What's the Right Price? Pricing Tasks for Finishing on Time

Siamak Faridani
University of California, Berkeley
faridani@berkeley.edu

Björn Hartmann
University of California, Berkeley
bjoern@eecs.berkeley.edu

Panagiotis G. Ipeirotis
New York University
panos@stern.nyu.edu

Abstract

Many practitioners currently use rules of thumb to price tasks on online labor markets. Incorrect pricing leads to task starvation or inefficient use of capital. Formal pricing policies can address these challenges. In this paper we argue that a pricing policy can be based on the trade-off between price and desired completion time. We show how this duality can lead to a better pricing policy for tasks in online labor markets. This paper makes three contributions. First, we devise an algorithm for job pricing using a survival analysis model. We then show that worker arrivals can be modeled as a *non-homogeneous Poisson Process (NHPP)*. Finally using NHPP for worker arrivals and discrete choice models we present an abstract mathematical model that captures the dynamics of the market when full market information is presented to the task requester. This model can be used to predict completion times and pricing policies for both public and private crowds.

Introduction

One of the most important challenges for task requesters on crowdsourcing markets like *Amazon Mechanical Turk (AMT)* is to properly price and schedule their tasks (or "HITs," which stands for "Human Intelligence Tasks"). Improper pricing or scheduling often results in task starvation and loss of capital on these markets. For example it is believed that workers have an expected hourly wage in mind and they tend to not accept *underpriced* tasks that need more time per unit reward than what they have in mind. Tasks that are not accepted stay in the system (they are often called *"starved HITs"*). Starved HITs may be canceled or reposted by the requester resulting in expenditure of more time and money than planned for the task. Overpriced tasks are also undesirable since requesters can invest excess capital in quality assurance for the data that they have collected. By using a survival analysis model we devise an algorithm for determining the optimal reward for a crowdsourced task. Even though we focus just on reward setting on this paper, our approach is generalizable and practitioners can use it to optimize the other task characteristics, such as task length, bonus, and even keyword selection for their tasks.

Survival analysis can yield the expected task completion time and optimal reward for a candidate task by using a model that is trained on the historical data of the market. However, survival analysis provides no insight into how the market works, how workers arrive to the system and how they decide to perform a task. In the second half of the paper we focus on the worker/task dynamics that characterize individual workers. For this second section, we assume that requesters are exposed to complete information about the market — they can access snapshots of the tasks posted on the market and get information about task completion by individual workers. Private crowds are examples of these type of markets where requesters often have access to historical data about arrivals and task choices for the workers. By looking at quantitative data from Mechanical Turk we show that worker arrivals can be modeled with a non-homogeneous Poisson Process (NHPP).

Building a proper model for worker behavior also requires a descriptive model of how workers decide to take on and finish a task. Workers often select their tasks from a desirable task pool. Our observation shows that workers often have preferences for the types of tasks they like to accept. We use this concept to develop a *discrete choice based model* for a better pricing policy and scheduling for crowdsourced tasks. In cases where complete, or even partial, information of the market is available, a requester can optimize her task attributes to increase the likelihood of workers accepting the task. Discrete choice models can provide a framework to optimize the attributes of a task and therefore increase its desirability to the user. One convenient aspect of discrete choice models is that this change in desirability can be captured, quantified and used for attribute optimization.

Terminology and Definitions

Before continuing we define some of the terminology used in this paper. We define ***workers*** as individuals who accept tasks on a crowdsourcing market. A crowdsourcing ***market*** is the place, usually an online website, where workers find and perform ***tasks*** often for a financial ***reward***. In the literature, workers are occasionally called ***Turkers***, a description of workers who perform tasks on ***Amazon***

Mechanical Turk (AMT). Crowdsourced tasks are posted to the market by individuals or companies. In this paper, the entity that posts tasks to the market is called a ***requester***. A task may be composed of atomic subtasks, ***HITs (Human Intelligence Tasks)***. HITs are completed by workers. Some of the HITs are never picked by workers and they stay on the market until canceled by the requester or the market administrator. We call these ***starved HITs***. The ***optimal reward*** is the minimum amount of money that the requester can pay for each HIT and still have the task ***completed*** by his desired completion time.

Data Set

We have been monitoring the AMT marketplace and taking snapshots of the market since January 2009. For a description of the process and of the dataset, please see (Ipeirotis 2010). For the purpose of this paper, we used a smaller dataset, containing 126,241 HIT groups from 7,651 requesters. Our dataset contains 4,113,951 individual HITs that are worth $344,260. We use *Latent Dirichlet Allocation (LDA)* and requesters' selected keywords to capture the type of the work (Blei, Ng, and Jordan 2003). The reputation of the requester is accounted for by using their historical number of posted HITs, amount of rewards that they have spent in the market, the number of different HIT groups that they have used and the first time that the requester has posted to the market. Market condition is also captured by counting the number of competing HIT groups and competing rewards that were available when each HIT was posted.

A Brief Introduction to Survival Analysis

Survival analysis, frequently used in epidemiology and biostatistics, is a general term for statistical techniques to determine the time until a particular ***event*** occurs. Time can be represented in any units (hours, minutes or years). What constitutes an ***event*** depends on context. For instance, in epidemiology an event usually refers to the *death* of the individual. In the context of maintenance scheduling, an *event* can be referring to a machine breakdown. A ***survival function***, $S(t)$, is the probability that the survival time is longer than t. The survival function $S(t)$ is often defined through a hazard function $h(t) = -\frac{S'(t)}{S(t)}$, with $S'(t)$ being the first derivative of $S(t)$. The hazard function captures the rate of death at time t, across the population that survived until that point. A Cox proportional hazard (CoxPH) model is a ***semi-parametric*** model in which the hazard function for an individual with predictors $\mathbf{X}$ is defined as:

$$\log(h(t, \mathbf{X})) = \log(h_0(t)) + \sum_i \alpha_i \cdot X_i \tag{1}$$

where $h_0(t)$ is the "baseline hazard function" and can have an arbitrary form.

In the basic form of CoxPH, the predictor variables are assumed to be *time-independent*. Extensions of the Cox model use time-dependent predictors (Kleinbaum and Klein 2005). In our work, we used the CoxPH implementation available in R that considers time-dependent variables and multiple events per subjects, by using the counting process formulation introduced by Andersen and Gill (Andersen and Gill 1982).

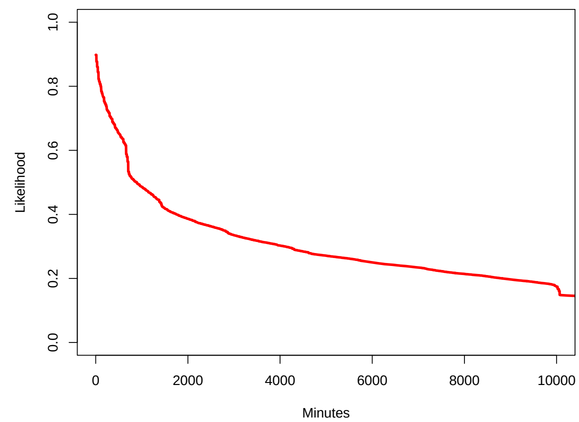

Figure 1: Survival curve for a crowdsourcing task with reward = $0.25, with 30 HITS, from a requester that has posted $1,341 worth of tasks and 7100 total tasks. The task was posted on a Monday with 917 other competing tasks on the market.

Figure 1 shows the survival curve for a task with $0.25 reward, derived by fitting a Cox proportional hazard model to the data. The x-axis represents the age for the task, while the y-axis shows the probability that the task is still "alive." The details and a complete evaluation of the CoxPH model for predicting the completion time for crowdsourced tasks are in (Wang, Faridani, and Ipeirotis 2011) and we refer interested readers to that paper for a detailed discussion. What we will use in this paper is the fact that we have a functional form to connect completion time with assigned reward, controlling for factors such as the history of the requester, the type of the task, and the competition in the market. In the next section, by using a Cox proportional hazard model, we present a pricing policy for posting tasks to online labor markets.

Pricing Policy Based on Survival Analysis

In Figure 2, we vary the reward from 1 cent to 1 dollar and calculate the expected completion time for the task described in Figure 1. A continues curve is fitted to data points for better visualization. As we see, the graph is monotonically decreasing, for increasing values of the reward. This behavior, in conjunction with the desired completion time, can be used to develop an procedure to determine the price for a task, in order to finish right before the desired completion time. Algorithm 1 shows an optimization algorithm for finding the price for a crowdsourcing task.

Algorithm 1 uses a bisection search on the results of a Cox proportional hazard model to find the appropriate reward for the desired completion time. The appropriate reward is defined as the minimum value of the reward

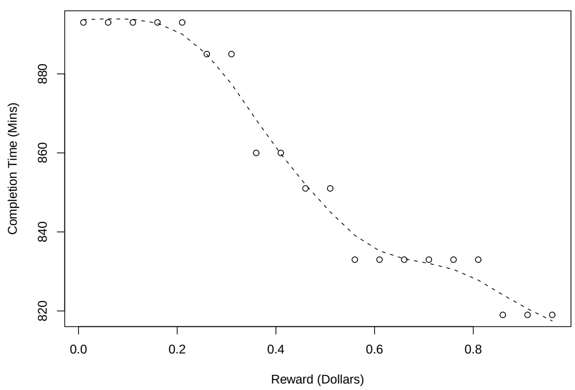

Figure 2: Expected completion times to the task presented in Figure 1 when the reward varies from \$0.01 to \$1.00. The curve is monotonically decreasing.

that ensures the completion by the desired completion time t_{max}. In this example we have only used the most important attribute (reward) but this approach can be easily extended to a multivariate optimization model that includes other attributes of the task like number of HITs and even keywords.

Towards a Better Theory Model for Market Behavior

In the previous section, by using a CoxPH model, we provided an algorithm for a pricing policy based on the attributes of the task. The procedure uses the trade-off between the reward and desired completion time to come up with the lowest reward that ensures a proper completion for the task. CoxPH is used as the module that provides the values of completion times to the algorithm. In this section, we further study the market dynamics and provide a model that can eventually replace the CoxPH model in our algorithm. We first focus on worker arrivals and show that worker arrivals to a crowdsourcing labor market follow a *Non-Homogeneous Poisson Process (NHPP)*. We then show that the likelihood of the task being selected by a worker can be maximized by using a discrete choice model such as the multinomial logit model (MNL). In order to find the completion times of a certain task, we can simulate the NHPP arrivals and simulate their choice behavior based on the MNL model. Providing a closed form equation for completion times is out of the scope of this paper but is a topic of interest for future research.

Stochastic Arrival Model

Figure 3 shows the amount of activity for workers based on the day of the week. The result indicates different levels of activity, and suggests that the assumption of time homogeneity is not justified. To alleviate the assumption of homogeneity, we consider a *non-homogeneous Poisson Process (NHPP)* as the arrival model. This means that workers arrive at the labor market according to a Poisson model with a varying rate $\lambda(t)$. Unlike the Poisson model,

Algorithm 1: Algorithm for calculating reward for the desired completion time.
R stands for reward and **CT** stands for completion time

Input: Dataset A that contains historical information about different tasks, their posting date, reward, requester,...
Attributes of the new task h
Desired completion time t_{max}
Maximum payable reward R_{max}
Precision value ϵ
Output: Reward amount
begin
$R_{min} \leftarrow 0$
$R_{max} \leftarrow R_{max}$
$R_{mid} \leftarrow (R_{min} + R_{max})/2$
$CT_{R_{min}} \leftarrow SurvFit(A, h_{R_{min}})$
$CT_{R_{max}} \leftarrow SurvFit(A, h_{R_{max}})$
$CT_{R_{mid}} \leftarrow SurvFit(A, h_{R_{mid}})$
while $|CT_{R_{min}} - CT_{R_{max}}| > \epsilon$ **do**
if $CT_{R_{mid}} \geq t_{max}$ **then**
$R_{max} \leftarrow R_{mid}$
$R_{mid} \leftarrow (R_{min} + R_{max})/2$
if $CT_{R_{mid}} < t_{max}$ **then**
$R_{min} \leftarrow R_{mid}$
$R_{mid} \leftarrow (R_{min} + R_{max})/2$
$CT_{R_{min}} \leftarrow SurvFit(A, h_{R_{min}})$
$CT_{R_{max}} \leftarrow SurvFit(A, h_{R_{max}})$
$CT_{R_{mid}} \leftarrow SurvFit(A, h_{R_{mid}})$
return R_{mid}

/* Function *SurvFit(A, h)* */
Input: Dataset A and new task h as defined above;
Output: Expected completion time
begin
Completion Time = Find the completion time for h by fitting CoxPH to A (i.e., by using survfit(coxph(Surv(A),h)) in R language)
return Completion Time

in a *NHPP* arrivals of two workers are not independent and they both depend on a latent variable, *time* t.

Traditionally used for counting data, *Poisson regression* is a subclass of *generalized linear* models where we fit a distribution from the exponential family to experimental data. Generalized Linear Models were introduced as a regression tool for the random variable of the exponential family of distributions (Nelder and Wedderburn 1972). This family includes the normal, Poisson, Gamma, inverse Gaussian and binomial distributions. Many statistical methods are a subclass of a generalized linear model. To formulate this problem we first use classical stochastic process arguments to show that worker arrivals to a labor market can be modeled as NHPP arrivals.

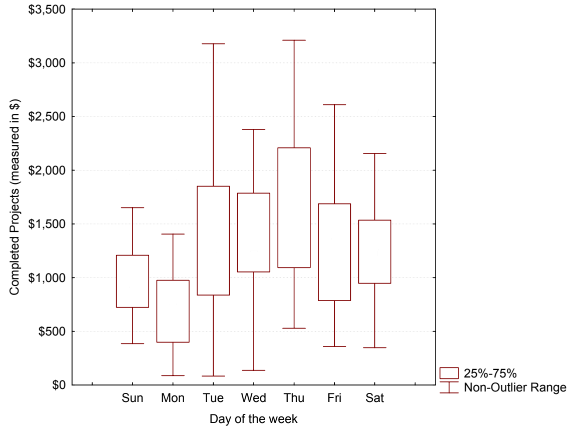

Figure 3: The amount of activity over different days varies, highlighting the fact that a homogeneous arrival model for workers is inappropriate.

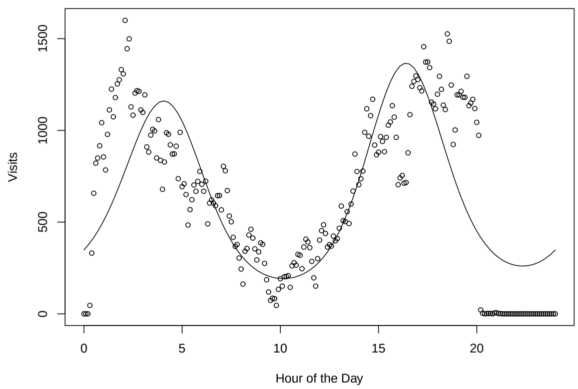

Figure 4: Fitting a NHPP to a visitor data retrieved from one of our tasks on Mechanical Turk. $\lambda(i)$ is found by using GLM.

Worker arrivals to the labor market are NHPP

A Poisson distribution is used for modeling counting processes. We show that worker arrivals to an online labor market follow a NHPP process. Using empirical data, (Gunduz and Ozsu 2003) showed that the number of visits to a web page can be modeled with a Poisson model. Faridani et. al. (2009) study their private crowd of volunteers and demonstrate an NHPP model for their worker arrivals.

Poisson Regression Poisson regression is a subcategory of the generalized linear model in which a non-homogeneous Poisson distribution with a varying rate $\lambda(t)$ is fit to the data. The goal of regression is to find the value of the function $\lambda(t)$ for different values of t.

In Figure 4 we have used a Poisson regression to fit a regression line to more than 131,000 worker arrivals to Amazon Mechanical Turk. This number of workers only represents a portion of the workers on the market but we can argue that this is a *thinned* Poisson process. As a result the original arrival of workers to the market is also a Poisson model (the superposition of Poisson processes is also Poisson).

Choice Based Crowd Dynamics

Discrete choice models have been used extensively in economics and revenue management (Train 2003; Vulcano, van Ryzin, and Ratliff 2008; Vulcano, van Ryzin, and Chaar 2010). A discrete choice model assumes that the worker has a choice of tasks to work on, and chooses the task that optimizes its own "utility," whatever the definition of utility may be for each worker (e.g., it may be the hourly wage). The model adopts utility as a modeling concept, and explicitly assumes that utility is latent and we only observe the actions that are the result of optimizing this utility. In our case, the different choices for the workers are the different tasks on the labor market, with different attributes. Attributes include the reward for the task, time of the day that the task is posted, number of total HITs in that task, and other properties. These attributes make a task more desirable or less desirable for a worker. Of course, the workers' decision to accept a task is also influenced by the other tasks that are available on the market at that moment. For example, a worker may decide to not accept a transcription task for $1 if a survey task with a $1 reward is available on the market. However, the same $1 transcription task may be desirable if the only tasks available in the market are other, comparable transcriptions tasks worth 50 cents are available. We may also have the case that a worker may decide not to accept any tasks and leave the market without completing anything.

This *dependent behavior* can be modeled with discrete choice models. One aspect of such models is that the likelihood of accepting a task can be updated as the attributes of available tasks on the market change. We assume that workers are utility maximizers and in order to capture the preference behavior of workers we use a *Logit Model*. In this paper we assume that the crowd is homogeneous in terms of task preferences. An extension of this model, explicitly modeling the fact that there are groups of workers with different preferences and skills is the BLP model (Li, Ghose, and Ipeirotis 2011). While we do not cover BLP-style models in this paper, it is definitely a direction for interesting future research (Berry, Levinsohn, and Pakes 1995).

Choice Based Model for Task Selection for Crowd Workers

In the previous section we showed that workers arrive to the system according to a *NHPP* process. The question that we answer in this section is "How do workers select a task to work on." In our framework, as described above, we assume that workers are ***utility maximizers*** and work toward maximizing the value of their own utility. One of the advantages of this viewpoint is that it does not require information about individual decisions of workers (such information is not observable on platforms like AMT) but relies on just observing the aggregate output to infer the aggregate preferences of the workers. Aggregated market data can be used to estimate the weights of individual attributes in the aggregated utility. To analyze the choice model for workers we define two utility values:

- *Utility of Task*: The utility that the worker will gain by selecting and finishing a task, which it typically the utility of money earned, but also can include aspects such as learning something, having fun, etc.
- *Utility of Time*: The utility that the worker will lose by spending time on a task (and thus not being able to accept and finish other tasks).

Intuitively, a worker works on a task only if the *utility of task* is larger than the *utility of time*. The assumption of a rational worker implies that workers select the task that maximizes the difference between these two values. The worker will also pick the tasks that maximizes this difference.

Assume that n tasks are available on the market and denote j as the index of task H_j. If we assume that a worker has T units of time at hand to spend on working on tasks, then we denote the utility of task for task j as $U_h(X_j)$ and the utility of time as $U_t(t_j)$ in which t_j is the amount of time that takes for the worker to finish task j. The *rational worker* assumption implies that workers maximize the value of $U_h(X_j) - U_t(t_j)$. In this formulation X_j is a multidimensional attribute vector for the task ($X = \langle x^1, ..., x^k \rangle$). For our analysis, we consider $U_h(X_j)$ to be a linear combination of weighted utilities from observable task attributes, plus an unobservable stochastic term ξ to capture the utility of the unobserved characteristics of the task and is typically assumed to be independently and identically distributed according to a Gumbel distribution. In this case, the utility of tasks are formulated as:

$$U_h(X) = \sum_{k=1}^{K} \beta^k x^k + \xi \tag{2}$$

Note that β is positive for desirable attributes (e.g., number of HITs) and is negative for undesirable attributes (e.g., required time to finish the task). To make the formulation simpler we assume that the utility of time is a term with negative β in the utility of task (Equation 2). The main challenge for this formulation is to estimate the value of parameters β from recorded market information.

Homogeneous Workers Model (Logit Model)

We assume that the crowd is *homogeneous* meaning that the values of β are the same for all workers (Train 2003; McFadden 1972). Workers arrive to the labor market according to a non-homogeneous Poisson Process and each of the workers selects a task from available tasks with certain probability that is determined by a logit model. For example the worker w is then presented with a set C_w of tasks to work on. The worker can also decide not to accept any task, an option that we denote with C_0. In our setting, the probability that worker i decides to work on task j is:

$$P(choice^i_j) = \frac{e^{\beta \mathbf{X^i_j}}}{\sum_{j \in C_w} e^{\beta \mathbf{X^i_j}} + 1} \tag{3}$$

In Eq 3 the number one in the denominator is due to the zero utility for C_0 cases when worker decides to not accept any tasks ($e^{\beta \cdot 0} = 1$). For homogeneous workers Equation 3 is equal to the market share of the task j. McFadden shows that β values can be found by using a logistic regression (Li, Ghose, and Ipeirotis 2011).

Results

Figure 5 shows our preliminary results for the model. As expected, as we increase the number of HITs for a task it becomes more likely that it will be picked up by workers, resulting in increased demand for the product. Increasing the number of competing projects decreases the demand for the task. A potentially surprising outcome is that increasing the reward decreases the demand for the task. While this appears counter-intuitive, it is the direct result of the choice model for the market. High reward tasks usually mean more complex and more involved tasks and that decreases the utility of high reward tasks for the worker. Effectively, we observe the Simpson's paradox in our analysis. In the future, we are planning tease out these confounding factors, by incorporating a topic model that will capture the inherently difficulty of each available task.

Our choice model can now be used to price tasks. Instead of changing prices for survival time, we can change prices to adjust demand for the task. By simulating NHPP arrivals and simulating demand for the task over time, we can achieve the same effect and price the task for being completed on time.

Conclusion and Future Work

Heavy tail distributions of completion times cause traditional machine learning algorithms in software packages like Weka to fail in predicting the numerical value of completion times for crowdsourced tasks. This heavy tail distribution is detailed in (Barabasi 2005) and also studied for AMT in (Wang, Faridani, and Ipeirotis 2011). Cox proportional hazard regression models and survival curves are typically used to model these heavy tail behaviors. There is a nonlinear relationship between the value of completion time and the predictors that are used to train the model. We show that this value is monotonically decreasing for increasing reward values. This property is used to design an algorithm for finding the reward for a candidate task. This reward ensures the minimum payment for the task with a desired completion time.

Using the empirical data from Mechanical Turk and examples from private crowds like CONE (Faridani et al. 2009) we show that arrivals follow a NHPP model. This enables us to simulate arrival of the workers to the market. We then use discrete choice models and multinomial logit model in particular to show how a requester can optimize her task by increasing the likelihood of the task being picked by workers.

We are interested in exploring the discrete choice model further and extending it to a closed form formulation that combines both the arrival model and the logit model to estimate the completion times.

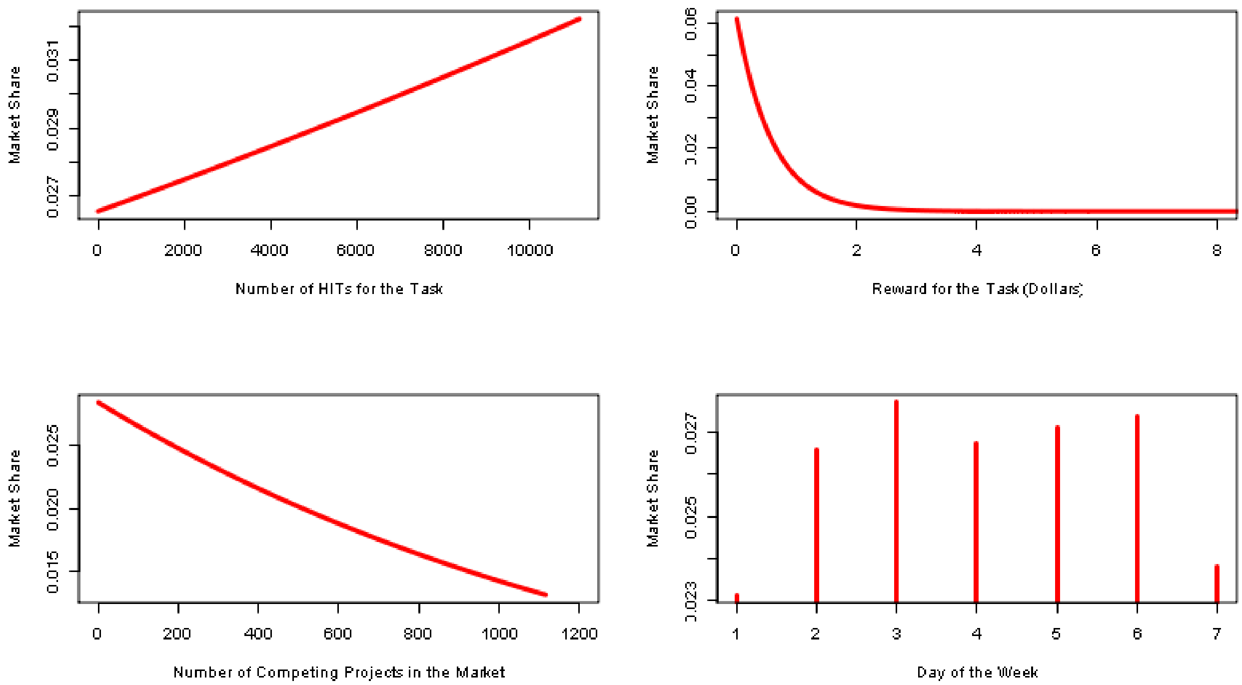

Figure 5: Training a logistic regression model on the market data. Plots show predictions for a typical task with a 50 cents reward that contains 100 HITs and is posted on a Monday on 9AM where there were 100 other competing projects on the market. Graphs are results of experiments where we have varied each predictor and predicted the likelihood

References

Andersen, P., and Gill, R. 1982. Cox's regression model for counting processes: a large sample study. *The annals of statistics* 10(4):1100–1120.

Barabasi, A. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435(7039):207–211.

Berry, S.; Levinsohn, J.; and Pakes, A. 1995. Automobile prices in market equilibrium. *Econometrica* 63(4):841–890.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

Faridani, S.; Lee, B.; Glasscock, S.; Rappole, J.; Song, D.; and Goldberg, K. 2009. A networked telerobotic observatory for collaborative remote observation of avian activity and range change. In *Proceedings of the 2009 IFAC Workshop on Networked Robotics (Moore KL, Ed.). International Federation of Automatic Control. Elsevier, Oxford, United Kingdom*. Citeseer.

Gunduz, S., and Ozsu, M. 2003. A poisson model for user accesses to web pages. *Lecture Notes in Computer Science* 332–339.

Ipeirotis, P. G. 2010. Analyzing the Amazon Mechanical Turk marketplace. *XRDS* 17:16–21.

Kleinbaum, D., and Klein, M. 2005. *Survival analysis: a self-learning text*. Springer Verlag.

Li, B.; Ghose, A.; and Ipeirotis, P. G. 2011. Towards a theory model for product search. In *Proceedings of the 20th international conference on World wide web*, WWW '11, 327–336. New York, NY, USA: ACM.

McFadden, D. 1972. *Conditional logic analysis of qualitative choice behavior*. Institute of Urban & Regional Development, University of California.

Nelder, J., and Wedderburn, R. 1972. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)* 370–384.

Train, K. 2003. *Discrete choice methods with simulation*. Cambridge Univ Press.

Vulcano, G.; van Ryzin, G.; and Chaar, W. 2010. OM Practice—Choice-Based Revenue Management: An Empirical Study of Estimation and Optimization. *Manufacturing & Service Operations Management* 12(3):371–392.

Vulcano, G.; van Ryzin, G.; and Ratliff, R. 2008. Estimating primary demand for substitutable products from sales transaction data. Technical report, Working paper.

Wang, J.; Faridani, S.; and Ipeirotis, P. 2011. Estimating the Completion Time of Crowdsourced Tasks Using Survival Analysis Models. *Crowdsourcing for Search and Data Mining (CSDM 2011)* 31.