Skill Networks and Measures of Complex Human Capital: Supplemental Information

Alternative Link Weights

In the Methods section of the main text, we discuss the use of similarity weights in the skill networks. We have chosen this weighting scheme because it provides insight into the problem we address. There are several alternatives that may be useful in alternative analyses. The simplest is a *frequency weight*: $w_{ij}^{freq} = n_{ij}$. While this weighting reflects the overlap between the two skills in the labor market, it is not normalized, and does not provide as much information about the functional synergies between skills, because they only reflect the popularity of the skills.

Another alternative is to normalize the frequency by the overall prevalence of both skills: $w_{ij}^{JSI} = \frac{P(s_i \cap s_j)}{P(s_i \cup s_j)} = \frac{n_{ij}}{n_i + n_j - n_{ij}}$. We call this a *Jaccard Similarity Index (JSI) weight.* It is important to note that JSI weights are higher when there is a large overlap between two common skills than when there is a large overlap between a rare skill and a common skill. The upper bound on w_{ij}^{JSI} is set by the difference in how often the two skills occur.

The difference between these three weighting schemes is illustrated in Figure 1.

One might also consider cosine similarity: a weighting scheme often used in citation networks. In that context, cosine similarity is given by $w_{ij}^{cosine} = \frac{\sum_{papers} k_{ip} k_{jp}}{\sqrt{\sum_{papers} k_{ip}^2} \sqrt{\sum_{papers} k_{jp}^2}}$, where k_{ip} is the number of citations of discipline i in paper p. Here, each skill set will contain only one copy of each skill, and thus cosine similarity becomes a function of counts: $w_{ij}^{cosine} = \frac{n_{ij}}{n_i n_j}$. This weight is inappropriate in skill networks because some skills are much less common than others (see Figure 3 below), making w_{ij}^{cosine} very close to $\frac{1}{n_i}$. It is possible that this weighting scheme is useful in some context. However, w_{ij}^{JI} is likely the superior weighting scheme for most applications.

Data

For the illustration in this paper, we have used data drawn from the largest online freelance labor market in the world: UpWork. There are two sides to the UpWork market: worker profiles and job postings. Workers apply for jobs and employers hire and pay workers through the site. Worker profiles include a list of skills and a list of previous jobs with associated hourly wages. Job postings contain information about the job, including a list of required skills.

The data used in this paper was collected from UpWork over a period of three months, between November 2013 and January 2014.[1] We collected a total of 33,592 worker profiles and 365,561 job

[1] UpWork is the result of a merger between oDesk and their largest competitor–Elance–in December 2013. However, the two competitors maintained separate platforms until May 2015, when the two platforms merged to become

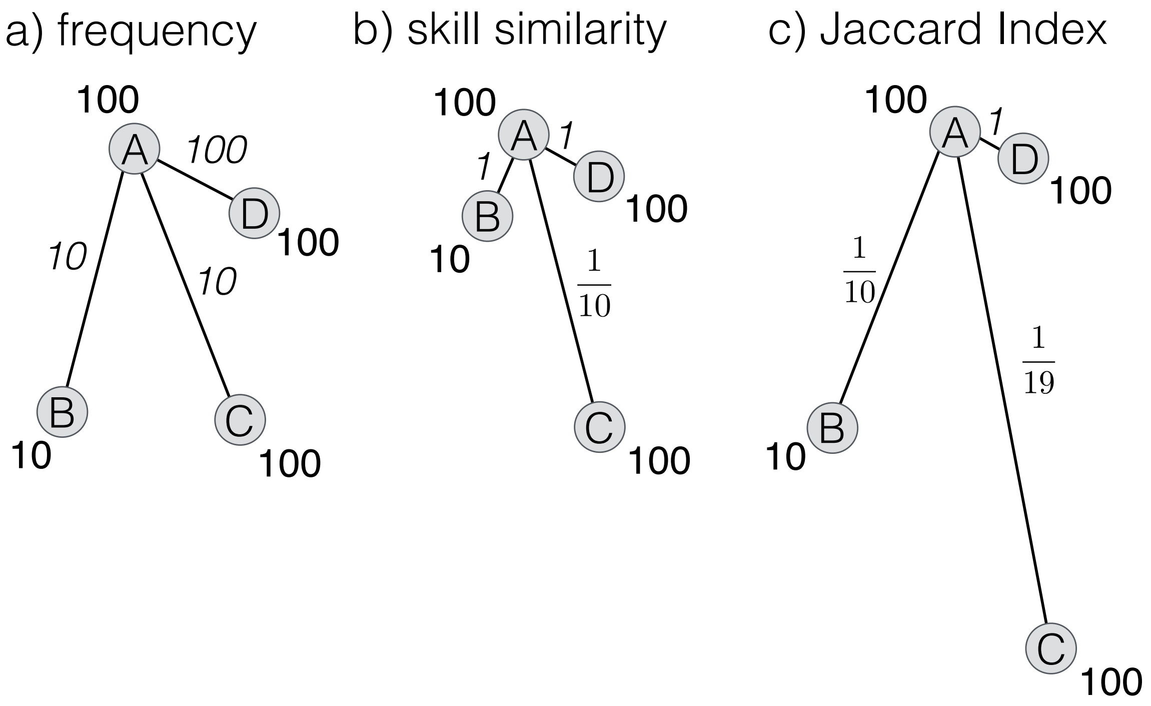

Figure 1: An illustration of three possible weighting schemes for links in a human capital network

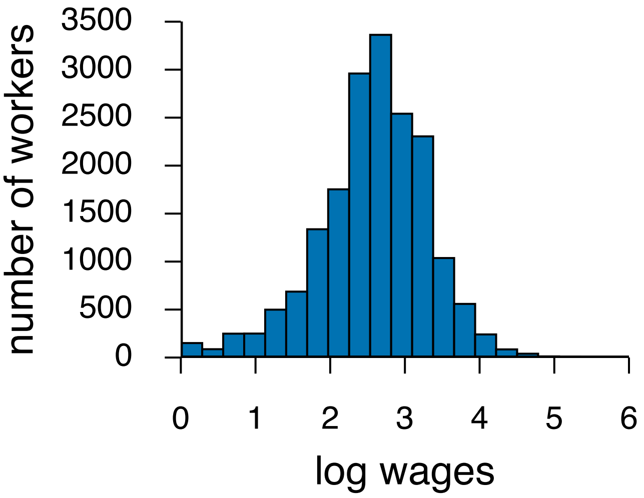

Figure 2: Distribution of Hourly Wages

	Workers with wages and skills
count	18283
# of skills	6.7 (0.02)
avg wage	$16.84 (0.10)
jobs worked	18.0 (0.25)
hrs worked	768 (13.1)

Table 1: UpWork worker summary statistics

listings at random from the public part of the website. From that sample, we have dropped 5266 workers who listed no skills (17% of the sample), because they are irrelevant in our analysis. We have also dropped workers with more than 10 skills because there is evidence that these profiles are not actually held by individual workers, but rather conglomerates.. Only 66 workers fall into this category (0.2% of the sample). This leaves an initial sample of 26,046 workers. We use this full population to construct our skill networks, because all of these workers are visible to employers, and thus are part of the labor market. When making statements about wages, we will restrict our sample to those with a wage history on the site–there are 18,165 of these workers (70% of the pool). The distribution of workers' hourly wages shows typical inequality (see Figure 2). Summary statistics can be found in Table 1).

The UpWork market is ideal for this analysis, because workers and employers are limited in the skills they can list on their profile: workers and employers must choose their skills from the site's master database of allowable skills, and must make a formal request that a new skill be added. This eliminates any ambiguity in skills due to spelling errors or synonymous entries [1]. There are 2197

UpWork. The data collected here was exclusively from the oDesk platform, before the two were merged. However, the site has remained essentially the same post-merger.

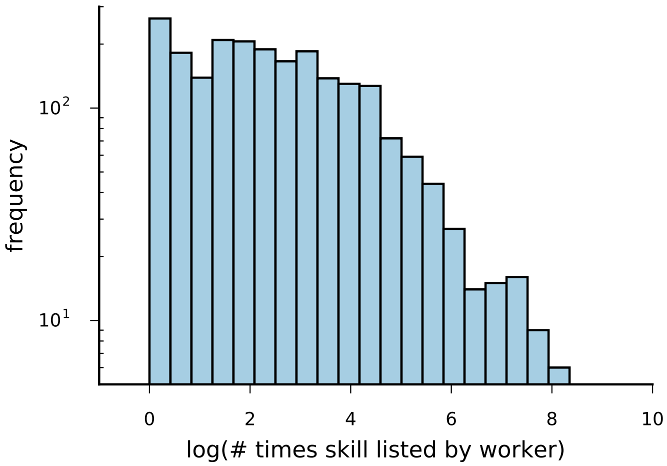

Figure 3: Distribution for the number of times a skill is listed by workers.
Most skills are very uncommon, while a few are held by a huge number of workers and required for a huge number of jobs.

skills listed by workers in our full sample and 2477 listed by job postings. The prevalence of these skills is highly uneven–most skills are very uncommon, while there are a few skills that appear in a huge number of worker profiles and job listings (see Figure 3 for the case of workers). 20% of skills are held by less than .1% of workers, and 12% of skills are held by only *one* person in this sample. Among job postings, 57% of skills are required by less than .01% of jobs and 7% of skills are unique. Table 2 shows the most common skills among workers and jobs.

Most common job skills	Most common worker skills
php	adobe photoshop
wordpress	php
graphic design	javascript
adobe photoshop	html
content writing	mysql
article writing	css
javascript	microsoft excel
data entry	wordpress
internet research	data entry
web design	article writing

Table 2: The most common skills among worker profiles and job listings.

	Worker network	Job network
N	1933	2293
average degree	107	110
density	0.056	0.048
number of communities	6	12
modularity of partition	0.47	0.50

Table 3: Basic statistics about the worker and job skill networks

Admin/Writer/Sales	Artist/Designer	Programmer/Technical			Testing
		General	Mobile/Stats	IT/Network Admin	
microsoft excel	adobe photoshop	php	c#	linux systems admin	software testing
data entry	adobe illustrator	javascript	c++	perl	software qa testing
article writing	web design	html	.net framework	windows admin	manual testing
microsoft word	graphic design	mysql	asp.net	apache admin	functional testing
blog writing	adobe indesign	css	android app dev	network admin	usability testing
creative writing	logo design	wordpress	c	lamp admin	regression testing
internet research	adobe flash	jquery	microsoft access	amazon web services	atlassian jira
customer service	illustration	java	ios dev	unix shell	black box testing
proofreading	adobe dreamweaver	ajax	objective c	computer networking	automated testing
editing	photography	html5	ms visual basic	unix system admin	web testing

Table 4: Top skills in each worker category.

Analysis

Nodes in the two human capital networks are skills and a connection is made whenever two skills are both listed by a worker (in the case of the supply-side network) or a job (in the case of the demand-side network). The links are weighted by skill similarity: $w_{ij}^{sim} = P\left(s_i|s_j\right) = \frac{n_{ij}}{n_j}$ where n_i and n_j are the number of workers who have skills i and j respectively and $n_j < n_i$. For clarity, we drop skills that occur only once in our sample. This will have no effect, because the link weights from those skills to others are, by definition, 1. After dropping unique skills, we are left with a worker network with 1933 nodes and a job network with 2293 nodes. Basic information about these two networks can be found in Table 3.

We divide both networks into communities using the Louvain method ([2]). The modularity of this partition is 0.47 in the worker network and 0.50 in the job network, which suggests significant community structure. We define the skill categories by hand, using the most common 100 skills in each. The top 10 skills in each skill cluster can be found in Tables 4 (workers) and 5 (jobs). Table 6 shows the number of skills in each of these clusters as well as the number of workers who specialize in that area.

An image of the full worker and job networks can be seen in Figures 4 and 5, respectively. The size of the nodes corresponds to the number of workers who have them. The colors represent the different skill categories.

There is a significant amount of overlap between the skills in the two networks: 1826 skills are present in both worker profiles and job postings. This allows us to create a mapping between the

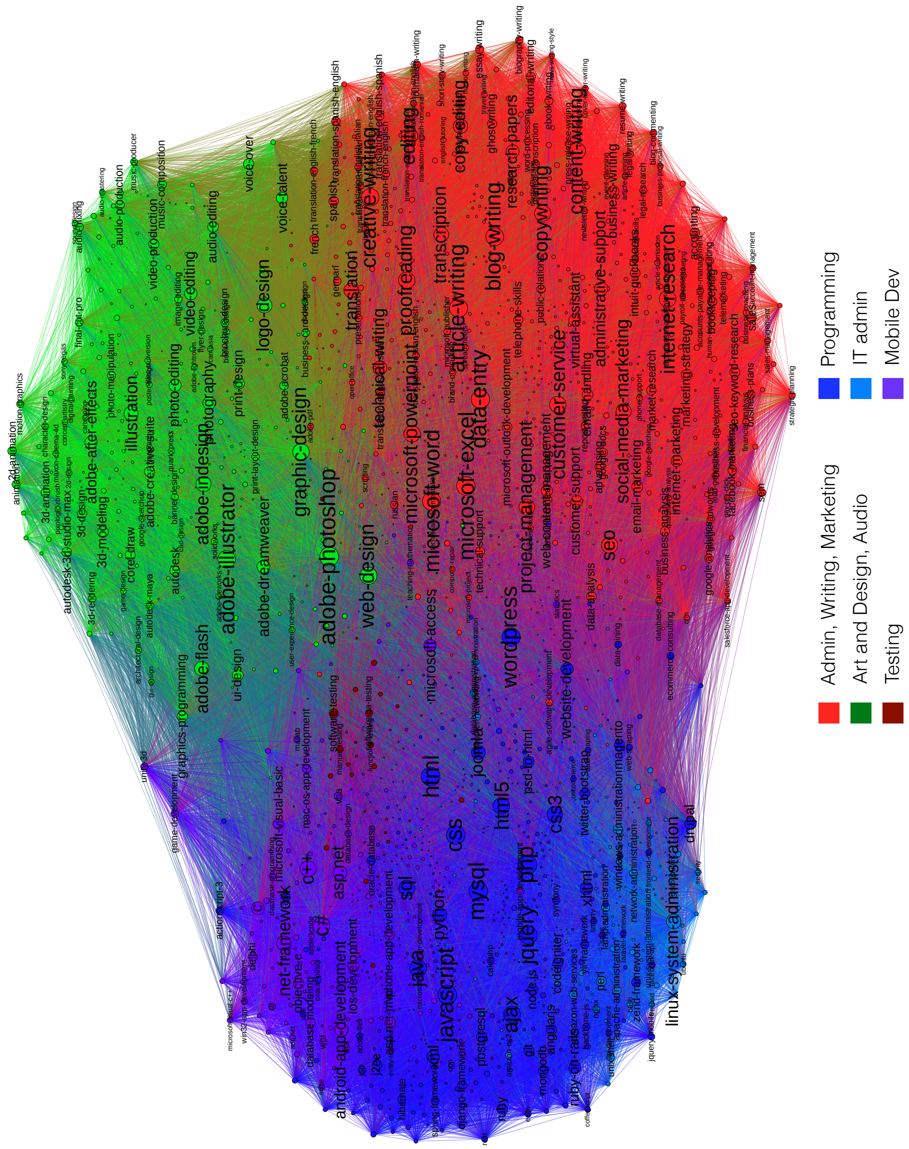

Figure 4: Worker Human Capital Network: Nodes are sized according to the number of workers who list them, and they are colored according to their category.

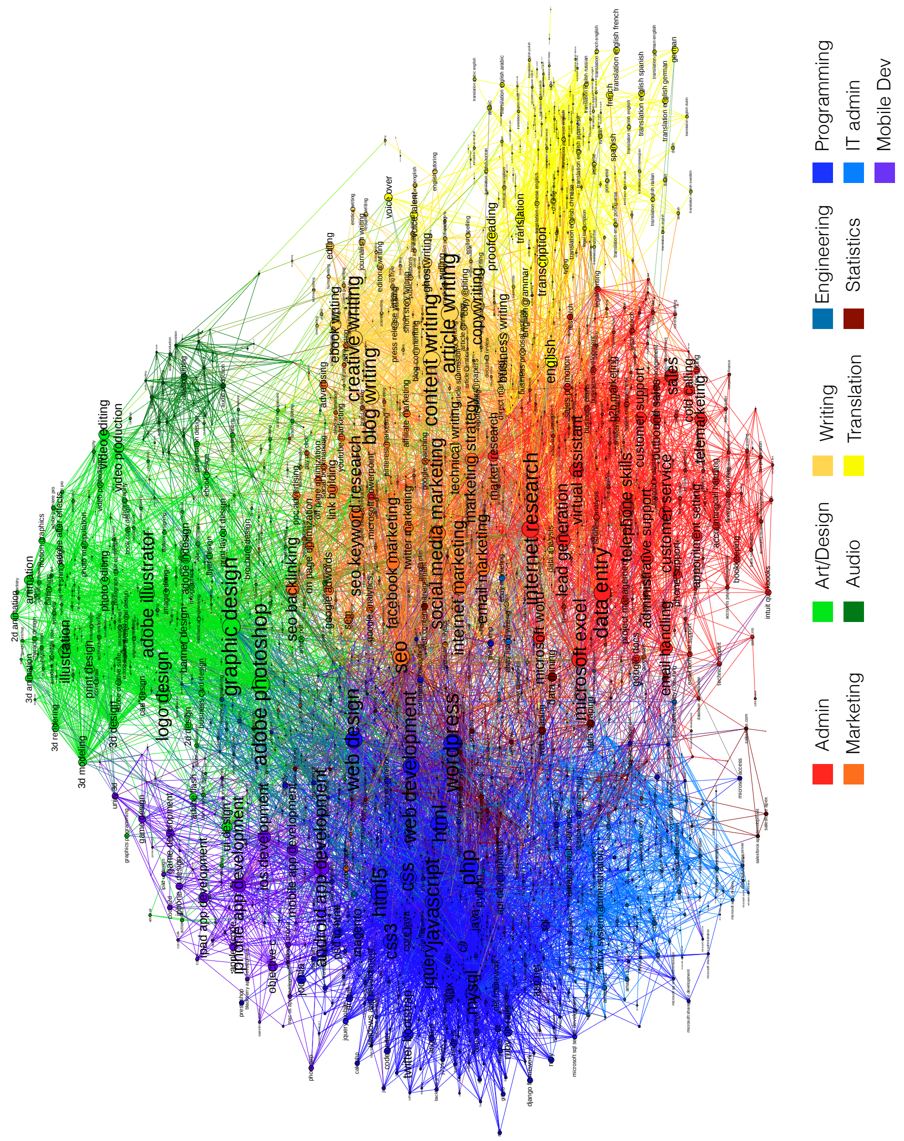

Figure 5: Job Human Capital Network: Nodes are sized according to the number of job postings that list them, and they are colored according to their category.

Administrative	Writing	Translation	Art/Design	Music/Audio	Marketing
data entry	content writing	english	graphic design	audio editing	seo
internet research	article writing	transcription	adobe photoshop	audio production	social media market
microsoft excel	blog writing	proofreading	adobe illustrator	audio mixing	internet marketing
lead generation	creative writing	translation	logo design	audio post-product	seo keyword research
sales	copywriting	voice over	illustration	music composition	email marketing
telephone skills	technical writing	voice talent	video editing	audio mastering	marketing strategy
virtual assistant	ebook writing	french	UI design	audio engineering	seo backlinking
customer service	business writing	spanish	video production	music producer	facebook marketing
email handling	ghostwriting	german	print design	music arrangement	twitter marketing
telemarketing	english grammar	english to french	animation	sound editing	link-building

Programming/Technical					Testing
General	Mobile Dev	IT/Network Admin	Engineering	Data/Statistics	
php	android app dev	linux system admin	c	data mining	software testing
wordpress	iphone app dev	amazon web services	electrical engineering	data scraping	QA testing
javascript	ios dev	ebay listing writing	electronics	web scraping	web testing
web design	mobile app dev	ebay marketing	matlab	web content mgmt	manual testing
html5	ipad app dev	network admin	pcb design	scripting	functional testing
html	objective c	cpanel	arduino	web crawler	automated testing
css	game dev	email deliverability	microcontroller prog	website wireframing	usability testing
website dev	apple xcode	amazon ec2	circuit design	salesforce apex	database testing
css3	game design	network security	embedded systems	salesforce.com	localization
mysql	iphone UI design	voip software	electrical drawing	salesforce app dev	selenium

Table 5: Top skills in each job category. One very small category has been omitted.

	number of skills	number of specialized workers	average wage
admin/writer	709	4896	$13.51 (0.183)
art and design	334	1840	$16.54 (0.264)
programming	413	1634	$19.93 (0.324)
mobile dev/stats	267	300	$19.93 (0.645)
IT admin	174	151	$23.04 (1.09)
testing	36	38	$9.35 (.849)

	number of skills	number of specialized workers	average wage
admin	306	1096	$11.01 (0.395)
writing	193	509	$14.70 (0.434)
translation	193	408	$13.54 (0.558)
art and design	343	1295	$15.52 (0.280)
audio	45	71	$16.54 (1.30)
marketing	144	137	$18.63 (1.45)
programming	579	2396	$19.56 (0.254)
mobile dev	104	43	$22.10 (2.01)
IT admin	214	125	$22.32 (1.26)
engineering	55	15	$23.87 (4.81)
data/statistics	70	13	$21.59 (4.62)
testing	42	34	$9.27 (0.935)

Table 6: Statistics for different worker and job categories

		Categories of worker						
		admin	prog.	IT admin	mobile/stats	art/design	testing	total
Job categories they qualify for	admin	234	7	3	4	4		252
	writing	136	2		22	5		165
	translation	148			1	3		152
	marketing	75	18	2	4	4		103
	programming	26	297	20	109	5	3	460
	IT admin	20	13	130	6	1		170
	mobile dev	6	10		53	7	3	79
	art/design	25	8	1	7	238		279
	engineering	2	3	1	34	1		41
	audio	1				39		40
	data/statistics	15	15	3	10	4	1	48
	testing	4	2	3	1		26	36
	total	692	375	163	251	311	33	1825

Table 7: Types of jobs a worker of a given type might be qualified for. The last row and column show the total number of skills in each worker and job category (respectively).

types of specialized workers in the labor pool, and the types of specialized jobs that they qualify for (Table 7). The correspondence between worker categories and job categories is quite strong: worker skills identified as "programming" tend to be useful for "programming" jobs. This is further evidence that the categories we have measured are a meaningful partition of skills.

There is also a correspondence between UpWork's exogenous categories and our network-based categories. The differences are highlighted in Table 8. In some cases, UpWork's categories encompass several of our endogenous categories. In others, UpWork's categories provide more detail. Interestingly, some of the job skill categories lumped together on UpWork pay significantly different wages. For example, workers who qualify for software testing jobs earn much less than workers in any other category, which suggests that those jobs should not be included with programmers. Also, it appears that workers who qualify for audio jobs do not necessarily qualify for art and design jobs, and vis-versa, suggesting the UpWork would improve their ability to match workers to jobs by separating those two categories of jobs. Finally, our method has combined several of UpWork's Sales and Marketing categories and Administrative categories into a single category of workers. This suggests that there is significant overlap in the skills required for each of these UpWork categories, in turn suggesting overlap in the labor pools. For example, many of the same skills required for legal jobs will also be required for accounting and consulting jobs, suggesting there may be substantial number of workers who will qualify for both.

Human Capital and Wages

Our next comparison is the wages for workers with different levels of skill diversity. The columns of Table 9 lists average wages for workers who specialize in one area and who have more diverse skills. The workers with more diverse skills have higher wages, both in aggregate (the last row) and when divided according to how many jobs they cross. The same is true when we look at the intensive margin. Figure 9 shows wages for individuals with different levels of skill diversity. The size of the

Network-Based Job Skill Category	UpWork Job Category
Web, Mobile and Game Dev.	Web, Mobile and Software Dev.
Software Dev.	
Software Testing	
IT and Network Administrative	IT and Network Admin
Data Science and Analytics	Data Science and Analytics
Art and Design	Design and Creative
Music and Audio	
Engineering and Physical Design	Engineering and Architecture
Sales and Marketing	Sales and Marketing
	Accounting and Consulting
	Legal
Admin Support	Admin Support
	Customer Service
Writing	Writing
Translation	Translation

Table 8: The job categories suggested by the job-side human capital network (left) are similar–but not identical–to the categories provided by the managers of the labor market (right).

	Specialized Skills	Diverse Skills	total
Fit one type of job	$\substack{\$15.91 \\ (.174)}$ (n=5253)	$\substack{\$18.96 \\ (.371)}$ (n= 889)	$\substack{\$16.36 \\ (.159)}$ 6142
Fit multiple types of jobs	$\substack{\$15.36 \\ (.219)}$ (n=3606)	$\substack{\$17.82 \\ (.150)}$ (n=8417)	$\substack{\$17.08 \\ (.124)}$ 12023
total	$\substack{\$15.69 \\ (.136)}$ (n=8859)	$\substack{\$17.93 \\ (.139)}$ (n=9306)	

Table 9: Average wages for workers with different combinations of skills

circles indicates the number of workers who have that combination of skills. The colors represent the average wage of individuals in that bin. The first column shows wages for specialized workers. Scanning from left to right shows the wages for workers whose skills span an increasing number of worker areas, suggesting increasing skill diversity. We see that there are fewer workers with diverse skills, but they tend to have higher wages than the specialists.

Worker wages also vary in the number of jobs they qualify for. In Table 9, we see that specialized workers gain no advantages from qualifying for more than one type of job. However workers with diverse skills earn higher wages when they qualify for a small number of jobs. This is presumably because workers whose diverse skills qualify them for different types of job are more likely to be using their skills independently, while those who qualify for a single job area are likely using a rare combination of skills in an area with high demand.

Putting workers on a simplex provides another view of how skill breadth affects wages. In a simplex, each worker will have a vector $s_i =< s_i^a, s_i^b ... s_i^M >$ where $s_i^x = \frac{k_i^x}{k_i}$ is the fraction of worker i's skills that fall into category x, and (necessarily) $\sum_{x=1}^{M} s_i^x = 1$. This gives each worker a location on a plane in M dimensional space. For simplicity, we will divide worker skills into just three categories: art and design, programming (including IT and mobile development), and administrative (including testing). This means that every worker will have a vector $\left\langle s_i^d, s_i^p, s_x^a \right\rangle$, with entries corresponding to

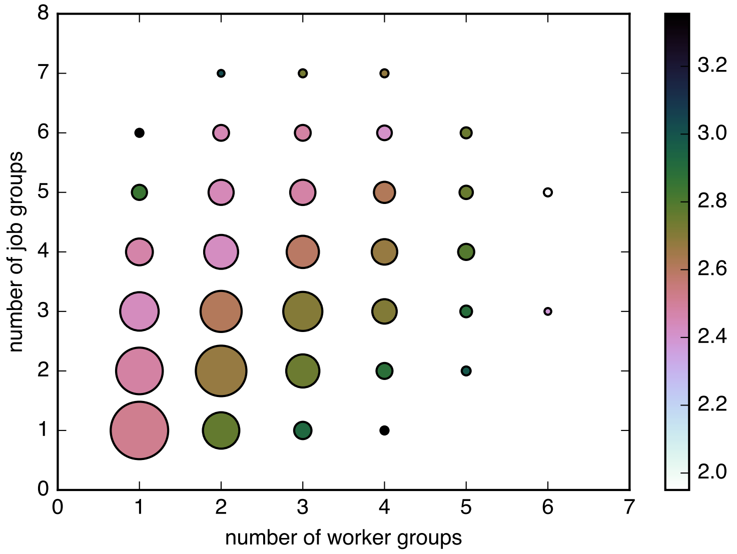

Figure 6: Log wages for workers with different distributions of skills across the worker and job networks. The size of each point indicates how many workers are in each group, while colors represent log wages. The workers in the first column on the left have skills in one area, the workers in the other columns have skills in multiple areas. The vertical axis represents the number of categories of jobs a worker qualifies for.

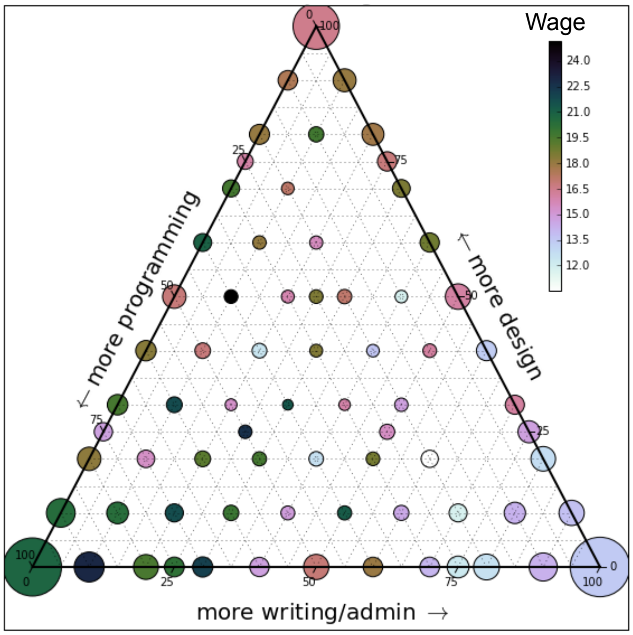

Figure 7: Average wages for workers with different combinations of skills. Each point on this simplex, $< s_d, s_p, s_a >$ represents a distribution of skills across design/programming/administration space. The color of each point indicates the average wage for workers with that skill combination. The size indicates the number of workers in that bin.

the fraction of the worker's skills in each of those three areas, and each worker occupies a point on a plane. Figure 7 projects this plane into two dimensions, illustrating average wages for workers with different combinations of skills. The size of the circles indicates the number of workers who have that combination of skills. The circles in the corners are pools of workers whose skills are specialized (fall into a single area). The circles between the corners are pools of workers with more diverse skill sets. The colors represent the average wage of individuals in that bin. We see that there are fewer workers with diverse skills, but they tend to have higher wages than the specialists.

The value of considering skill combinations

Tables 10 and 11 show the results of the regressions referenced in the text. Model 1 (the first column of Table 10) is a linear model of log wages with dummies for the individual skills. When too many dummies are included, the terms become colinear, so we restrict ourselves to the most common skills: those held by at least 2% of the population. Due to the number of dummies involved, for each of these models, we include a random selection in the table. As expected, some of these coefficients are significant and others are not. In Model 2 (the second column of Table 10), we also include dummies for worker categories identified through the network analysis: mobile development, testing, programming, IT, and art and design. We omit administration as a comparison. In this case, we restrict our sample to workers who specialize in a single area.[2] The worker category is a significant factor for all of the areas, with coefficients in the expected directions. The effect sizes are quite large. For example, workers in programming fields earn 50% more than administrative workers with the same skills (~$5.40), while workers in software testing earn 30% less (~$4.00). In addition, the adjusted R^2 of this model is higher, indicating the network-based terms explain variance in wages, beyond that explained by the skills independently.

In Model 3 (the third column of Table 10), we consider the number of job areas the worker crosses. Again, in spite of controlling for the skills individually, the coefficient on the network-based measure is significant, and the adjusted R^2 value is higher. The effect size here is smaller, but still signficant: qualifying for jobs in one additional area increases your wages by approximately 5% over others with your same skills (~$0.65). In Model 4 (the fourth column of Table 10), we include the number of worker categories crossed. Again, the coefficient on this variable is significant and the adjusted R^2 value goes up. The effect size is similar to that of the job measure. Workers who cross an additional skill area earn 5% more (~$0.62) than workers with similar skills.

These results are the same, even when controlling for larger numbers of individual skills. When we regress log wages against the skill categories and an increasing number of skill dummies, the coefficients remain significant (See Table 11). One would expect these coefficients to become insignificant as the number of dummies increases: with over 1000 dummies it becomes difficult for any one of them to explain a significant amount of variation in the dependent variable. That they do not is strong evidence for the benefits of considering skill interactions.

Ideally, we would also examine a model with a more traditional one-dimensional measure of human capital, such as years of education or experience. Unfortunately, in this data, we observe neither. The number of worker skills is sometimes used in the study of wage effects. However, it is problematic in the context of the UpWork labor market. Clearly some skills will be much more valuable than others, making a simple count a very poor proxy for human capital. Nonetheless, in models with the number of skills standing in for the skill vector, the results are similar to above (see Table 12). The coefficients on the network-based measures are significant, have meaningful effect sizes, and explain variation beyond that explained by the number of skills (in some cases, quite a lot more, presumably because the number of skills is such a poor human capital measure). However, for the reasons above, it is important to not place too much weight on these results. We report them here for completeness.

References

[1] Horton, J. J. (2010) *Online labor markets.* (Springer), pp. 515–522.

[2] Blondel, V. D, Guillaume, J.-L, Lambiotte, R, & Lefebvre, E. (2008) Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008.

[2] The results are very similar if we use non-specialists as the reference group.

	Model 1	Model 2	Model 3	Model 4
N	18165	8859	18165	18165
Adj R^2	0.178	0.190	0.180	0.183
constant	2.6***	2.36***	2.54***	2.51***
program.		0.51***		
IT admin.		0.59***		
mobile dev.		0.47***		
art & design		0.26***		
testing		-0.3***		
number of skill areas			0.05***	
number of job categories				0.05***
html	-0.08***	-0.08**	-0.09***	-0.09***
illustration	0.12***	0.15***	0.12***	0.13***
social media marketing	0.04*	0.12***	0.04	-0.0
mysql	0.08***	0.01	0.07***	0.07***
project management	0.18***	0.24***	0.16***	0.14***
adobe photoshop	-0.13***	-0.17***	-0.14***	-0.15***
microsoft word	-0.15***	-0.11***	-0.15***	-0.16***
css	0.04**	-0.0	0.04**	0.05**
data entry	-0.48***	-0.37***	-0.48***	-0.49***
blog writing	-0.04*	0.01	-0.05**	-0.06**
ajax	0.03	0.0	0.04	0.04*
technical writing	0.05*	0.08*	0.04	0.02
article writing	-0.11***	-0.08***	-0.11***	-0.12***
creative writing	-0.11***	-0.06*	-0.11***	-0.12***
logo design	-0.06**	-0.08	-0.05*	-0.06**
adobe illustrator	0.07**	0.09**	0.08***	0.08***
copy editing	0.09***	0.12***	0.09***	0.09***
administrative support	-0.08**	-0.0	-0.07**	-0.08***
web design	0.08***	0.12***	0.06***	0.05**
customer service	-0.23***	-0.18***	-0.23***	-0.23***
jquery	0.1***	0.01	0.11***	0.11***
javascript	0.06***	0.08**	0.06***	0.06***
editing	0.1***	0.16***	0.1***	0.08***
graphic design	0.07***	0.03	0.07***	0.06**

Table 10: Regressions including network-based skill measures and dummies representing the workers' individual skills. Model 1: dummies alone. Model 2: dummies and skill categories determined from the network. Model 3: dummies and number of skill areas. Model 4: dummies and number of work groups crossed.

	Model 1	Model 2	Model 3	Model 4
N	18165	8859	18165	18165
Adj R^2	0.030	0.108	0.031	0.039
constant	2.31***	2.40***	2.17***	2.28***
number of skills	0.04***	0.03 ***	0.03***	0.04***
program.		0.46***		
IT admin.		0.64***		
mobile dev.		0.56***		
art & design		0.31***		
testing		-0.25**		
number of skill areas			0.03***	
number of job categories				-0.07***

Table 11: Regressions with network-based measures in combination with the number of skills. Model 1: number of skills alone. Model 2: number of skills and skill categories determined from the network. Model 3: number of skills and number of skill areas. Model 4: number of skills and number of work groups crossed.

	Model 1	Model 2	Model 3	Model 4
N	8859	8859	8859	8859
Adj R^2	0.190	0.216	0.236	0.262
constant	2.36***	2.28***	2.25***	2.22***
program.	0.51***	0.49***	0.52***	0.54***
IT admin.	0.59***	0.67***	0.69***	0.65***
mobile dev.	0.47***	0.55***	0.55***	0.51***
art & design	0.26***	0.3***	0.27***	0.29***
testing	-0.3***	-0.23**	-0.19*	-0.22
html	-0.08**	-0.07*	-0.07*	-0.06
illustration	0.15***	0.16***	0.13***	0.12***
proofreading	0.09***	0.1***	0.1***	0.1***
content writing	0.07**	0.07**	0.07**	0.07**
social media marketing	0.12***	0.04	0.03	0.03
project management	0.24***	0.21***	0.2***	0.19***
adobe photoshop	-0.17***	-0.14***	-0.1**	-0.1**
adobe illustrator	0.09**	0.09**	0.09*	0.09**
microsoft word	-0.11***	-0.09**	-0.09**	-0.06
wordpress	-0.11***	-0.07*	-0.07*	-0.07
data entry	-0.37***	-0.35***	-0.34***	-0.31***
microsoft excel	-0.11***	-0.12***	-0.12***	-0.14***
technical writing	0.08*	0.1**	0.11**	0.12***
article writing	-0.08***	-0.07**	-0.07**	-0.04
copywriting	0.29***	0.26***	0.26***	0.25***
creative writing	-0.06*	-0.05*	-0.04	-0.05
logo design	-0.08	-0.09*	-0.08	-0.08
copy editing	0.12***	0.12***	0.12***	0.1**
web design	0.12***	0.04	0.06	0.06
business writing	0.26***	0.18***	0.15***	0.13***
internet research	-0.06*	-0.05*	-0.05	-0.04
customer service	-0.18***	-0.16***	-0.14***	-0.13***
3d modeling	0.1**	0.06	0.08	0.08
adobe indesign	0.08	0.09*	0.09*	0.07
editing	0.16***	0.17***	0.17***	0.17***
javascript	0.08**	0.03	0.03	0.03
adobe after effects	0.07	0.05	0.01	0.07
virtual assistant	-0.08	-0.08*	-0.06	-0.05

Table 12: Regressions including the skill categories determined from the network and an increasing number of dummies.