Externalizing Worker Qualifications

Background

Labor market platforms ranging industries and styles of work (e.g. Uber, TaskRabbit, UpWork, and Amazon Mechanical Turk) have struggled for years with persistent (generally growing) challenges relating to worker qualifications. These issues range a wide spectrum: in some cases, worker qualifications are non–transferable, leading requesters to "re–invent the wheel" as they attempt to determine in their own way whether a potential worker is qualified and reliable.

Problems determining a worker's qualifications start on day 1; labor markets begin their relationships with new workers almost entirely uncertain about the worker's competence in any type of task. Gathering this information through qualification exams is generally time–consuming and costly.

Challenges mount as workers' skill sets develop; work requiring more training and skill (for example, translating or programming) are either verified by individual *requesters* (e.g. Amazon Mechanical Turk) or are verified by the *platform* itself (e.g. UpWork). While the *UpWork* model avoids needless repetitive work by generally consolidating qualifying exams at the platform level, these labor platforms nevertheless find themselves in the unenviable (and often unexpected) position of having to develop new qualifications exams to outpace would–be cheaters.

A (Potential) Solution

In offline labor markets involving skilled workers, credentials are sometimes managed by external, trusted organizations: the state requires electricians to serve in apprenticeships and pass licensing exams; lawyers take exams administered by the American Bar Association (ABA); doctors take similar exams given by their own oversight organization (the NBME).

This approach may prove useful in alleviating the burden online labor markets are increasingly taking on. By externalizing worker qualifications, a number of benefits can emerge:

1. New workers can provide some evidence of their work history to labor platforms, mitigating or even resolving the "cold start" problem of not knowing the trustworthiness or competence of a new user.
2. Aggregated (even heterogeneous) ratings sourced from differing labor markets can provide a more holistic picture of a worker's areas of competence. Workers can use this information to identify specialization more easily, reducing wasted time *searching for* or *working on* suboptimal tasks. Requesters can benefit from this information by making more informed decisions about which worker to contract.
3. By turning the reputation management of workers into an external entity (agnostic to the labor platform), designers and developers can focus more fully on the marketplace itself.

Measurable Variables

We see two measurable variables that might be relevant to the goal of better, or higher–quality, crowd (and other on–demand) work. The first, accuracy of predicting worker acceptance rates, is fairly straightforward. The second, determining whether workers better self–evaluate in their selection of work based on more comprehensive awareness of their reputation (or credibility). We'll call this latter dimension "worker

self–specialization" to allude to the emergent property of workers identifying their own niche expertises and selecting for tasks in those domains.

Worker Accuracy

Determining a worker's likelihood of completing a task correctly is fundamentally the goal of using past worker approval rates (like on AMT, Upwork, and other platforms) as a low pass filter for workers. The thinking goes that past performance is an indicator towards future performance.

The baseline approaches requesters have used typically falls into two paths: 1. Relying principally on the platform's rating system as a broad–spectrum filter, and 2. Implementing one's own rating or qualification system as a determiner of whether or not to allow someone to do tasks of similar nature.

We can conduct post–facto analyses of workers' data (aggregated from their various work platforms) to see if any combination and/or weighting of their worker profiles from multiple platforms more accurately predicts their likelihood of doing good work than the "baseline" (that is, informed solely by the platform on which the work was solicited).

To run this analysis, we would need to solicit new information from workers than they've traditionally shared. Specifically, workers would need to be willing to share approval rates from other work platforms (such as AMT or CrowdFlower).

Worker Self–Specialization

The next dimension that we could evaluate would be to use this data to better–inform workers about their approval trends. The best way to run this experiment would be to randomly assign some number of workers to a control group and some others to an experimental group, wherein these workers have the opportunity (after sharing their aggregated profiles with us) to see at a higher level on what *types* of work they typically fare better.

The operationalization of this dimension would almost certainly be the measurement of overall approval rates over time after intervention. To make this more concrete, we would provide workers with guidance on what sorts of work they typically get approved, perhaps helping to guide them to those kinds of work more actively than otherwise.

There are problems with the experimental design described at the outset. For one thing, finding willing participants might be difficult, and randomly sampling for potential participants would make recruitment of a sufficiently large cohort difficult. Moreover, the self–selection of workers into this group (or rather, out) would almost certainly compromise the "randomness" of the assignment approach.

Instead, it may be better simply to allow a limited number of interested participants to sign up to participate in a "pilot program", incorporating their work data from other platforms. While it's true that this would otherwise be "giving up" on randomized experimentation, the question of what types of workers would prefer this sort of work is its own important area of discussion that we shouldn't try to evade.

First Steps

The first step is to solicit workers (either on AMT, Upwork, or another platform) who may be interested in participating in a study of this nature. We'll ask these participants to self–report their own approval rates from each platform, perhaps asking for an annotated history of their tasks (if feasible).

Two methods of evaluation may be viable, depending on the homogeneity of the tasks. One approach would be to train a machine learning model on a subset of the aggregated data and test it against other subset of that data.

The other approach would be to investigate the nature of the tasks that are being done, categorize them by similarity, and focus on trying to predict worker quality from the aggregated approval rates across platforms (for instance, looking at the worker's approval rates for translations across AMT and Upwork).

A later step is to build a prototypical server serving the basic needs of an aggregated reputation service. That is, if a client queries the service for information about a worker on Upwork, the client will receive a reasonably comprehensive snapshot of the worker's history on other platforms.