Information in the Job Application Graph*

John J. Horton
Leonard N. Stern School of Business
New York University

October 3, 2015

Abstract

When a worker chooses to apply—or not to apply—to a particular job opening, she implicitly reveals information about her estimated productivity if hired for that job. These application decisions, taken collectively, can be used to calculate measures of worker similarity, analogous to those used in recommender systems for products. I show that these similarity measures have an economic interpretation when the worker application decision is modeled as conventional utility-maximizing discrete choice problem—namely that the similarity between two workers is a monotonic transformation of the posterior probability that two workers are of the same "type." I also provide a way to estimate parameters of the discrete choice model using the generalized method of moments. Finally, I validate the usefulness of the worker similarity approach by reporting results from a field experiment in which employers were recommended additional workers based on similarity measures. This treatment substantially increased recruiting and was more effective for relatively harder-to-fill jobs that required a more diverse set of skills.

1 Introduction

Compared to product markets, labor markets generally have high search costs. Each "unit of supply"—the individual worker—is a unique bundle of productivity-relevant attributes. Learning about these attributes is costly to employers; workers have to learn about how different employers value different attributes. Despite the importance of search costs in labor markets, most of the innovation around reducing search costs has occured in product markets; aggregated consumer reviews/ratings and recommender systems (Adomavicius and Tuzhilin, 2005) have become commonplace in electronic product markets and a substantial literature documents there usefulness and importance.

One explanation for the product market/labor market innovation gap is simply a matter of data. The rise of electronic commerce for products gave platforms detailed micro-data on transactions—collected essentially without error and with minimal cost. With this data, recommendations become feasible. A similar development has not happened to nearly the same extent for labor markets. However, with the rise of LinkedIn, Glassdoor, more modern job listing sites, online labor markets such as Upwork and computer-mediated marketplaces like Uber and Lyft, much more of the labor market is becoming computer-mediated in an analogous way.

A particularly powerful approach to recommendations is to use purchase data to define item-based notions of similarity that leverage the "purchase graph" (Linden et al., 2003) : "Customers who bought A also bought B." A great advantage of this approach—compared to say a hedonic approach that tries to measure customer valuation of various product features—is that the platform does not have to "know" what it is recommending—it is simply surfacing statistical relationships. This "know nothing" approach is particularly useful when the number of goods is very large and the platform is relatively ignorant compared to the market participants. The analogous approach in a labor scenario would be an employer with a job posting considering a single worker and then the platform recommending similar workers to consider. How would a similar statistical, graph-based approach be applied to a labor market?

*Thanks to Panagiotis Adamopoulos for helpful comments and suggestions.

An immediate problem with the "purchase graph" approach when applied to labor markets is that each "item" is generally hired for a single job opening, making it typically impossible to make statements like "Employers who hired A also hired B." However, hiring is typically the ultimate step of a process where "edges" are not nearly so sparse—even if the firm hires no one, they could easily consider dozens or even hundreds of applicants. This application graph is generally not observed, but when it is, it becomes possible to use graph-based measures of worker similarity— "Employers considering worker A also considered worker B."

In this paper, I explore using this job application graph for recommendations both theoretically and empirically, through a field experiment in an online labor market. I develop a simple model of the worker's job application decision. In the model, workers and jobs have different "types" that are unobserved by the researcher but affect which job openings a worker applies to, with workers being more likely to apply to jobs that match their type. I then show a commonly used measure of statistical similarity (e.g., the number of times two workers make the same application decision divided by the total number of decisions) is a monotonic transformation of the Bayesian posterior probability that two workers are of the same "type." As such, recommendations ordered by these graph-based measures of similarity will be consistent with our best estimate of similar workers. Under some assumptions about the market distribution of types, the model can be used to estimate a number of primitives, such as the size of match-specific productivity gains, which in turn affects the probability of application.

For the empirical portion of the paper, I report the results of a field experiment in which graph-based measures were used to make recommendations to employers in an online labor market. These recommendations were effective at increasing employer recruiting and were more effective for harder-to-fill openings, as measured by the number of skills required.

2 Model

A worker i, if hired for job opening j will have productivity

$$y_{ij} = y_0 m_{ij} + \epsilon \tag{1}$$

where $m_{ij} \in \{0,1\}$ is a measure of whether the worker matches the "type" of the job opening. Workers and firms observe m_{ij} but the researcher does not. The ϵ term has a cdf $\Phi(\cdot)$. The error term is also *not* observed by the researcher but is observed by the market participants themselves, and so workers know y. The outside option of not applying is fixed at $\underline{u}$.

Assume that there are two types and that they are in equal proprtion on boths sides of the market: 1/2 of jobs are Type 1 and the other half are Type 2. A worker will apply to a job opening for which she matches types ($m_{ij} = 1$) with probability p_H and a job opening where she does not match types ($m_{ij} = 0$) with probability p_L, where

$$p_H = \Phi(y_0 - \underline{u}); \quad p_L = \Phi(-\underline{u}). \tag{2}$$

As expected, $p_H > p_L$.

For each worker, we observe a binary vector of length J, where J is the number of job openings. The elements of the vector are indicators for whether the worker applied to that particular opening, i.e.,

$$A_i = (a_{i0}, a_{i1}, a_{i2} \ldots a_{iJ}). \tag{3}$$

The application vectors collectively define the application "graph." For any two workers, we can compute various measures of similarity between the workers. Let $S_{ii'}$ be the realized count of choice matches for workers i and i'. The simple matching coefficient (SMC) is defined as $S_{ii'}/J$.

The probability that two workers randomly selected from the pool of workers will match their choices (i.e., either both apply or not apply) for a given job opening also selected at random is

$$p_M = \frac{1}{2}\left(p_H^2 + (1-p_H)^2\right) + \frac{1}{2}\left(p_L^2 + (1-p_L)^2\right). \tag{4}$$

whereas the probability of not matching is

$$p_{NM} = \left(p_H p_L\right) + (1-p_H)(1-p_L). \tag{5}$$

As expected, $p_M > p_{NM}$.

For two workers selected at random, the prior probability that they are the same type as each other is $\frac{1}{2}$. The realized number of choice matches between the two—out of a total of J possible—is a binomial random variable, $S_{ii'}$. If the two workers are of the same type, then

$$S_{ii'}|\text{Type}(i) = \text{Type}(i') \sim \text{Binom}(J, p_M); \quad S_{ii'}|\text{Type}(i) \neq \text{Type}(i') \sim \text{Binom}(J, p_{NM}). \tag{6}$$

We can use the realized number of choice matches between two workers to compute the posterior that they are of the same type:

$$Pr\{\text{Type}(i) = \text{Type}(i')|S\} = \frac{1}{1 + \left(\frac{p_{NM}}{p_M}\right)^S \left(\frac{1-p_{NM}}{1-p_M}\right)^{J-S}}. \tag{7}$$

Let q be the posterior $Pr\{\text{Type}(i) = \text{Type}(i')|S\}$. With some algebra, we can re-write the expression for the posterior as

$$\log(1/q - 1) = J\left(SMC \times \log x + \log y\right) \tag{8}$$

where $x = \frac{p_{NM}(1-p_M)}{p_M(1-p_{NM})}$ and $y = \frac{1-p_{NM}}{1-p_M}$. We can see that the posterior probability that two workers are the same type is monotonically increasing in the computed SMC for those two workers, as

$$q'(SMC) = -J\left(\frac{1-q}{q}\right)\log x > 0, \tag{9}$$

which follows from that fact that $x < 1$ for all values of p_{NM} and p_M.

We can readily take the model to data and estimate p_M and p_{NM} and hence p_H and p_L and $y_0/\underline{\mu}$. As we are pairing workers at random, the posterior that any two workers are in fact of the same type must, over all pairings simply be the same prior probability of a match, and so $\mathbb{E}[q] = \frac{1}{2}$ and thus $\mathbb{E}[1/q - 1] = 1$. Further, the expected number of matches per job opening for two workers must match the population estimates of p_{NM} and p_M. This gives us two moment conditions:

$$\mathbb{E}\left[\left(\frac{p_{NM}}{p_M}\right)^{S_k}\left(\frac{1-p_{NM}}{1-p_M}\right)^{J-S_k}\right] - 1 = 0 \tag{10}$$

$$\mathbb{E}\left[S_k/J\right] - 1/2\left(p_M + p_{NM}\right) = 0. \tag{11}$$

where k indexes the $(1, 2, \ldots I \times (1, 2, \ldots J)$, where I is the number of workers

3 Experimental design and results

Employers on oDesk with job openings were randomized to receive “more like this” recommendations on workers they viewed in search or who had applied to their job openings. These “more like this” recommendations were determined solely from graph-based measures calculated using historical application data.

Table 1, Column (1) reports a regression where the dependent variable is the number of recruiting invitations sent. In Column (2), the dependent variable is the log number of invitations sent, conditional upon sending any. We can see from Column (1) that the treatment significantly increased the number of invitations sent. In percentage terms, Column (2) shows that the treatment increased recruiting by about 5 percentage points. In Column (3), I test whether the number of required skills for the employer job opening affects the treatment. Column (3) shows that the effect of the treatment on recruiting was strongly increasing in the number of skills required.

4 Conclusion

This paper shows that purely graph-based measures of worker similarity are valued by employers. The theory modeling provides a potential explanation of why these graph-based measures are useful.

Table 1: Effect of the MLT treatment on employer recruiting

	Dependent variable:		
	Number of invites	Log number of invites	Number of invites
	(1)	(2)	(3)
Treatment (More Like This)	0.034^{**}	0.053^{***}	-0.033
	(0.017)	(0.012)	(0.031)
Number of skills req'd			0.016
			(0.011)
Number of skills req'd × Treatment			0.042^{***}
			(0.016)
Intercept	0.807^{***}	0.565^{***}	0.782^{***}
	(0.012)	(0.009)	(0.022)
Observations	37,356	12,967	37,356
R^2	0.0001	0.001	0.001
Adjusted R^2	0.0001	0.001	0.001
Residual Std. Error	1.627 (df = 37354)	0.700 (df = 12965)	1.627 (df = 37352)
F Statistic	4.173^{**} (df = 1; 37354)	18.539^{***} (df = 1; 12965)	10.643^{***} (df = 3; 37352)

Notes: The table reports OLS regresses where the dependent variable is the number of recruiting invitations sent by the employer. Columns (1) and (3) are in levels, whereas Column (2) is in logs.

References

Adomavicius, Gediminas and Alexander Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *Knowledge and Data Engineering, IEEE Transactions on*, 2005, *17* (6), 734–749.

Linden, Greg, Brent Smith, and Jeremy York, "Amazon. com recommendations: Item-to-item collaborative filtering," *Internet Computing, IEEE*, 2003, *7* (1), 76–80.