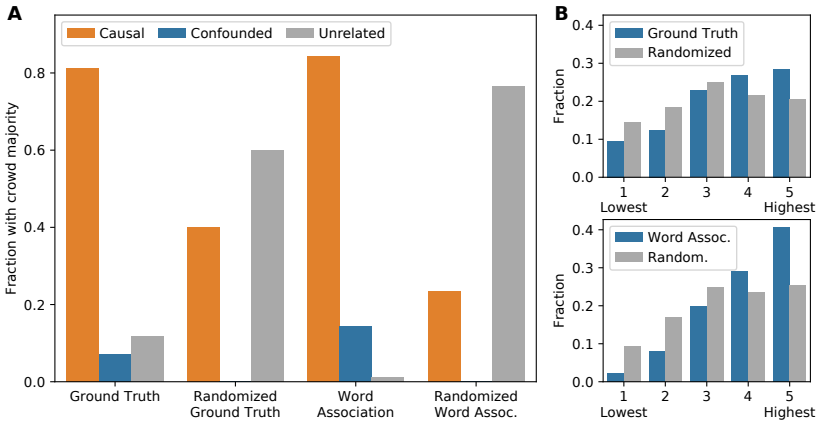


Fig. 2. Results of Experiment 1. (**A**) The proportion of majority responses for the four datasets. (**B**) Self-reported confidences given by workers for their answers for the four seed datasets. Confidence was indicated by a 1–5 Likert scale.

Workers were asked to report their self-confidence in their responses on a 1–5 Likert scale (1-lowest confidence). Figure 2B reports the distributions of their scores for the first four datasets. Workers were more confident overall for the word association data than the ground truth data, plausible as the latter were taken from more technical domains. Workers responding to randomized variants of both datasets typically had lower confidence than workers responding to pairs taken from the original datasets

4 EFFICIENT EXPLORATION AT SCALE

A drawback of the single-link learning scheme studied in Experiment 1 is lack of scalability. To build a causal network requires asking potentially many workers to support or falsify each link within a single task, which may become cost prohibitive. To address this, here we introduce a multi-stage algorithm that efficiently explores a network such as a causal attribution network. The algorithm proceeds by first asking workers to develop a **pathway** or chain of causes and effects rooted by an initial seed term. This pathway is then taken by workers and modified into new pathways by adding new terms, removing old terms, and so forth. This process iterates, repeatedly developing new pathways from old ones. Lastly, the final network is taken from the union of all worker-derived pathways.

We describe the algorithm in the context of causal attributions, but this scheme can be applied with little or not modifications to any network where workers possess enough knowledge and context to offer helpful exploration. The full algorithm is summarized in Alg. 1. We implement this algorithm in our second experiment (Sec. 5) and explore how well it samples a known network computationally in Sec. 7.

4.1 Cause Proposal (CP)

The algorithm begins from a single seed cause. A researcher can choose this cause as a starting point based on her needs or research interests, or she can ask the crowd to propose a cause that is of interest to them. For the latter case, used in this study, the proposal phase can be broken down into two steps: first, workers propose a collection of potential causes, and then workers can rank that list of causes according to some desired criteria. This ranking can be developed by showing

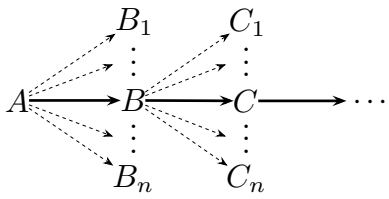


Fig. 3. Greedy Pathway Expansion (GPE) to learn the causal chain $A \rightarrow B \rightarrow C \rightarrow \cdots$. Starting from a seed cause A, n candidate effects $B_1, \ldots, B_n$ are proposed and validated by the crowd. The highest quality B_i, meaning the effect B_i for which the crowd most agrees on $A \rightarrow B$, is then chosen as the first link in the pathway. This B is then used as the start for the next iteration, which continues until N links along the pathway have been determined.

the workers a list and asking for their top choice, asking the workers to sort the list themselves, or by performing pairwise comparisons and then using a ranking algorithm [18] such as Rank Centrality [48] to select the top cause. The latter is most useful if many causes are proposed. The top ranked cause (or causes) can then be used as the seed for the next phase.

4.2 Greedy Pathway Expansion (GPE)

This phase begins from one or more previously developed seed causes. Workers are asked to propose effects for a given cause (*"What do you think is caused by 'poverty'?"*). The crowdsourcer develops n potential effects $\{B_i\}_{i=1}^{n}$ for cause A. Then, workers are asked to rank these effects from most plausible to least plausible and the top-ranked $B_i \equiv B$ is chosen. That cause-effect term pair (A, B) forms the first link in a chain of causes and effects. This process then repeats starting from B to learn the next link, and this continues until a full pathway $P = (A, B, C, \ldots)$ of length N is developed. An illustration of this greedy pathway expansion is shown in Fig. 3.

These propose-and-rank steps are very similar to those used in the previous proposal phase, and workers can easily be shared between these tasks, although one may wish to use different sets of workers for different phases if desired, for example, if there is a concern about possible bias in their responses.

It is also worth noting that workers asked to rank effects will see each link of the pathway independently, meaning they are asked to evaluate potential link (A, B_i), link (B, C_i), etc. They do not see, for example, the longer chain of causes and effects (A, B, C_i), and seeing the context of A may potentially change their selection of the best candidate effect C_i for a given cause B. The next phase of our algorithm addresses this, by allowing workers to see and modify longer pathways of causes and effects.

4.3 Iterative Pathway Refinement (IPR)

The goal of this phase of the crowdsourcing is to develop a larger set of pathways starting from the pathway(s) developed during the previous GPE phase. Here workers are shown a pathway using a web form that allows them to quickly edit and update the pathway. Updates can include insertions, deletions or rearrangements of the terms within the pathway, and workers can accomplish all of these tasks with a simple drag-and-drop interface in order to be as efficient as possible.

Each worker participating in this phase is shown a chosen pathway P_i, either the original GPE pathway or a subsequently developed pathway introduced by an earlier IPR worker. The worker uses the web form to modify P_i into a new pathway P_{i+1}. This new pathway is inserted back into the pathway set and the algorithm repeats for the next worker.

ALGORITHM 1: Multi-phase crowdsourcing algorithm for (causal) network exploration. All phases of the algorithm can run in parallel: the crowd can develop multiple seed causes, work simultaneously on multiple GPE pathways, and so forth.

(1) Cause Proposal (CP):
  (a) Workers propose n root causes A_i.
  (b) Workers rank these A_i's by interest, select top ranked $A \in \{A_i\}_{i=1}^{n}$ as the seed cause.

(2) Greedy Pathway Expansion (GPE) (Fig. 3):
  (a) Workers propose candidate effects B_i for seed cause A, leading to potential causal links $A \rightarrow B_i$.
  (b) Workers vote on these $A \rightarrow B_i$. Choose the top ranked $B \in \{B_i\}_{i=1}^{n}$, leading to causal link $A \rightarrow B$.
  (c) Repeat from 2a with B as the new root cause until a pathway $P_0 = (A, B, \ldots)$ of length N is achieved.

(3) Iterative Pathway Refinement (IPR):
  (a) Initialize a set of pathways $\mathcal{P}$ seeded by the initial GPE pathway, i.e. $\mathcal{P} = \{P_0\}$.
  (b) Workers iteratively refine pathways. A new worker is shown a selected pathway $P_i \in \mathcal{P}$, and uses a drag-and-drop interface (Fig. 4) to add, remove, and reorder terms in P_i to create P_{i+1}. The new pathway is added to the set: $\mathcal{P} \leftarrow \mathcal{P} \cup \{P_{i+1}\}$.
  (c) Repeat from 3b until a desired number of pathways are developed.

(4) Network Extraction:
  (a) Define a network $G = (V, E)$ by taking the union of all pathways developed in 3. Specifically,
  $V = \{T \mid T \in P_i, \text{ where } P_i \in \mathcal{P}\}$ and
  $E = \{(T_j, T_{j+1}) \mid T_j, T_{j+1} \in V, \exists P_i \in \mathcal{P} \text{ s.t. } (T_j, T_{j+1}) \text{ is a substring of } P_i\}$.

Because workers develop entire cause-effect pathways per microtask, this task structure provides considerably more information about the network per task than could be attained by the single-link (SL) learning experiment (Experiment 1). Of course, IPR is a more complex task than SL, so it is important to assess the *speed* of workers when investigating the efficiency of IPR, which we do in Experiment 2 (Sec. 5). By taking the union of all the pathways, a single causal attribution network can be estimated. Specifics details on network extraction are given in Algorithm 1.

Extensions and modifications

The multi-phase algorithm we propose here is modular and extensible in several ways, and a crowdsourcer is free to adapt the different components to meet her needs. For example, a crowdsourcer may skip the proposal of root terms if she already has a seed set to use. Likewise, one can bypass the Greedy Pathway Expansion phase and go directly to Iterative Pathway Refinement, either by using a pathway of one term as the original pathway seed set, or by starting from one or more predetermined pathways if they are available. The algorithm is easily parallelized, allowing the same crowd to contribute simultaneously to different network explorations, for example starting from different seeds. Lastly, while the algorithm is described and applied in this work to the case of directed causal attribution networks, it is by no means limited to such cases, and can serve as an effective crowd exploration method for any networks where the crowd is suitable for exploration. Indeed, Iterative Pathway Refinement is a special case of a more general exploration method—*Iterative Motif Refinement*—where workers construct and modify small subgraphs known as motifs [46], of which directed pathways are one such motif. We discuss such generalizations further in Sec. 8.

5 EXPERIMENT 2 — CAUSAL ATTRIBUTION AT SCALE

To demonstrate our crowdsourcing algorithm (Sec. 4), we implemented it on Amazon Mechanical Turk (AMT) as a collection of interrelated Human Intelligence Tasks (HITs). Our implementation proceeded primarily along the lines of the algorithm delineated in Sec. 4, with a few practically-motivated key differences. For example, multiple causes were proposed via the Cause Proposal HITs and multiple pathways were constructed simultaneously during the Greedy Pathway Expansion HITs.

5.1 Materials and methods

We implemented three HIT web interfaces for workers corresponding to the Cause Proposal (CP), Greedy Pathway Expansion (GPE), and Iterative Pathway Refinement (IPR) phases of the crowdsourcing algorithm detailed in Sec. 4. These are known as "external" HITs on AMT as the web forms are hosted on our own server that workers access through an iframe within the AMT website. At each phase we chose various criteria described below for the number of responses to collect, balancing the quantity of data needed with budgetary limits. Practitioners applying our algorithm or one similar to it will likely face similar decisions but this will generally depend on their particular circumstances. For all tasks, workers could not respond multiple times to the same information; they could not vote on the same cause-effect pair more than once, for example. This research procedure has been approved by our IRB (determination number CHRBSS: 15-039).

The CP task showed workers an example of a cause-effect pathway and asked the worker either to propose or to rank 3–5 causes that they believed "may have very surprising or unintended or interesting effects" (see App. A for screenshots of task interfaces with instructions). Care was taken with these instructions: we wanted the workers to have simple instructions that demonstrated causal relationships but did not prime them in terms of quantifying how interesting or important the causes should be. While we could have given them detailed selection criteria to follow (we plan to study this in future work), here we are mainly interested in determining what crowd workers consider important, as free from our influence as possible. This goal is in contrast with most crowdsourcing research that emphasizes the importance of detailed, exact instructions. In this phase, we sought 50 proposed causes, 20 worker rankings per cause, and we selected the top eight ranked causes. These eight root or seed causes were then passed to the GPE HITs.

For the GPE task, we used a HIT very similar to the CP HIT but now asking workers to either (i) propose effects for a given cause, or (ii) vote on effects by either agreeing or disagreeing with a given cause-effect pair. See App. A for the GPE instructions. At each step when growing a GPE pathway from a given cause, we sought 10 unique effects for that cause (part i), and we sought 5 different workers to validate each of the 10 effects (part ii). When finished, the effect which received the highest proportion of affirmative votes was selected as the next piece of the GPE pathway (any ties were broken at random) and the GPE task restarted with the newly chosen effect as the given cause. Note that the GPE phase was conducted on all 8 pathways in parallel, and workers could propose and/or vote on any tasks available in any of the chains. We continued this process until the GPE pathways had an average length of 5 terms. These pathways then seed the IPR HITs.

The IPR task is the most important phase of our algorithm, and subsequently had the most detailed, interactive web interface. Figure 4 shows a screenshot of the IPR interface. Unlike tasks in the previous phases, there is no switching between propose and vote subtasks. Workers were simply shown a causal pathway selected at random from the set of available pathways and asked to modify the pathway. The pathway was presented as a vertical list that workers could rearrange using drag-and-drop operations. Workers could also insert terms into the pathway and delete terms from the pathway using the same task interface. Upon receiving the modified pathway from the

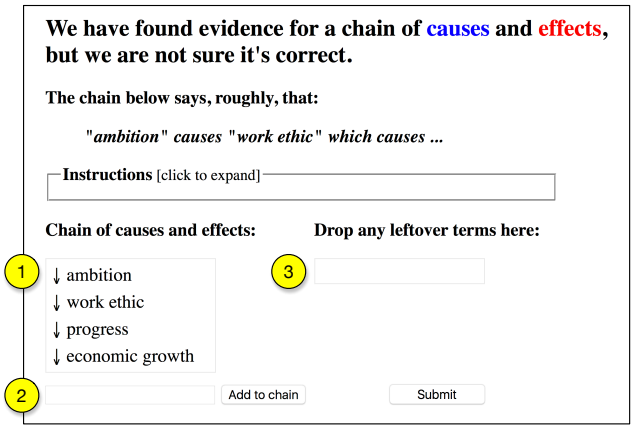


Fig. 4. Screenshot of the task interface for workers participating in Experiment 2's Iterative Pathway Refinement task. The pathway of terms listed in the lower left (Callout 1) is reorderable using drag-and-drop operations. Insertions and deletions can be made using the form as well (Callouts 2 and 3, respectively). The instructions box is collapsed in this view but is shown expanded before the worker chooses to accept the HIT (see Appendix A).

worker, we inserted it back into the pool of pathways alongside the original pathway. To prevent workers from modifying old pathways too often, pathways are flagged as 'unavailable' if more than 5 modified pathways were derived from it by workers. We sought 1500 IPR responses from workers.

5.2 Results

The three phases of the crowdsourcing algorithm (Cause Propose, Greedy Pathway Expansion, Iterative Pathway Refinement) were run sequentially on Amazon Mechanical Turk. We did not filter or manually remove any responses across the tasks. However, there is often a lag between giving a worker a task and receiving a response, so we often receive slightly more responses than necessary as the "stragglers trickle in". We consider this acceptable as it has a minor effect on our budget, but in principle a crowdsourcer could more carefully reserve tasks deployed to workers to prevent this. Altogether, 27 workers submitted 48 unique (50 total) root causes. We selected the top 8 ranked causes to initialize 8 GPE pathways. These pathways were then expanded by workers until they had an average length of 5 terms; we received 41 distinct terms across the 8 pathways. Lastly, IPR workers provided 1567 pathways, including the 8 original GPE pathways.

The net result is a collection of 1567 pathways representing 394 unique terms and 1329 distinct cause-effect term pairs. We summarize the numbers of responses, numbers of workers, and rewards per HIT for each phase in Table 1.

Likewise, Table 2 summarizes the eight initial seeds and subsequent pathways built by crowd workers during the first two phases of the experiment. The first term of each pathway is a seed developed during the first phase, while the pathway itself was developed during the second or GPE phase. All terms here and throughout the causal network were introduced by workers. Interestingly, there is a roughly even split between positive sentiment ("ambition", "education") and negative sentiment ("inequality", "fear") seed terms.

The GPE pathways shown in Table 2 were used to seed the pathway set developed during the Iterative Pathway Refinement (IPR) phase. Workers in this phase developed a total of 1567 distinct

Table 1. Summary of crowdsourcing tasks for Experiment 2. Rewards are in USD.

Task	# responses	# workers	reward per HIT
Cause Proposal:			
— Proposal	17	17	$0.17
— Rank	926	98	$0.17
Greedy Pathway Expansion:			
— Proposal	591	84	$0.07
— Vote	2136	122	$0.07
Iterative Pathway Refinement	1567	96	$0.17

Table 2. Initial pathways developed using Greedy Pathway Expansion in parallel during Experiment 2. The seed terms developed during the Cause Proposal task are the first terms of these pathways. All terms were introduced by workers. Pathways are listed in arbitrary order.

1) inequality → resentment → contempt → anger → mistakes → accidents → personal injury → pain → suffering
2) wealth → stability → comfort → security
3) poverty → hunger → starvation → death suffering misery → sadness
4) fear → sweating → dampness → mold → illness
5) curiosity → ideas → inventions → new products
6) ignorance → stupidity → bad mistakes → bad results
7) ambition → work ethic → economic growth → progress → growth
8) education → knowledge → curiosity → discovery → invention growth change

causal attribution pathways. Only 5 pathways contained a duplicate term and only one was a self-loop ("economic growth" caused "economic growth").

5.2.1 ***Patterns of pathway refinement.*** We explore the patterns of pathways and how workers edit pathways in Fig. 5. Using the IPR interface, workers given a path can make any of four ***edit operations***: insertion of a new term, deletion of an old term, substitution of an existing term with a new term (counted as a separate operation although it is a deletion and subsequent insertion at the same position), or the transposition of two existing terms. The first two edit operations are the most common choices workers make (Fig. 5A, B). At least one insertion occurred in over 50% of responses while at least one deletion occurred in over a third of responses (Fig. 5B). The mean length of pathways submitted by workers is 5.17 terms; the distribution of pathway lengths is shown in Fig. 5C.

Tracking the locations of the two most common edit operations performed by workers, insertions and deletions, both operations are more likely to occur towards the beginning of a given pathway than expected from the overall distribution of pathway length (Fig. 5D). Further, insertions occur more often than deletions at the end of pathways.

A bias in favor of edits located near the beginning of a pathway could be the result of cognitive mechanisms by which individuals process chains of causes and effects. But it could instead simply be evidence that workers are ***satisficing***, finding means to acceptably complete the IPR task as quickly as possible. If the latter, this could indicate that workers are preferentially ignoring the later terms in a pathway.

To investigate satisficing, in Fig. 6 we study the elapsed time workers spend on a given IPR task and the average positions of insertions and deletions within the pathway, both as functions of the initial length of the pathway shown for that task. If workers are satisficing we expect to observe an

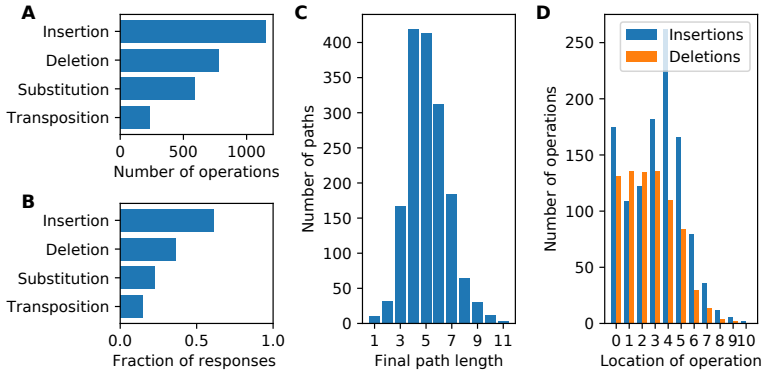


Fig. 5. Iterative Pathway Refinement (IPR) edit operations and cause-effect pathway lengths. (**A**) The numbers of each type of IPR edit operation. (**B**) The fraction of responses containing at least one edit operation. (**C**) The distribution of pathway lengths. (**D**) The distributions of locations of the two most common edit operations.

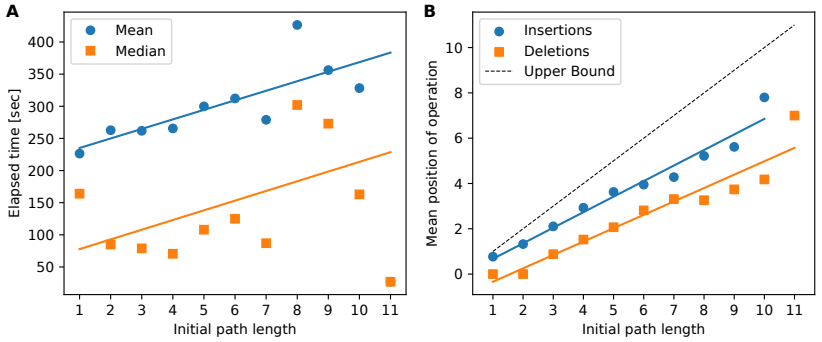


Fig. 6. Most workers do not "satisfice" the IPR task by focusing only at the start of a pathway, even when shown longer pathways. (**A**) The estimated elapsed time (in seconds) for IPR workers to submit their response as a function of the length of the path they are presented. (**B**) The average location of edit operations grows as the initial path length increases. Best-fit lines provide guides for the eye.

approximately flat trend in elapsed time versus initial path length, at least for longer pathways. Likewise, if satisficing, the average position of insertions and deletions would also tend towards a constant with pathway length. Neither occurs: the elapsed time grows roughly linearly with initial path length (Fig. 6A) as does the mean location of edits (Fig. 6B). Further, if some workers were satisficing by only studying the tail of the pathway instead of the head, we would observe a non-monotonic trend in Fig. 6, which does not occur. While satisficing is likely still occurring for some workers, these trends provide evidence that many if not most workers are taking the time to process and understand the full length of the pathway.

5.2.2 ***The causal attribution network.*** Taking the union of all IPR pathways in the Network Extraction phase (Sec. 4) yields the final causal attribution network developed by this experiment. This is a weighted, directed network where each edge (i, j) indicates that i "causes" j, and the edge weight w_{ij} associated with this edge denotes the number of IPR pathways containing the directed link (i, j). We visualize this network in Fig. 7.

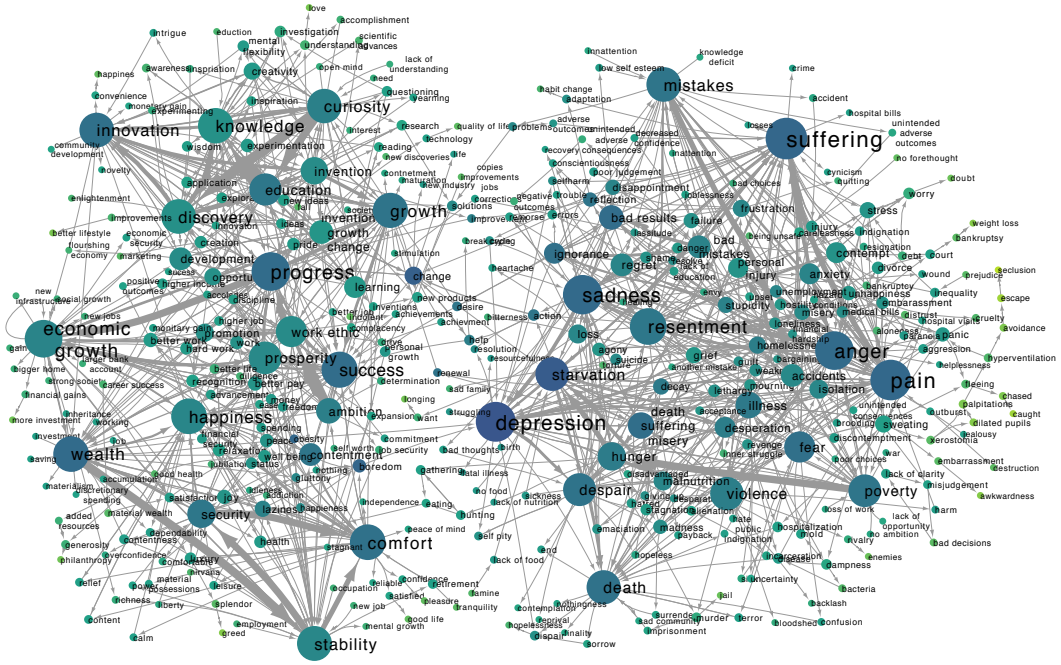


Fig. 7. A network visualization of the directed causal attribution network constructed by workers. Node size is proportional to degree (indegree + outdegree), node color proportional to closeness centrality, edge width proportional to edge weight.

We remark upon a number of interesting features of this causal network. Particularly intriguing is that the network is roughly clustered into two groups, one focused on positive terms ("progress", "growth", "success", etc.), and another around negative terms ("suffering", "resentment", "depression"). This clustering, which comes from the initial GPE pathways (Table 2) likely reflects known emotional and valence effects underlying causal attribution [7]. We see that workers are more likely to ideate causal relationships that are at the emotional extremes, either very positive or very negative.

Within the clusters, multiple causal links appear strongly supported by the crowd, for example, the links between stability, security, and comfort in the lower left of the network or the link from poverty to hunger (and not vice versa). There are also a number of cause-effect pairs where the crowd is split on the directionality: for example, "discovery causes knowledge" occurs in 57 pathways while "knowledge causes discovery" occurs in 64 pathways.

Turning our attention from the qualitative view of the network (Fig. 7), we next investigate quantitative properties of the network. Figure 8 shows the distributions of indegree, outdegree, edge weight (number of pathways containing a given cause-effect pair) and betweenness centrality (how often a term is in a shortest path in the network). In all cases, these statistics follow heavy-tailed, non-Gaussian distributions. Interestingly, after rescaling each quantity by its mean value (Fig 8, bottom row), the different quantities follow similar distributions.

Lastly, to evaluate the significance of the causal network structure, at least along one dimension, we consider the prevalence of ***feedforward loops*** and ***feedback loops*** in the network (Fig. 9A). Feedforward loops are an important structural motif in directed networks [46] and feedback loops are a key component of causal homeostasis [33]. We used the statistical procedure of Milo *et*

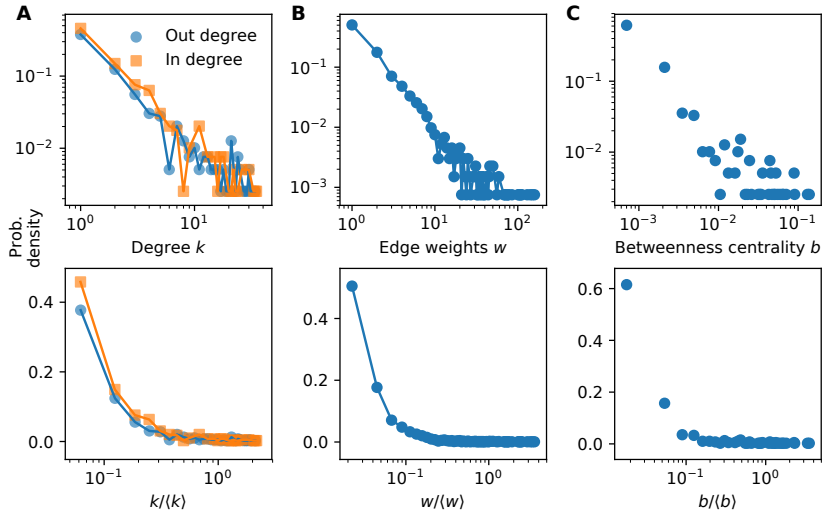


Fig. 8. Properties of the crowdsourced causal network shown in Fig. 7. (**A**) Degree distributions, (**B**) edge weight distribution, (**C**) Betweenness centrality distribution.

al. [46] to evaluate the significance of the three-node feedforward loop and feedback loop motifs. Specifically, we enumerated all feedforward and feedback loops in the crowdsourced network and compared that number in Fig. 9B with the same quantity computed from 5000 null networks generated using an in- and outdegree-preserving randomization procedure [46]. The observed network has significantly more feedforward loops and feedback loops than expected according to Milo *et al.*'s null model ($z \approx 30.99$, $p < 10^{-200}$ for feedforward loops; $z \approx 19.30$, $p < 10^{-80}$ for feedback loops). Expanding on our analysis of feedforward and feedback loops, Fig. 9C shows all the statistically significant three-node *motifs*, including their occurrence frequency and z-scores relative to the null model. Seven of the 16 possible three-node motifs were significant. Motifs in Fig. 9C were found using FANMOD with default parameters [64]. Understanding these structural motifs allows us to better characterize properties of the crowdsourced causal attribution network.

5.2.3 ***Speed of workers.*** The IPR task will explore more of a network with fewer worker responses than the "single link" task used in Experiment 1 (Sec. 3). But this is expected, as each IPR task is more complex and provides more information. But since the IPR task is so complex, it is likely that workers require more time to complete IPR tasks than single link tasks. Indeed, examining in Fig. 10 the distributions of ***elapsed times***, the delay between when a worker receives and submits a task shows exactly that: IPR workers require more time than single link workers to complete their task.

However, while slower elapsed times for the IPR task seems to imply that the crowdsourcing algorithm is not efficient in terms of time-to-explore, the direct comparison between these elapsed times is not necessarily appropriate. As workers provide information on multiple links within a single IPR task, it is better to compare the elapsed time ***per link*** for IPR workers with the elapsed time of single link workers. Even those links in the pathway that remain unmodified by the worker still receive an additional "vote" when taking the union of the pathway set, so normalizing the elapsed time relative to the total length of the IPR pathway accounts for the total information provided by the worker. Figure 10 shows that IPR workers provide information more quickly per link than single link workers. IPR workers had a mean elapsed time per link of 59.00 seconds,

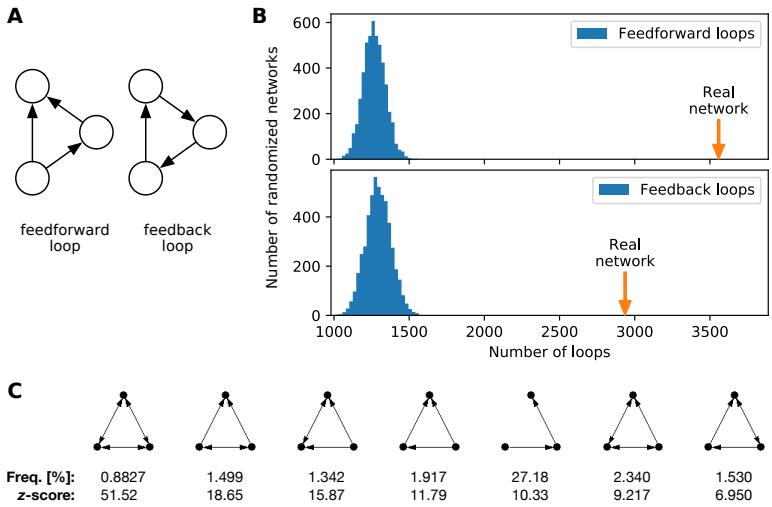


Fig. 9. The crowdsourced causal attribution network has significant feedforward loop, feedback loop, and other motif structure.

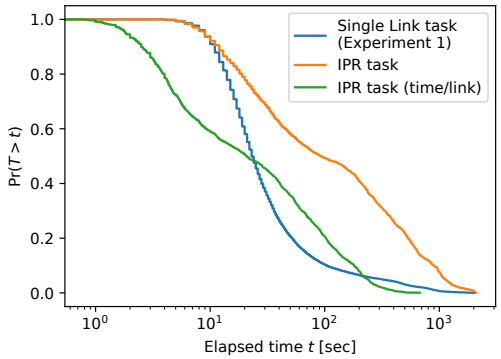


Fig. 10. Cumulative distributions of the elapsed time (as reported by AMT) for workers to complete tasks.

compared with 68.03 seconds on average for single link workers. This difference in distributions was significant (Mann-Whitney U test, $p < 10^{-15}$). Thus, the rate of information gain for IPR workers is significantly faster than single link workers.

5.3 Experiment 2 Discussion

Taken together, the crowd was able to generate a relatively large and meaningful causal attribution network using our multi-phase algorithm. Of course, the causal network as it stands after this experiment is not the final form of these data, and more work should be taken to filter the network. Specifically, terms that are overly vague ("unintended adverse outcomes") should be improved, the network likely contains synonymous terms that should be aggregated, and there remain a large number of terms on the periphery of the network that should be further explored. That said, the data collected from workers provide useful cause-and-effect relationships, possesses

statistically significant network structure, and the edit patterns of workers can help researchers further understand cognitive aspects of causal attribution.

The IPR task forming the core of the crowdsourcing algorithm enabled workers to be significantly faster at providing network information than workers asked to validate a single link per task, with IPR workers completing the task approximately 9 seconds faster per link on average. Such a time difference when compounded over a large network exploration can have a sizable impact on the overall completion time of the crowdsourcing. Further improvements on the interface of the IPR task, such as allowing workers to edit pathway terms in place instead of using drag-and-drop operations to update terms, will likely improve worker speed even more.

6 EVALUATING RESPONSE QUALITY

Experiment 2 (Sec. 5) provides evidence that workers can build a graph more efficiently with our algorithm (Sec. 4) than they could working one link at a time (as per Experiment 1 in Sec. 3). But does this efficiency come at a cost? Are we trading off quality for quantity?

While Experiment 1 studies how workers attribute cause and effect using a small ground truth dataset, it is challenging to systematically investigate the quality of Experiment 2's network as no ground truth is available. In light of this and to provide at least some information on quality, we performed a followup crowdsourcing survey asking workers to examine results generated during Experiment 2. We presented new crowd workers with potential cause-effect pairs taken from the causal attribution network (using questions of the form "*Do you think that A causes B?*") and asked them if they agree or disagree (or were uncertain) with these attributions. The goal of this survey is for workers to signal some degree of relative quality (or at least consistency) by assessing previously generated causal attributions. This survey task is identical in form to the 'vote' phase of Greedy Pathway Expansion (Sec. 4). To provide an independent assessment, workers who participated in Experiment 2 were excluded from this followup task.

For this new task we extracted two groups of $n = 50$ cause-effect pairs ($n = 100$ total) sampled at random from Experiment 2's graph:

(1) In the first group, each cause-effect pair (i, j) was a directed link within the causal attribution graph. Half ($n = 25$) of the sampled links had low weight ($w_{ij} \leq$ median link weight) while half had high weight ($w_{ij} >$ median link weight). Since link weight w is the number of worker responses containing the link, we assume "stronger" links are more likely to be approved.
(2) Our second group focuses on disconnected nodes, asking workers to agree/disagree with potential causal attributions not captured in the current network. Half ($n = 25$) of the node pairs were chosen at random from those pairs at distance $d = 2$ in the network, meaning they were only two hops apart on the network. It is plausible that some or even many of these pairs have a causal relationship yet to be introduced by workers since, if the link between them existed, they would form a feedforward loop (Sec. 5.2.2). The second half of the group consists of pairs of nodes at distance $d > 2$, capturing more distant nodes. We expect these node pairs to be less likely to hold a causal relationship, although some may have a relationship.

We collected 984 responses from 114 workers with each of the $n = 100$ cause-effect pairs examined by at least 8 workers. Workers were compensated \$0.07 per response and each worker was limited to at most 25 responses to provide room for many different workers to respond. This research procedure has been approved by our IRB (determination number CHRBSS: 15-039).

The results of these followup surveys are summarized in Fig. 11 where we present the fraction of workers who approved or were uncertain about a potential cause-effect pair. Most workers tended to approve of links already present in the network, even the low weight links (although workers were more uncertain about low-weight links). Workers also approved causal attributions for most

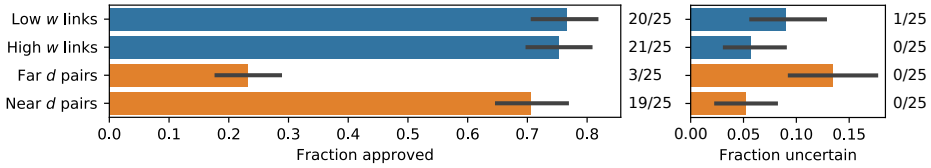


Fig. 11. A followup survey asking workers who did not participate in Experiment 2 if they agree, disagree or were uncertain about cause-effect links and unlinked potential cause-effect pairs taken from the network generated by Experiment 2. Errorbars denote 95% C.I. Plotted bars denote the response fractions pooled across all surveyed links or pairs; numbers to the right denote counts of links or pairs where the *majority* of responses to that individual link or pair agreed.

of the near distance ($d = 2$), unlinked pairs but were significantly less likely to approve cause-effect pairs at high distance (z-test on proportions: $z = -10.5, p < 10^{-25}$) (and workers were significantly more likely to be uncertain about high distance pairs; $z = 3.14, p < 0.002$). While further study is needed, these survey results do fit with our expectations if the data generated during Experiment 2 were of high quality, providing some evidence that the efficiency gains from our algorithms were not offset by lower quality or at least lower consistency, as perceived by other crowd workers.

7 COMPUTATIONAL STUDY — BIAS AND ROBUSTNESS OF NETWORK EXPLORATION

In this section we explore the differences between a true network and one derived from a sampling procedure based on Iterative Pathway Refinement (IPR). We introduce a simplified computational model to simulate workers generating pathways and apply this model within the IPR algorithm on a known network topology to test IPR's performance. As such, there are three components to this computational study: (i) a model of the underlying graph being explored (Sec. 7.1), (ii) a model for how crowd workers develop pathways taken from the underlying graph (Sec. 7.2), and (iii) a set of network metrics to compare the true underlying graph with the sampled graph derived from a finite number of worker pathways (Sec. 7.3). We describe the results of our computational study using these components in Sec. 7.4.

7.1 Modeling the underlying network

We study two directed network models for the underlying graph G_{true}:

Random graph The directed analog of the undirected Erdős-Rényi (ER) graph [19]. Here N nodes are introduced and each possible edge (i, j) exists independently with constant probability p. To make this network directed, both edge (i, j) and (j, i) exist with probability p, giving a maximum of $2\binom{N}{2}$ possible edges.

Scale-Free graph The most common model for a scale-free graph is the Barabási-Albert (BA) graph [4]. Here N nodes are introduced one at a time starting from a small seed graph. Each new node attaches to m existing nodes with probability proportional to the existing nodes' degrees. This leads to a rich-get-richer feedback mechanism as nodes with higher degree receive more incoming links giving them higher degree still.

Traditionally, the BA model is taken as an undirected network. The most obvious way to form a directed graph is where each newly introduced node i forms a directed link (i, j) when it attaches to a preexisting node j. Unfortunately, this creates a "temporal" hierarchy of newer nodes pointing backwards to older nodes, and makes it impossible for any local search algorithm to explore the entire network. To address this, we instead model directed

edges as being equally likely to point in either direction, i.e., when a new node i attaches to a preexisting node j, directed edge (i, j) is created with probability 1/2, otherwise directed edge (j, i) is added to the graph.

We took $N = 1000$, $p = 5/999$, and $m = 3$ to simulate these networks, and only consider the giant connected component of G_{true} if it is disconnected.

The choice of these two model networks is meant to cover the extremes of possible degree distributions of the underlying network G_{true}. If there is a strong dependence between the degree distribution and the efficiency of exploration, then it is likely that custom exploration algorithms will be needed for different networks. Conversely, however, if there is at most a weak dependence, then that means the exploration algorithm we propose may be able to efficiently explore a general class of networks.

7.2 Modeling pathway refinement

Given a true, latent network to be explored, we seek a simple model for how workers could reveal that network structure via individual IPR microtasks. Network search models, where an agent moves locally over a network exploring its structure, provide inspiration. Here we treat the pathway refinement problem as one of a short-term local search of a network. A searcher, starting from a randomly chosen node in G_{sampled} and respecting edge directions, takes a small number of moves to connected nodes, either in G_{sampled} or in G_{true}. The trace of the searcher can then mimic a new pathway generated by a worker proposing existing or novel terms.

Specifically, we model pathway generation using **self-avoiding walks** (SAW). A SAW on a graph is a variant of a random walk where the random walker is constrained such that it cannot visit the same node more than once [16, 24, 41]. Of course, independent SAWs my overlap, traversing some of the same nodes and links as other SAWs. Short-length SAWs can then represent causal pathways that do not repeat terms—almost no pathways generated during Experiment 2 contained duplicate terms (Sec. 5). Self-avoiding walks have good properties for efficiently exploring unknown networks [1, 65], providing a theoretical underpinning for the IPR algorithm.

To simulate IPR exploring a latent network with this model, at each timestep a new pathway is generated by an independent self-avoiding walker initialized at a node within G_{sampled}. This walk proceeds for a number of steps L, and any new nodes or links the walker happens to travel over—those in G_{true} not yet in G_{sampled}—are added to G_{sampled}. In our simulations, the length of each pathway L was drawn from a shifted Poisson distribution with fixed mean $\lambda = 1$, i.e. $L \sim \text{Pois}(\lambda = 1) + \ell$. We selected constant values of $\ell = 4, 6, 8$ that bracket the mean path length observed in Experiment 2 (see Fig. 6). The stochasticity introduced by this distribution is intended to better mimic the variation in length of actual crowdsourced pathways, although the exact form of the statistical model is not crucial. Further, choosing (constant) $L = 1$ corresponds to the Single Link task studied in Experiment 1 (Sec. 3), so we can compare simulated IPR exploration with Single Link exploration. This model can thus build G_{sampled} out of G_{true} over time by taking the union of all the simple model pathways created, providing a plausible, though idealized, simulation of the Iterative Pathway Refinement crowdsourcing phase.

7.3 Network exploration metrics

Suppose G_{sampled} is derived from simulating our IPR model (Sec. 7.2) until a total of $|P|$ pathways have been developed. To understand the effects of generating a network via this algorithm requires comparing G_{sampled} with the true, latent network G_{true}. Of course, in practice G_{true} is unavailable for such comparisons, but the synthetic analysis performed here helps us understand, in some manner, the performance and properties of our crowdsourcing algorithm.

To compare G_{sampled} and G_{true} we use the following network metrics:

Fraction of nodes and links observed Here we track how much of the underlying network G_{true} is captured in G_{sampled} by comparing the numbers of nodes N and links M in the observed graph with the sizes of the true graph: $N_{\text{sampled}}/N_{\text{true}}$ and $M_{\text{sampled}}/M_{\text{true}}$ as functions of $|P|$.

Average degree We also track the mean indegree and outdegree $\langle k \rangle$ of nodes in G_{sampled} compared with the same quantities for G_{true}. In the context of causal attribution networks, indegree captures the number of causes an effect has, while outdegree captures the number of effects that are due to a cause. This metric captures what portion of the neighbors of nodes are retained by the sampling algorithm.

Betweenness centrality The betweenness centrality $B(v)$ of a node v in G is the sum of the proportion of all shortest paths that pass through v [10]:

$$B(v; G) = \sum_{s,t \in V} \frac{\sigma(s, t|v)}{\sigma(s, t)} \tag{1}$$

where $\sigma(s, t)$ is the number of shortest paths connecting nodes s and t, and $\sigma(s, t|v)$ is the number of those paths passing through another node v ($v \neq s$, $v \neq t$). To compare G_{sampled} and G_{true}, we take the ratio $B_{\text{sampled}}/B_{\text{true}}$ using the mean node betweennesses $B_{\text{sampled}} = \frac{1}{N_{\text{sampled}}} \sum_{v \in V_{\text{sampled}}} B(v; G_{\text{sampled}})$ and $B_{\text{true}} = \frac{1}{N_{\text{true}}} \sum_{v \in V_{\text{true}}} B(v; G_{\text{true}})$.

The focus here is on relatively basic metrics that quantify global differences in the networks, but many other network metrics and statistics are available. For example, one statistic commonly used to quantify random walkers searching a network, particularly in the statistical physics literature, is the mean first-passage time, measuring how long it takes an individual random walker to reach a particular node [53, 65]. Yet, this measure is most suitable for a single random walk that moves over the network, whereas our model focuses on the collective action of a large number of shorter walks, so we instead focus on metrics suitable to quantifying the proportion of the true network that has been revealed, and if and how properties of G_{sampled} differ from those of G_{true}.

7.4 Results

Figure 12 presents the results of our computational study (simulating 100 independent runs for each combination of parameters). For both the random and scale-free networks, IPR was able to explore the latent G_{true} more quickly than the single link task: The fraction of nodes discovered, fraction of links discovered, and average node degree all converged more quickly to their true values, especially for longer average L. For example, IPR with $\ell = 4$ revealed 90% of the nodes in the Scale-Free graph on average with only ≈ 940 IPR tasks while over 4200 single link tasks were necessary to reach the same proportion, a factor of 4.6 more tasks. Of course, longer pathways should necessarily reveal the network more quickly, because each pathway provides more subnetwork structure, but in practice this will be complemented by the faster speed of IPR workers per link observed in Sec. 5.2.3. Overall, IPR is effective at exploring different types of networks.

Surprisingly, unlike the other metrics, the mean betweenness for G_{sampled} is generically *higher* than it is for G_{true}. This reveals an important bias: short-duration explorations will over-report the centralities of nodes, even though betweenness centrality (Eq. (1)) already accounts for the total number of shortest paths in G_{sampled}. The origin of this bias is that there are fewer links early on in the exploration, leading to fewer alternate shortest paths and so more paths will "flow" through the currently observed shortest paths. As more nodes and links are added, new shortest paths will appear, taking away some of the load carried by the existing shortest paths. Thus, a crowdsourcer exploring a network should account for biases in betweenness centrality and possibly other centrality measures. This biased overestimation of betweenness at low numbers of samples is

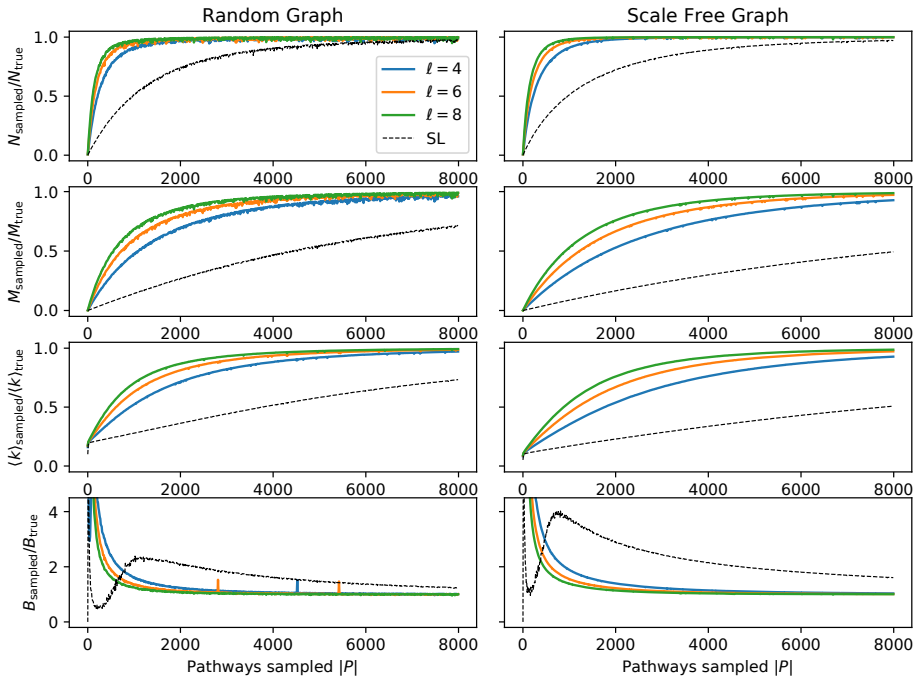


Fig. 12. Convergence of G_{sampled} to G_{true} using the self-avoiding walk (SAW) model of IPR. The Single Link (SL) learning baseline represents the task studied in Experiment 1 (Sec. 3).

far stronger for the single-link sampling method than for the IPR model, further emphasizing the usefulness of the IPR method.

8 DISCUSSION AND FUTURE WORK

Establishing causal relationships is one of the most important and challenging tasks of science. This work focused on crowdsourcing causal attribution networks, directed networks where links $A \longrightarrow B$ denote that a term A is a cause and the term B is its effect. However, much of the methodology contributed in this work can inform the exploration of other types of networks, for example, surveying individuals and their social ties to map out a social network [23]. Many technical networks can be explored if the participants possess the relevant expertise. A team of cardiology researchers and bioinformaticians, for instance, could collaboratively build a network of gene pathways related to cardiac arrest. Such domain-specific cases are an especially interesting avenue to combine our network exploration method with statistical methods and data, such as gene expression datasets for the cardiology example.

The crowdsourced causal attribution network derived by this study warrants further refinement, by merging synonymous terms, filtering out unsupported links, "backfilling" incomplete portions of the network as needed, and otherwise accounting for the human perceptual biases that influence causal reasoning. Indeed, while our first experiment provided evidence that workers are generally able to recognize true causal relationships, workers also make false positives, attributing cause and effect relationships where none exist. Measuring and accounting for this bias remains an important challenge when developing larger causal networks.

Our multiphase network crowdsourcing algorithm enables ***indirect collaboration*** between workers, as workers only interact with the responses given by previous workers when viewing those responses as part of a subsequently constructed task. Direct collaboration, where teams of individuals work together to build causal pathways, is an interesting area we plan to explore. This approach may lead to better causal attribution data, novel information on how teams perform, and improved understanding of the perceptual and attributional biases of causal relationships.

Iterative Pathway Refinement (IPR) is the core of our crowdsourcing algorithm. Our experiments, followup survey, and computational study provide evidence that IPR can leverage crowdsourcing to explore a large network relatively efficiently. IPR enabled workers to provide information efficiently and is well motivated by prior theoretical work on exploring networks using self-avoiding walks. Workers complete IPR tasks by modifying short linear causal pathways or chains. IPR is a special case of the more general ***Iterative Motif Refinement*** (IMR), as these linear chains can easily generalize to other types of motifs. Chains were chosen for IPR as they are relatively easy to understand in the context of cause-and-effect, but other network motifs can be used as well.

Indeed, studying motifs as part of IMR is a particularly fruitful line of future research, as different motifs may work best when exploring different types of networks, particularly when accounting for differences in human judgment and perception across the different types of networks. For example, a chain (dipath) may be most appropriate for a causal graph, while a triangle motif may be better when crowdsourcing a social network, as triadic closure is one of the most important properties of social networks [52]. If training data are available, one could even in principle use supervised learning to determine the ideal motif for a given network automatically.

Iterative Pathway Refinement can also be improved in other ways. We focused on one of the *simplest possible forms* of IPR, where existing pathways were chosen uniformly at random to show to workers. Although the same pathway was prevented from being shown too frequently (by limiting the total number of responses to a given pathway), IPR would be more efficient still if existing pathways were selected for updating by incorporating our current understanding of the quality or correctness of the given pathway. This would better distribute the crowd by preventing repeated refinement of known good pathways and instead guide the crowd towards those pathways that are the most fruitful to explore. Likewise, network-specific algorithms such as IPR can be complemented by non-network crowdsourcing algorithms such as efficient response aggregation strategies [15, 32, 40]. Algorithms that balance an exploration-exploitation trade-off, such as Thompson sampling [12], are also likely to improve the exploration efficiency of IPR even further.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful comments and the crowd workers whose participation made this research possible. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1447634.

REFERENCES

[1] Lada A Adamic, Rajan M Lukose, Amit R Puniyani, and Bernardo A Huberman. 2001. Search in power-law networks. *Physical review E* 64, 4 (2001), 046135.

[2] Cecilia R. Aragon and Alison Williams. 2011. Collaborative Creativity: A Complex Systems Model with Distributed Affect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 1875–1884. 00033.

[3] James P Bagrow. 2018. Crowd ideation of supervised learning problems. *arXiv preprint arXiv:1802.05101* (2018).

[4] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

[5] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2015. Soylent: a word processor with a crowd inside. *Commun. ACM* 58, 8 (2015), 85–94. 00607.
[6] Kirsten E. Bevelander, Kirsikka Kaipainen, Robert Swain, Simone Dohle, Josh C. Bongard, Paul D. H. Hines, and Brian Wansink. 2014. Crowdsourcing Novel Childhood Predictors of Adult Obesity. *PLOS ONE* 9, 2 (2014), e87756. 00019.
[7] Gerd Bohner, Herbert Bless, Norbert Schwarz, and Fritz Strack. 1988. What triggers causal attributions? The impact of valence and subjective probability. *European Journal of Social Psychology* 18, 4 (1988), 335–345.
[8] Josh C. Bongard, Paul DH Hines, Dylan Conger, Peter Hurd, and Zhenyu Lu. 2013. Crowdsourcing predictors of behavioral outcomes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 43, 1 (2013), 176–185. 00024.
[9] Daren C Brabham. 2008. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence* 14, 1 (2008), 75–90.
[10] Ulrik Brandes. 2008. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks* 30, 2 (2008), 136–145.
[11] Roger Brown and Deborah Fish. 1983. The psychological causality implicit in language. *Cognition* 14, 3 (1983), 237–273.
[12] Olivier Chapelle and Lihong Li. 2011. An Empirical Evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2249–2257.
[13] Justin Cheng, Jaime Teevan, Shamsi T Iqbal, and Michael S Bernstein. 2015. Break it down: A comparison of macro- and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 4061–4064.
[14] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1999–2008. 00117.
[15] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics* (1979), 20–28.
[16] Pierre-Gilles De Gennes. 1979. *Scaling concepts in polymer physics*. Cornell university press.
[17] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. 2015. The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 238–247.
[18] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*. ACM, 613–622.
[19] Paul Erdös and Alfréd Rényi. 1959. On random graphs, I. *Publicationes Mathematicae (Debrecen)* 6 (1959), 290–297.
[20] Enrique Estellés-Arolas and Fernando González-Ladrón-De-Guevara. 2012. Towards an integrated crowdsourcing definition. *Journal of Information science* 38, 2 (2012), 189–200.
[21] Roxana Girju, Dan Moldovan, et al. 2002. Text Mining for Causal Relations. In *FLAIRS Conference*. 360–364.
[22] Clive W J Granger. 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica: Journal of the Econometric Society* (1969), 424–438.
[23] Douglas D Heckathorn. 1997. Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems* 44, 2 (1997), 174–199.
[24] Carlos P Herrero. 2005. Self-avoiding walks on scale-free networks. *Physical Review E* 71, 1 (2005), 016103.
[25] Steven M Hill, Laura M Heiser, Thomas Cokelaer, Michael Unger, Nicole K Nesser, Daniel E Carlin, Yang Zhang, Artem Sokolov, Evan O Paull, Chris K Wong, et al. 2016. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature Methods* 13, 4 (2016), 310.
[26] Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65.
[27] Jeff Howe. 2006. The rise of crowdsourcing. *Wired magazine* 14, 6 (2006), 1–4.
[28] David Hume. 2012. *A Treatise of Human Nature*. Courier Corporation.
[29] Jason T Jacques and Per Ola Kristensson. 2013. Crowdsourcing a HIT: measuring workers' pre-task interactions on microtask markets. In *First AAAI Conference on Human Computation and Crowdsourcing*.
[30] RB Joynson. 1971. Michotte's Experimental Methods. *British Journal of Psychology* 62, 3 (1971), 293–302.
[31] Immanuel Kant and Paul Guyer. 1998. *Critique of Pure Reason*. Cambridge University Press.
[32] David R Karger, Sewoong Oh, and Devavrat Shah. 2011. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*. 1953–1961.
[33] Frank C Keil. 2006. Explanation and understanding. *Annu. Rev. Psychol.* 57 (2006), 227–254.
[34] Harold H Kelley. 1967. Attribution Theory in Social Psychology.. In *Nebraska symposium on motivation*. University of Nebraska Press.
[35] Hyun Duk Kim, Malu Castellanos, Meichun Hsu, ChengXiang Zhai, Thomas Rietz, and Daniel Diermeier. 2013. Mining Causal Topics in Text Data: Iterative Topic Modeling with Time Series Feedback. In *Proceedings of the 22nd ACM*

international conference on Conference on information & knowledge management. ACM, 885–890.
[36] Aniket Kittur. 2010. Crowdsourcing, collaboration and creativity. *XRDS: crossroads, the ACM magazine for students* 17, 2 (2010), 22–26.
[37] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 453–456.
[38] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 43–52.
[39] Jon M Kleinberg. 2000. Navigation in a small world. *Nature* 406, 6798 (2000), 845.
[40] Justin Kruger, Ulle Endriss, Raquel Fernández, and Ciyang Qing. 2014. Axiomatic analysis of aggregation methods for collective annotation. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1185–1192.
[41] Bin Li, Neal Madras, and Alan D Sokal. 1995. Critical exponents, hyperscaling, and universal amplitude ratios for two-and three-dimensional self-avoiding walks. *Journal of Statistical Physics* 80, 3-4 (1995), 661–754.
[42] Qi Li, Fenglong Ma, Jing Gao, Lu Su, and Christopher J. Quinn. 2016. Crowdsourcing High Quality Labels with a Tight Budget. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)*. ACM, New York, NY, USA, 237–246.
[43] Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. 2010. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 57–66.
[44] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Andrej Aderhold, Richard Bonneau, Yukun Chen, et al. 2012. Wisdom of crowds for robust gene network inference. *Nature methods* 9, 8 (2012), 796.
[45] Thomas C. McAndrew, Elizaveta Guseva, and James P. Bagrow. 2017. Reply & Supply: Efficient crowdsourcing when workers do more than answer questions. *PLOS ONE* 12, 8 (2017), e69829. arXiv:1611.00954
[46] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: simple building blocks of complex networks. *Science* 298, 5594 (2002), 824–827.
[47] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research* 17, 1 (2016), 1103–1204.
[48] Sahand Negahban, Sewoong Oh, and Devavrat Shah. 2017. Rank Centrality: Ranking from Pairwise Comparisons. *Operations Research* 65, 1 (2017), 266–287.
[49] Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* 36, 3 (2004), 402–407.
[50] Judea Pearl. 2009. *Causality*. Cambridge university press.
[51] Robert J. Prill, Julio Saez-Rodriguez, Leonidas G. Alexopoulos, Peter K. Sorger, and Gustavo Stolovitzky. 2011. Crowdsourcing Network Inference: The DREAM Predictive Signaling Network Challenge. *Science Signaling* 4, 189 (2011), mr7–mr7.
[52] Anatol Rapoport. 1953. Spread of information through a population with socio-structural bias: I. Assumption of transitivity. *The bulletin of mathematical biophysics* 15, 4 (01 Dec 1953), 523–533.
[53] Sidney Redner. 2001. *A guide to first-passage processes*. Cambridge University Press.
[54] Martin Rolfs, Michael Dambacher, and Patrick Cavanagh. 2013. Visual Adaptation of the Perception of Causality. *Current Biology* 23, 3 (2013), 250–254.
[55] Donald B Rubin. 2011. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *J. Amer. Statist. Assoc.* (2011).
[56] Matthew J Salganik and Karen EC Levy. 2015. Wiki surveys: Open and quantifiable social data collection. *PLOS ONE* 10, 5 (2015), e0123483.
[57] Brian J Scholl and Patrice D Tremoulet. 2000. Perceptual causality and animacy. *Trends in cognitive sciences* 4, 8 (2000), 299–309.
[58] Pao Siangliulue, Kenneth C Arnold, Krzysztof Z Gajos, and Steven P Dow. 2015. Toward collaborative ideation at scale: Leveraging ideas from others to generate more creative and diverse ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 937–945.
[59] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 254–263.
[60] Shelley E Taylor and Susan T Fiske. 1975. Point of View and Perceptions of Causality. *Journal of Personality and Social Psychology* 32, 3 (1975), 439.

[61] Jaime Teevan, Shamsi T. Iqbal, and Curtis von Veh. 2016. Supporting Collaborative Writing with Microtasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2657–2668.
[62] Jeffrey Travers and Stanley Milgram. 1967. The small world problem. *Phychology Today* 1, 1 (1967), 61–67.
[63] M. D. Wagy, J. C. Bongard, J. P. Bagrow, and P. D. H. Hines. 2017. Crowdsourcing Predictors of Residential Electric Energy Usage. *IEEE Systems Journal* PP, 99 (2017), 1–10.
[64] Sebastian Wernicke and Florian Rasche. 2006. FANMOD: a tool for fast network motif detection. *Bioinformatics* 22, 9 (2006), 1152–1153.
[65] Shi-Jie Yang. 2005. Exploring complex networks by walking on them. *Phys. Rev. E* 71 (Jan 2005), 016107. Issue 1.

A MECHANICAL TURK WORKER INSTRUCTIONS

Here are the instructional pages AMT workers were shown before they accepted each type of HIT comprising Experiment 2. Figure 13 shows the instructional page for workers asked to propose or rank seed causes, Fig. 14(a) shows the instructional page for workers participating in the Greedy Pathway Expansion phase, and Fig. 14(b) shows the instructional page for workers participating in the Iterative Pathway Refinement phase. Figure 14(b) shows the contents of the collapsed instructions box found in Fig. 4; the box was expanded for the instructional page but collapsed by default during actual HITs. The "vote" phase shown at the bottom of Fig. 14(a) is also the task interface used for the response quality survey (Sec. 6).

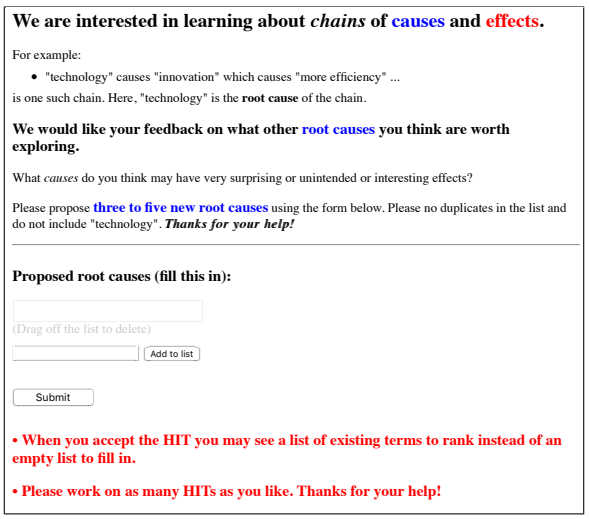
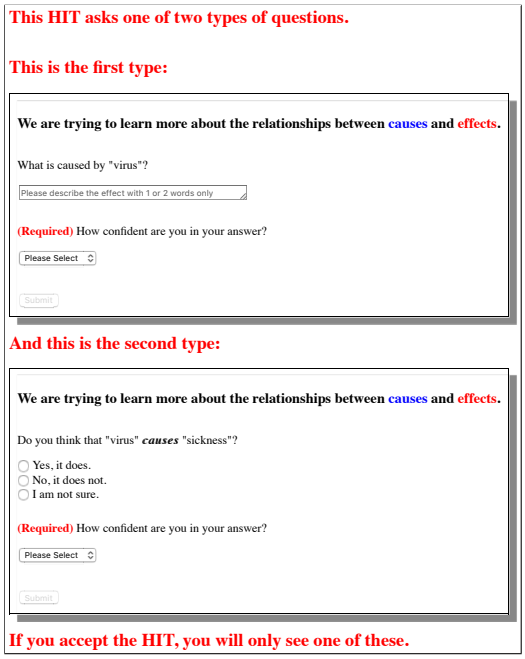
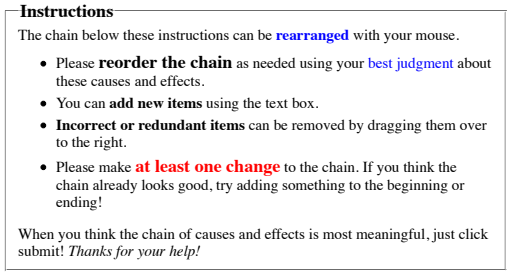


Fig. 13. Screenshot of the instructional page shown to Amazon Mechanical Turk workers before accepting the Cause Proposal task that initialized Experiment 2.

Received April 2018; revised July 2018; accepted September 2018



(a) Greedy Pathway Expansion



(b) Iterative Pathway Refinement (cf. Fig. 4)

Fig. 14. Screenshots of the instructional pages shown to Amazon Mechanical Turk workers before accepting (a) Greedy Pathway Expansion tasks and (b) Iterative Pathway Refinement tasks. For GPE tasks, a worker will be asked to either propose a new effect or to validate an existing cause-effect pair, depending on the state of the algorithm, and so worker instructions cover both tasks.