Preserving Privacy in Crowd-Powered Systems

Walter S. Lasecki[1], Mitchell Gordon[1],
Jaime Teevan[2], Ece Kamar[2], Jeffrey P. Bigham[3]

University of Rochester[1], Microsoft Research[2], Carnegie Mellon University[3]

Abstract. It can be hard to automatically identify sensitive content in images or other media because significant context is often necessary to interpret noisy content and complex notions of sensitivity. Online crowds can help computers interpret information that cannot be understood algorithmically. However, systems that use this approach can unwittingly show workers information that should remain private. For instance, images sent to the crowd may accidentally include faces or geographic identifiers in the background, and information pertaining to a task (e.g., the amount of a bill) may appear alongside private information (e.g., an account number). This paper introduces an approach for using crowds to filter information from sensory data that should remain private, while retaining information needed to complete a specified task. The pyramid workflow that we introduce allows crowd workers to identify private information while never having complete access to the (potentially private) information they are filtering. Our approach is flexible, easily configurable, and can protect user information in settings where automated approaches fail. Our experiments with 4685 crowd workers show that it performs significantly better than previous approaches.

1 Introduction

There can be considerable value in sharing potentially sensitive content. For example, people may want to share photo collections with their friends via social media while not revealing private details. Companies may want to share customer data without exposing themselves to legal liability or releasing competitive information. And the users of crowd systems may want the assistance of the crowd in managing their personal information [4,3,9,10] without malicious crowd workers using their data inappropriately [12]. If sensitive content could be easily and robustly filtered from large datasets, it would be possible to capture this value with limited risk of exposure. Unfortunately, even the best automatic approaches can fail because they require a rich understanding of the content and the ways that it might be used. For example, an automated system may help a user filter their account number from a picture of their bank statement by masking all of the numbers, but this approach would make it impossible to then extract the customer service phone number from the picture. This paper presents a crowd-powered system that uses human intelligence to filter sensitive information while limiting the amount of information any individual crowd worker encounters.

We aim to filter sensitive content from images. User can provide natural language descriptions of what to filter (e.g., “Remove the faces of children”), which requires no technical skills or knowledge of the system’s underlying processes. We then choose from multiple content segmentation approaches to distribute the filtering tasks to different crowd workers, showing each of them only a small portion of the original image.

The segment granularity is optimized by the system based on user-specified budget constraints using a novel method that employs multiple segmentation levels to avoid problems arising from unknown content granularity (e.g., the system does not know a priori how large a face is in a given image).

An increasing number of crowd-powered systems require workers to interact with user-generated data, such as audio recordings [10], personal photographs [4], email [9], documents [3], search queries [1], and handwritten text [14,5]. These systems can accidentally expose information users would like to remain private to the workers powering the system, but require exposing very similar content to function properly. For instance, a blind user of VizWiz [4] may want the crowd to identify the name of a prescription medicine from a photograph of the bottle. Because prescription labels typically contain the patient's name, the crowd can only provide an answer without learning the user's identity if the patient's name is obscured, but the medicine's isn't. To address this challenge, we introduce a pre-processing step in the crowd pipeline.

We explore several ways of dividing content to protect sensitive information during the filtering stage, and introduce a novel *pyramid workflow* that uses multiple segment sizes to help overcome the problems with fixed-size segmentation approaches that have trouble when sensitive information appears at different sizes.To avoid filtering content needed to complete the underlying task, workers add semantic labels to the masks they generate. Our experiments with 4685 Mechanical Turk workers show that we can mask a variety of types of sensitive content from images without exposing it in the process, and while making it possible for a disjoint crowd of workers to perform the base task.

2 Related Work

We begin by outlining the types of tasks that crowdsourcing platforms employ that require workers to interact with end user information, and discuss the threats crowd workers pose to such systems.

2.1 Crowd Platforms that Expose User Data

Many crowd-powered systems assist users in their daily lives, often using data from users. For example, Soylent helps users edit documents [3], and PlateMate determines how many calories meals contain based on photographs of them [15]. The intended tasks are generally not expected to contain sensitive information, but nevertheless may. For example, PlateMate may receive an image from a diner at a restaurant that accidentally includes a credit card on a table. Because crowd-powered systems can easily be confused with automated systems by end users, exposure can happen unintentionally. Interactive crowd-systems that respond to users in real time [11] also make it easy to mistakenly capture sensitive information. Getting responses from the crowd in a few seconds [2] means there is little time for users to review the content they are sending.

Assistive technologies are a natural match with crowdsourcing because they provide mediated access to human assistance. Scribe [10], for example, provides deaf and hard of hearing users with real-time captions, and VizWiz [4] allows blind users to get answers to arbitrary visual questions. These systems can have a profound impact on how

users access the world around them. VizWiz has answered over 80,000 questions for thousands of users. However, users of these systems may be unable to effectively avoid capturing sensitive information. For example, a blind user might not be able to tell that they have inadvertently captured a billing statement in an image sent to VizWiz, and a deaf user might not be able to tell that their account information could be overheard in speech until after it has been captioned by Scribe [10].

The identity of crowd workers on platforms like Amazon Mechanical Turk tend to be unknown to the requesters that hire them [13]. This anonymity, coupled with a range of worker skill levels and the need for workers to complete large numbers of tasks to earn a reasonable wage, create the need for quality control systems [3,7]. These approaches increase the overall quality of the work, but at a cost; they tend to increase the number of workers who will see each piece of information contained in a task. Crowd-powered systems that use personal information potentially put users at risk for identity theft, blackmail, and other information-based attacks.

2.2 Crowd-Based Privacy and Security Threats

Concerns with issues related to the privacy and security of sensitive information used in crowd-powered systems have led to some initial work exploring the types of problems that may arise. Harris et al. [6] bring up the idea that ordinary workers might be hired for potentially malicious tasks. Lasecki et al. [12] outline a variety of different individual and group (both coordinated and uncoordinated) attacks that are possible on current platforms, and demonstrate that workers can be hired to do seemingly-malicious tasks (such as copy a credit card number from another task), even if some percentage of workers will abstain from such tasks. Teodoro et al. [16] also found similar hesitation to potentially illicit tasks, such as mailing lost cell phones to a service promising to find their owner and return them. Forums and other worker communities also help discourage this behavior. Our experiments investigate protecting image data of the kind dealt with by VizWiz when answering visual questions for blind users, where information from bank accounts, to names and addresses, to accident revealing images (e.g., accidentally taking a picture) may arise. For much of the valuable information, there is the potential for malicious workers to begin targeting such systems as their popularity grows and these incidences become more frequent [12].

2.3 Approaches to Preserving Privacy

To preserve privacy in crowd systems, Wang et al. [18] studied how to detect malicious workers. However, most approaches have focused on protecting sensitive content. Varshney [17] proposed using visual noise and task separation to preserve the privacy of content in images. This could help protect some types of information, but in many cases information needed to complete the final task (e.g., read a label for a blind user) is lost. Little and Sun [14] looked at protecting privacy in medical records by asking workers to first annotate a blank record to indicate where field values are entered, then using this information to divide a real medical record into pieces that workers could help transcribe without being able to see too much information.

We attempt to counter these threats by ensuring that no worker individually is able to see enough information to do the end user harm. It differs from existing approaches in that: i) the final task being completed does not need to be unknown a priori, ii) it uses a general model in which we do not have an initial template that can be used to advise the division of future tasks, and iii) the division algorithm progressively zooms out while applying partial masks along the way, overcoming many of the context-based challenges encountered in previous work (e.g., information referenced in other pieces). While few systems can prevent coordinated groups from attaining information from tasks, protecting against individual worker threats drastically decreases the threat to end users. To our knowledge, ours is the first work to explore such approaches to general, task-independent privacy preservation using an implemented system.

3 CrowdMask

Our approach protects end users from an important class of privacy/security threats that arises in crowd-powered systems by filtering potentially sensitive content that is unnecessary for the system to operate before it is sent to the crowd. Users define "filters" in natural language for content sent to the crowd. Filters may be defined *a priori* or at sending time, depending on the application and preference. Our approach acts as a filtering step preceding a regular crowd task, instead of having workers try to answer the from a single smaller piece of the image, because the context available in the larger scene is often required to complete their task.

3.1 Dividing Content

The base approach used is to divide content into pieces that each contain incomplete information. This can greatly reduce the risks faced by end users, but is sensitive to the type of risk, granularity, and information available in specific instances. For instance, an image may contain multiple types of PII: a person's face, their name on a nametag, and a partial reflection of them in a mirror. Each of these sources of information can be in a different location, can be a different size, and may be identifiable in different ways.

The problem that we address is that of setting an appropriate granularity for the segmentation. A single segment granularity might allow these pieces of information to be separated from one another, but each still might be contained in a single segment. Setting the granularity of the segmentation higher might result in the person's name tag being filtered out successfully without anyone seeing their full name or job title, but also might result in the person's face being divided into pieces too small to identify that each one is part of a face – resulting in no piece of the face being filtered, meaning the user's face remains unmasked in the final image.

3.2 The Pyramid Workflow for Minimal-Knowledge Filtering

To solve the granularity identification problem, we use a *pyramid workflow* that first presents very small segments of the image and iteratively zooms out in order to identify visual information at different characteristic sizes. Rather than segmenting an image at

a single level of granularity, this multi-level zooming and masking approach segments an image at multiple granularities. Workers are first shown the smallest possible segments, which is least likely to reveal sensitive information. Once all segments of this size have been filtered, workers are shown larger segments with prior masks applied to the images. This process continues until workers have filtered all levels of granularity.

Consider the previous example of an image that contains both a face and a nametag. Applying multi-level filtering allows both the nametags and face to successfully be filtered out without any worker seeing the entire nametag or face. Initially, workers see only small segments of the image, which allows them to filter out the nametags. However, it can be difficult or impossible to tell that such small segments contain faces. Filtering faces requires subsequent, larger segments. These larger segments have the potential to reveal more information, but the nametags will be masked thanks to the filtering that took place at the smaller size. This allows workers to identify all regions that should be filtered without revealing the entire object.

3.3 Optimizing Segment Sizes

As mentioned before, content can vary greatly in size and type of information. To use our multi-level approach effectively, there must be a difference between each level so that workers can gain new context and recognize potential threats that they could not in the previous level, without simply revealing large pieces of the content.

To do this, we optimize the separation in segment sizes between different zoom levels, given a cost bound. We start with the maximum amount that a user is willing to pay to filter each query. Given the user's budget B, and the cost of a crowd task C, we can compute the total number of questions N that we can have answered, $N = B/C$.

Using this bound, we then select a number of levels L to use. The selection of this value is dependent on the type, size, and quality of content that will be used. For instance, high resolution images may require more levels to effectively filter because both small distant objects and large close objects may contain legible information. In preliminary trials we found that 3 levels effectively handled content seen in web images and those taken by most smart phones. Assuming $L = 3$, we can set up a linear system to find the segment sizes for each level that maximizes separation:

$$N = \sum_{i \in 0 \ldots L} N_i = N_1 + N_2 + N_3$$

Where N_i is the number of segments to be created at level i. Now, we want to find the growth factor G between levels. Redefining this linear system as a function of the growth rate gives:

$$N = N_1 + G * N_1 + G * (G * N_1)) = N_1 + G(N_1) + G^2(N_1)$$

We set a minimum division size for N_1 of M (where $M = 2 * 2 = 4$ is the smallest non-trivial segmentation). This is used as the division for the smallest size of the image at the lowest zoom level, while the number of segments along the other dimension of the image is calculated proportionaly to the aspect ratio of the image (e.g., a 2:1 aspect ratio image would end up divided into $2 \times 4 = 8$ segments). This gives:

$$N_1 >= M \text{ so } N_1 + G(N_1) + G^2(N_1) = M(1 + G + G^2)$$

Now we can factor this term to find the solutions for g. Note that while this case can be solved using the quadratic formula, not all selections of L will lead to such clean forms – for instance, $L = 6$ results in a fifth degree polynomial that cannot be factored. In these cases, there are numerical methods can find solutions well within a reasonable margin of error. Finally, we get our growth rate:

$$G = \sqrt{N/M - 3/4} - 1/2$$

While we use this algorithm to generate segments with maximal separation, which is the "optimal" case for our approach, users need only provide a price they are willing to pay prior to sending their source image.

3.4 Predicting Sensitive Content

In addition to directly showing workers images, we can also leverage their understanding of the scene to predict what is in adjacent segments even without showing them. For instance, if we are filtering for PII and workers observe someone's body in one image, then it can be reasonably assumed that their face may be showing directly above it. The worker interface includes a second stage (Figure 1(a)) that asks workers to indicate whether the segments around it are: *(i)* very likely to contain sensitive information, *(ii)* very likely *not* to contain sensitive information, or *(iii)* they are unsure of what would be contained (the default answer).

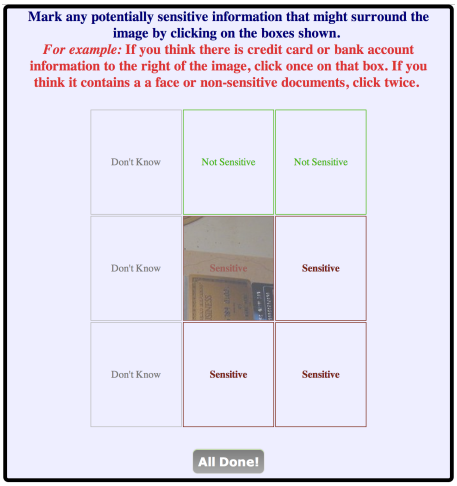
To best use workers' ability to predict content outside of their current view, we issue images in each level of our process in two interleaved "checkerboard" patterns, where each segment shown in the first pass is bordered directly above, below, left, and right by content that has not yet been seen by another worker. After this first pass has been completed, images from the alternate segments are issued to arriving workers. Doing this allows us to withhold asking about segments from the first pass with sufficient agreement between workers on content being or not being present. In the best case, over 50% of the segments from a level can be answered without the need for workers to even view the segment itself (some content can be filtered from the diagonal elements in the first level). While we expect such extreme cases to be rare, there is the possibility for large privacy gains and cost savings.

3.5 System

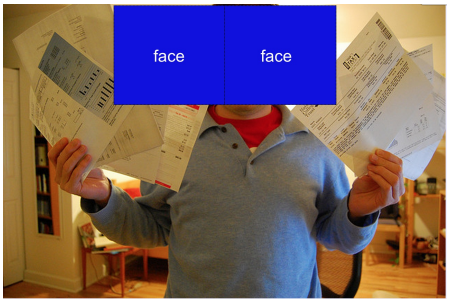
CrowdMask is comprised of three main components: an end user session creation process, a front-end interface for crowd-workers to complete their specified task, and server-side image modification framework.

The end user session creation process allows the user to specify an image to be filtered, the number of granularity levels to use, the maximum amount they would like their session to cost, and the instructions that are shown to workers. Their maximum cost determines what size segments each level of granularity will use.

The worker interface shows an image and a "yes" or "no" question, such as "Does this image contain any sensitive content?" Filtering is done on a per-segment basis to simplify the task and make consensus-finding more tractable, though it reduces precision. Workers then provide a short, textual label to describe what they saw in the image.



(a) The image content prediction UI allows workers to label the likely sensitivity of adjacent segments without seeing them.



(b) Image with personally identifying information (PII) filtered. Here, the bills do not contain PII because there is no name or other identity readable from this distance in the image.

Finally, workers are asked to predict whether there may be any sensitive content that surrounds the image, using a $3x3$ grid that contains the served image and surrounded by empty, clickable boxes that allow workers to mark the predicted content of each box as either "sensitive," "not sensitive," or "don't know." The system serves workers segments starting with the highest zoom level. Importantly, images are segmented and filtered server-side to prevent workers from being able to bypass our restrictions.

3.6 Other Filtering Options

The segmentation approaches we use can also be combined with other filtering approaches. For instance, segments can be blurred before showing them to workers. While blurring alone cannot ensure or attempt to optimize for the trade-off between hiding content and providing sufficient context to decide what to block, it can provide another protective layer in front of a task. Simple blurring might prevent text below a certain size from being read (if it is known that this might be a common case for sensitive information), and even when no additional information is available, it can be used as an initial pass, in much the same way that the predictive step described above provides a pre-filter for some content. In this case, the result would be the first pass in a given level being blurred, and any content that can still be identified through the blurring as potentially harmful is marked without ever showing workers the high-definition version. In cases where nothing can be identified, the de-blurred segment is shown to workers. This process can repeat at each segmentation granularity if desired. Overall, the goal remains the same as in the rest of the system – to avoid showing workers more than the minimum amount of information they need to flag content as potentially sensitive.

3.7 Semantic Labeling

To help workers – both those helping to filter content, as well as those who contribute to the final task – correctly complete their task, they need a certain amount of context. While our masking process prevents much of the context that may create a risk for the requester, this can be partially mitigated by providing a simple description of what content is being hidden from workers. To provide this, we also collect a 1-2 word label from workers after they mark an image as sensitive or not. This label can then be applied to the mask that covers the content in subsequent levels. While this information is only useful to future workers in the case where the worker marks something as sensitive, we always collect a label to prevent this additional effort from biasing workers towards labeling as not sensitive.

4 Preliminary Tests

As an initial exploration of the feasibility of our approach, we explored the level at which people can determine what class of object they are viewing, *e.g.*, an arm, a face, a keyboard, versus the level that they can identify a specific instance of an object, *e.g.* a particular person. We used images of people and objects that had been cropped to the size of the entire image (400px by 400px) so that the relative size of the element in the image would not be a factor. In real images, content will vary in size and may appear as a small piece or far larger than a fixed-size image segment, no matter what the chosen segment size is. The following experiments used Mechanical Turk workers, each paid \$0.14-\$0.16 per task.

Element Recognition To evaluate whether the crowd can recognize elements at increasing levels of granularity, we showed 60 Mechanical Turk workers images of both a face and credit card at three levels of granularity and asked them to identify what type of element they saw using a multiple choice question with five plausible answers. Each level of granularity received responses from 10 unique workers, and no worker could answer for more than one level of granularity of a given image.

When shown half of a credit card, 100% of workers were able to correctly identify it as a credit card. When shown $\frac{1}{5}$ and $\frac{1}{10}$ of a credit card, 80% and 70% of workers correctly identified it, respectively. However, when trying to recognize a face, the worker success rate dropped far more quickly. When showing $\frac{1}{2}$ and $\frac{1}{5}$ of a face, 100% of workers correctly identified that they had seen a face. However, at a granularity level of $\frac{1}{10}$, that number dropped to 40%.

We also conducted a second set of experiments with 60 more Mechanical Turk workers that used the segment's output, and asked workers whether or not the segment they were viewing contained either a face or a credit card, instead of a less clearly-defined query such as "sensitive information." We ran this using the same images as in the multi-level masking experiments, and at the same three levels of granularity. One image contained neither a face nor a credit card number, and the crowd correctly did not mask any of those segments. For the images that contained a face or credit card, the crowd masked them with an average precision of .92 and recall of .91.

Identity Recognition Evaluating the effectiveness of our approach at concealing information, such as personal identity, from crowd workers relies on how much of an effect image segmentation has on workers' ability to recognize and identify a person's face given different sized image segments. To evaluate this, we showed 125 Mechanical Turk workers 16, 6, and 3 segments of a face at three levels of granularity, respectively. 5 workers viewed each segment, and we asked them "Does this image contain a face?" Then, before ending the task, we presented workers with a police-lineup style interface showing six images of faces side-by-side in random order. One of these images was a different picture of the same person they saw in the previous screen, but in a slightly different setting. We avoided using the same image of the person to avoid other pieces of the scene being used to identify the matching image. The other five images in the lineup were all people of the same race, gender, and approximate age, but were identifiable as different people. We asked workers which, if any, of those faces they recognized from the image that they just filtered. We tested five different sets of images, spanning multiple races and genders.

We found that, when simply showing workers a full image of the face without any level of granularity, workers correctly recognized that person around 60% of the time. When segmenting the face into just two segments, the rate of recognition dropped to just 13%. When segmenting the face at a much higher level of granularity - divided into 16 segments, the recognition rate dropped even further to 7.9%.

4.1 Discussion

The key finding from these initial studies is that workers can identify the class of an element in an image even when it is divided in half (100% of workers got the answer correct), whereas only 13% were able to determine the identity of the person in the photo. In naive image segmentation (where a single level is used), the segmentation must divide content by luck to a sufficiently small size. Decreasing the size of segments will increase the chances of dividing arbitrary content into pieces too small to identify some piece of information, but it also increases the chance that some element is divided beyond workers' ability to accurately identify its class. This suggests that we use of a progressive, multi-level approach will allow the system to filter what can be identified with smaller pieces of information, *e.g.*, easier to identify elements and those with a smaller overall size, while not revealing too much, and then progress to a larger image (with previously masked content removed) in order to find content that could not be identified prior to that. Because the threshold for class identification is lower than entity identification, this approach is able to mask before identification is possible.

5 Experimental Setup

Our evaluation focused on images because they comprise a common content type in crowd tasks, and they clearly demonstrate challenges faced when trying to filter out specific information, such as variable size elements (faces, text, etc.). Furthermore, the only atomic units in images are pixels, but these are clearly too high granularity. This is unlike in text, where words or even letters are an easy-to-use minimum granularity.

We evaluate four main issues: *(i)* how varying the natural language question asked and the instructions provided to workers varies their responses and reliability, *(ii)* how well workers are able to identify and mark sensitive content given a query and different size image segmentation, *(iii)* how well workers can predict sensitive content that they cannot directly see in their current image, and *(iv)* how masking images affects the ability of a different group of workers to complete the initial task requested by the user.

5.1 Experimental Design

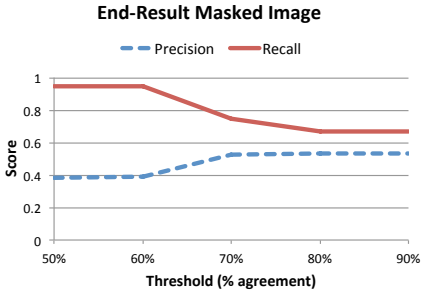
Within image tasks, we were interested in demonstrating that this method is effective even in use cases that not only involved a diverse set of scenarios with potential privacy risks, but also that hold real benefit in solving. As such, we focused on the types of images that blind users take when using VizWiz [4]. Our dataset consisted of 15 creative commons licensed images collected from the web. These images spanned a wide range of different content: from credit card numbers and bills, to faces and to police protests, to potentially embarrassing actions.

We recruited workers from Mechanical Turk and paid $0.14 to $0.16 cents per task, which were estimated to take around 30 seconds each, resulting in a pay rate of $16.80 – $19.20 per hour (far above the typical wage on Mechanical Turk). 3 workers were recruited for each image segment, except where stated otherwise. No worker filters were used, and all completed tasks were accepted (no filtering was done on the data prior to the thresholds we describe for each trial).

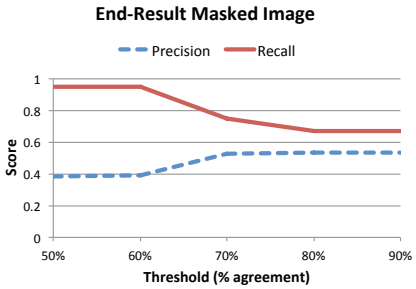
5.2 Exploring Parameters

Effects of Segmentation Granularity To help us determine the optimal granularity levels to use for our other studies, we filtered images with a very high max-cost, meaning that the highest level of granularity consisted of a very large number of small segments and a total of 112 segments across all levels, divided up as 88 small, 20 medium, and 4 large. We found that the crowd could effectively filter some content in phone-sized images at this level. For large objects such as faces, less of the overall face needed to be masked before the face was no longer recognizable. However, some smaller objects were not able to be masked until a later, lower granularity level because the smallest segments did not render those objects recognizable. Ultimately, results were effectively the same as the lower cost option for these type of images. Therefore, this very high level of max-cost was higher than necessary for our test set of phone-sized images.

Asking the Right Questions We tested four variants of instructions that the system presented to workers. Each instruction type was evaluated using the first three of the four images described in the main section of the Experimental Results and using a similar setup, but with one trial for each image + both PII and sensitive information instead of three. These images, segmented into three levels with 25 segments total, were run both for sensitive information and personally identifiable information, separately. Each segment was viewed by three Mechanical Turk workers (1821 total over four question types) paid $0.14 - $0.16 per task.



(c) Average precision and recall the 24 total runs on 4 images. We found that at 50% agreement, the resulting precision is 38.2% and recall is 94.8%. When the agreement is 80%, there is an increase in precision to 52.9% and an increase in recall to 66.5%.



(d) Average precision and recall scores from just the highest level of granularity in the main experiment. Averages are across all tests from all four instruction types.

- **(F1) Filter Question:** Workers are shown only the intended filter definition. For example, "Does this image contain any potentially sensitive information about the requester?"
- **(F2) Filter Question with Example:** The initial question, with examples that give workers an idea of the types of what should be filtered: "For example: If the image contains a face, name, address, or other contact information, click 'yes'."
- **(F3) Filter Question with Example and Non-Example:** The initial question, examples, and non-examples that give workers an idea of what type of should *not* filtered: "If it contains a company name or non-identifying documents, click 'no'."
- **(F4) Filter Question with Non-Examples and End Goal:** The question, examples, non-examples, and the task's goal, that tells workers what information we ultimately need from the image after filtering. The hope is that this information will prevent workers from electing to mask segments that make the goal (e.g., answering "Which of these is my library card?") impossible.

We observed that worker responses varied significantly by each instruction type. F1 was subject to each individual worker's idea of what the question meant. This led to often disparate answers, and generally resulted in both false-positives and false-negatives. This caused high disagreement among worker responses: when looking at the labels provided by workers, the crowd often correctly identified an object that they saw, but disagreed on whether that object needed to be filtered. F2 gave the workers an idea of what the question was looking for, but actually resulted in a larger number of false-positives. In 5 of the 6 runs, the crowd masked more than half of the image, significantly more than our baseline and more than any other instruction type. F3 attempted to rein in the false-positives, but we found that workers often did not listen to our non-example. For instance, the crowd masked a face even when specifically asked not to. Finally, F4 produced images closest to our baseline, with unwanted information masked, but enough information left to answer the end-goal. As a result of this preliminary experiment, all experiments were performed using F4 instructions.

6 Experimental Results

To evaluate how well workers can filter within the multi-level zooming and masking model, we ran an experiment that involved filtering for both sensitive information and personally identifiable information. Four images were tested for each of these conditions. The first image contained an assortment of cards lying on a table, two of which were credit cards. The second contained a man holding pieces of paper, some of which were bills with balances and other information on them (no account information was visible to workers). The third contained a fully addressed and stamped letter. The fourth image contained a police protest/arrest scene in which two faces of protesters were visible. We used three levels of granularity when filtering the images. The highest level was comprised of 16 segments, the middle level 6, and the lowest 3.

To measure our approach's performance, we calculated a baseline to compare results against. To validate the baseline, two coders completed the first image. Inter-rater reliability was calculated using a Cohen's kappa, with a score of .72.

We then completed a full run twenty-four times. Six times for each of the four images. For each image, three trials were run for sensitive information, and the other three for PII. Each image segment was viewed by 3 crowd workers, with a total of 1702 worker responses. Averaging the scores from the 24 total runs, we found that at a threshold of 50% agreement, the end result images had a precision of 38.2% and recall of 94.8%. When raising the agreement threshold to 80%, there is an increase in precision to 52.9% and an increase in recall to 66.5%. The magnitude of this change suggests that there is a high level of agreement between workers. Figure 1(c) shows the full span of precision and recall scores for the final, masked image that is generated. While precision is low (in part because we filter by segment), our high recall means that sensitive information is rarely left unfiltered.

Comparison to Single-Level Segmentation Figure 1(d) shows precision and recall scores when just considering the first (highest) level of granularity. At a threshold of 50% agreement, the single-phase filter had a precision of 46.2% and recall of 62.0%; at a 90%, precision is 54.6% and recall is 36.5%. Compared to the multi-level run, this resulted in a significantly lower recall but slightly higher precision. We also compared the F1 score (which aggregates precision and recall) for all four images of single-level to multi-level segmentation and found that there was a significant 29.8% improvement between single and multi-level, where F1 increased from .438 to .591 ($p < .01$).

6.1 Detecting Embarrassing Content

To evaluate the effectiveness of our approach at understanding highly subjective scenarios, we used three potentially embarrassing images, which contained scenes, such as a person picking their nose. We used three levels of granularity, and asked, "Does this image contain anything embarrassing?" Two images contained one segment with embarrassing content, while the third image did not. We received responses from 239 workers with each segment viewed by 3 workers, and observed that the crowd was able to identify both of the potentially embarrassing situations with a perfect score for both precision and recall, masking only the segments that contained the embarrassing scene.

6.2 End-Goal Answerability

Key to the success of our approach is the ability for workers to still answer a desired question by looking at an image that has been masked. To evaluate this, we used three images that were filtered, from the mutli-level masking experiment. We then presented these masked images to 30 workers, along with the image's associated question: "How many pieces of paper is this person holding?", "Which of these is my loan package checklist?", and, "What is in this picture?". We asked workers whether they believed they were able to answer the question with the masked image they have been shown, and then to simply answer the question. Each question was shown to 10 unique workers. 90% of workers replied that they thought they could answer the question, and 90% of workers answered the question correctly (though the two sets are not a necessarily a union). This demonstrates that content filtering, even using the somewhat course-grained (medium-low budget) divisions we did in our trials, can be done in a way that still allows for the end goal question to be answered.

7 Reducing the Cost of Masking

The cost of running our system is largely dependent on the levels of granularity that the user wishes to filter at, the size of the image, and the level of worker redundancy desired. We priced our tasks at an average of $0.15, and used around 25 segments per image in total, across all levels. While this means each image would cost $3.25 to filter, we specifically did not optimize for price, but instead focused on showing that it is possible to protect user information using a natural language-defined filter. For comparison, reducing the pay rate of these tasks to the U.S. minimum wage ($7.25), would result in images costing $1 each to filter.

7.1 Worker Prediction

To further reduce cost, we wanted to know if workers could effectively predict what is in segments adjacent to the image segment they were shown, without needing to actually show workers the content of those segments (similar to the handwriting completion by prediction in [19]). To evaluate this, we chose a test set of image segments that fit under five categories: *(i)* information partially in scene, *(ii)* information out-of-scene with clear indicator, *(iii)* information out-of-scene with expected content, *(iv)* information out-of-scene with expected non-content, and *(v)* information out-of-scene with unexpected content / no-content baseline. We selected two images in each category (total: 10 images), and got 157 worker responses, with each image shown to 15 workers.

Overall, a progressively selective threshold can be set depending on how cautious (or thrifty) of a policy is desired (Figure 1). To calculate segments that should be removed, we used a slightly more complex aggregation scheme than just setting a threshold: we selected the top-voted element (or multiple elements if there was a tie), as long as the maximum value was above 50%. Doing this gave an average precision across all images of 94.8%, with an average recall of 92.5%.

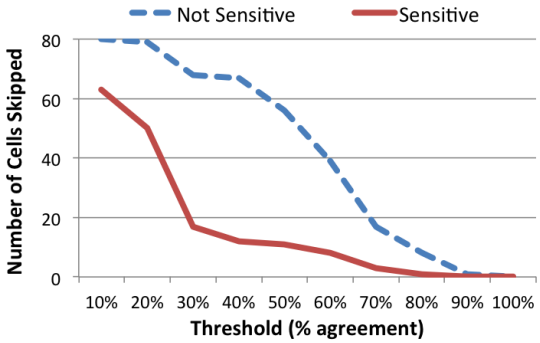


Fig. 1. Plot of how many tasks would be skipped based on worker prediction in our example scenario as agreement rate threshold is increased.

7.2 Reducing Task Size

By using the crowd's predictive capabilities, we can reliably reduce the number of segments that need to be individually marked by workers by 17% while retaining 90% accuracy. This reduces the number of tasks that workers need complete. The number of granularity levels and max-cost can also be optimized for images of certain sizes, quality levels, and content-types. Increasing the number of images shown to workers per task would allow the filter cost per segment to be decreased because workers can more easily complete sets of tasks. To ensure that this does not undermine our core approach, the image segments should be drawn from disjoint pieces of the image, and comprise only a relatively small percentage of the image per worker.

Automated systems can also be used as a 'first pass', filtering content of a type that might present a risk. Computer vision approaches cannot accurately distinguish the context in which certain types of information are sensitive, but it guide the human-filtering process by having workers look only at segments that contains potenitally harmful *types* of information. This can even guide the use of CrowdMask at a higher level. For instance, an image not containing numbers in unlikely to contain harmful account information, so no filtering must be done by human workers.

8 Discussion

Our results demonstrate that we can accurately and reliably filter sensitive content based on natural language definition – even with subjective queries – without revealing the information to workers along the way. We show that this content masking is done in such a way that the initial question behind the image can still be answered.

8.1 Worker Biases

We found that workers had particular biases to marking certain types of content as sensitive, even when it is not. For instance, an image of 10 cards lying on a table resulted over 50% of workers marking all segments that contained any card as sensitive, even though just 2 of the 10 cards were credit cards, and the rest non-sensitive cards, such as

a restaurant loyalty program membership. Similarly, when workers saw pieces of paper that contained numbers, charts, and graphs, some marked that content as sensitive even when they could not know what the text content actually was due to the low resolution of the image. Raising the worker agreement threshold eliminated this problem while still masking sensitive content. Given a more focused question, (e.g., embarrassing content) workers were adept at only filtering segments which contained such content.

8.2 Content Labels

When workers answered a question about an image segment, they also provided a 1-2 word label to describe the segment. We found that these labels were highly informative as to how the worker answered the question. The labels ranged from high-level descriptions to low-level parroting of the content that they saw. For instance, if a worker saw a segment containing what appeared to be a portion of a credit card number, a few would simply label it as "sensitive," most would label it as "credit card" or "credit card numbers", and a couple would provide the actual numbers that they saw.

8.3 Limitations

Our studies showed that our approach is able to successfully filter images in a way that prevents crowd workers from knowing the high-level information contained in the source image. Our analysis was performed using sample images and tasks derived from existing scenarios discussed in the literature. To fully understand how our approach would be used in practice requires recruiting people to create filters for their own potential tasks. However, many of the tasks that would benefit from our approach are not currently being run due to the risk of exposing private information. Running end-to-end experiments on an untested system poses substantial risk for the end user. For this reason we focus on understanding our system and its capabilities. By learning about how workers completed their task and establishing the capabilities of the system itself, we hope to be able to open a new option for crowdsourcing researchers and system builders to provide user protection.

While the our system is designed to thwart individual attacks, the information filtered remains at risk from coordinated group attacks [12]. If a sufficient number of workers colluded and shared images, they would likely be able to recreate the original content in the images. While communication channels are available to workers (e.g., forums) [8], they tend to ally with requesters against coordinated attacks.

We are unaware of existing solutions to thwart coordinated crowd-based attacks that make use of information from multiple workers or sources. Although we focus on instances where attackers learn everything within the task, our work significantly increases the difficulty of effectively attacking a system.

9 Future Work

Based on our results, we have implemented a version of this system that uses JCrop (`deepliquid.com/content/Jcrop.html`) to let workers select the exact regions

they think are sensitive by simply clicking and dragging. These masks can be used to train automated computer vision systems to take over the task of finding a certain type of sensitive information over time. Using the tighter bounds generated by the next version of our approach can make this process more effective by restricting the space even further.

We can also use our pyramid workflow to filter other types of content:

- **Text:** We have also extended our approach to allow text to be filtered using the same core segmentation algorithm, but over one dimension, measuring segment size in number of words. When workers see words or phrases that might constitute sensitive information, we allow them to click the word and it is filtered out in their view. We did not take structural elements of the text (e.g., punctuation) into account, though this could be used in future systems.
- **Audio:** Segmenting audio can be done in a similar manner to text. However, because the word boundaries are not clearly discernible, words may be truncated in a single pass, even with automated segmentation assistance. To solve this, the "check board" method used in our image masking approach can be used to stagger responses and ensure complete coverage. Selecting the length of audio clips has the same trade-offs as images, larger clips provide more context, which leads to more context as well as more risk. Text-to-speech can be used to apply semantic labels even for audio-only tasks.
- **Video:** Much like images, video requires a 2-dimensional filter, as well as temporal division to avoid over-filtering. If audio is present, that can be handled as a separate task (less context, but potentially more secure), or as time-synchronous content that uses the same temporal division as the rest of the video.
- **Other Media:** Other forms of media, including hypermedia and structured content (e.g., databases or knowledge graphs) can also be handled if the appropriate segmentation methods are added. Structured forms of data may have an advantage in terms of privacy protection and cost because there is more a priori knowledge about constraints and where sensitive information is likely to arise.

10 Conclusion

In this paper, we have introduced a task-independent approach for filtering potentially sensitive information using the crowd. While automated approaches can only remove content in a coarse-grain fashion (i.e., based on type), we have shown that it is possible to use human intelligence to filter content using simple natural language queries, without exposing information to the workers themselves. Our novel pyramid workflow progressively "zooms out" from a fine-grained content segmentation (many small pieces) to a more coarse-grained segmentation (fewer large pieces), filtering content as soon as it can be identified as something the user did not intend for inclusion in the final image. We showed that this approach can effectively filter content, while hiding each piece of sensitive information from the constituent workers. Future work aims to make this process faster and cheaper, while maintaining the same "minimal-knowledge" property for workers – preserving as much user privacy as possible, even when a system itself powered by human intelligence.

References

1. Bernstein, M., Teevan, J., Dumais, S. T., Libeling, D., and Horvitz, E. Direct answers for search queries in the long tail. CHI '12 (2012), 237–246.
2. Bernstein, M. S., Brandt, J., Miller, R. C., and Karger, D. R. Crowds in two seconds: Enabling realtime crowd-powered interfaces. UIST '11 (2011), 33–42.
3. Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. Soylent: A word processor with a crowd inside. UIST '10 (2010), 313–322.
4. Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., Miller, R., Tatarowicz, A., White, B., White, S., and Yeh, T. Vizwiz: Nearly real-time answers to visual questions. UIST '10 (2010), 333–342.
5. Chen, K., Kannan, A., Yano, Y., Hellerstein, J. M., and Parikh, T. S. Shreddr: Pipelined paper digitization for low-resource organizations. DEV '12 (2012), 3:1–3:10.
6. Harris, C. G. Dirty deeds done dirt cheap: A darker side to crowdsourcing. In *Privacy, security, risk and trust (passat)*, IEEE (2011), 1314–1317.
7. Ipeirotis, P. G., Provost, F., and Wang, J. Quality management on amazon mechanical turk. In *HCOMP Workshop* (2010), 64–67.
8. Irani, L. C., and Silberman, M. S. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. CHI '13 (2013), 611–620.
9. Kokkalis, N., Khn, T., Pfeiffer, C., Chornyi, D., Bernstein, M. S., and Klemmer, S. R. Emailvalet: Managing email overload through private, accountable crowdsourcing. CSCW '13 (2013).
10. Lasecki, W., Miller, C., Sadilek, A., Abumoussa, A., Borrello, D., Kushalnagar, R., and Bigham, J. Real-time captioning by groups of non-experts. UIST '12 (2012), 23–34.
11. Lasecki, W. S., Murray, K. I., White, S., Miller, R. C., and Bigham, J. P. Real-time crowd control of existing interfaces. UIST '11 (2011), 23–32.
12. Lasecki, W. S., Teevan, J., and Kamar, E. Information extraction and manipulation threats in crowd-powered systems. CSCW '14 (2014), 248–256.
13. Lease, M., Hullman, J., Bigham, J. P., Bernstein, M., Kim, J., Lasecki, W., Bakhshi, S., Mitra, T., and Miller, R. Mechanical turk is not anonymous. *SSRN* (2013).
14. Little, G., and Sun, Y.-a. Human ocr: Insights from a complex human computation process (2011).
15. Noronha, J., Hysen, E., Zhang, H., and Gajos, K. Z. Platemate: Crowdsourcing nutritional analysis from food photographs. UIST '11 (2011), 1–12.
16. Teodoro, R., Ozturk, P., Naaman, M., Mason, W., and Lindqvist, J. The motivations and experiences of the on-demand mobile workforce. CSCW '14 (2014), 236–247.
17. Varshney, L. R. Privacy and reliability in crowdsourcing service delivery. In *SRII 2012*, IEEE (2012), 55–60.
18. Wang, T., Wang, G., Li, X., Zheng, H., and Zhao, B. Y. Characterizing and detecting malicious crowdsourcing. In *SIGCOMM* (2013), 537–538.
19. Zhang, H., Lai, J. K., and Bcher, M. Hallucination: A mixed-initiative approach for efficient document reconstruction, 2012.