Crowds, Gigs, and Super Sellers: A Measurement Study of a Supply-Driven Crowdsourcing Marketplace

Hancheng Ge and **James Caverlee**
Department of Computer Science and Engineering
Texas A&M University
{hge, caverlee}@cse.tamu.edu

Kyumin Lee
Department of Computer Science
Utah State University
kyumin.lee@usu.edu

Abstract

The crowdsourcing movement has spawned a host of successful efforts that organize large numbers of globally-distributed participants to tackle a range of tasks. While many demand-driven crowd marketplaces have emerged (like Amazon Mechanical Turk, often resulting in workers that are essentially replace-able), we are witnessing the rise of supply-driven marketplaces where specialized workers offer their expertise. In this paper, we present a comprehensive data-driven measurement study of one prominent supply-driven marketplace – Fiverr – wherein we investigate the sellers and their offerings (called "gigs"). As part of this investigation, we identify the key features distinguishing "super sellers" from regular participants and develop a machine learning based approach for inferring the quality of gigs, which is especially important for the vast majority of gigs with little feedback.

Introduction

The crowdsourcing movement has spawned a host of successful efforts that organize large numbers of globally-distributed participants to tackle a range of tasks, including crisis mapping (Gao et al. 2011), translation (Zaidan and Callison-Burch 2011), and protein folding (Khatib et al. 2011). In one direction, we have seen the rise of crowdsourcing marketplaces that aim to connect task requesters with task workers. Successful examples of these crowdsourcing marketplaces include Amazon Mechanical Turk (AMT) and CrowdFlower, as well as niche markets like Microworkers and Shorttask. Many of these marketplaces are fundamentally *demand-driven*: that is, a requester posts a task (e.g., annotate an image with a set of related keywords) and workers are for the most part treated as replaceable commodities that can complete the task. Some tasks may require particular qualifications, but the workers themselves are subject to the demands of the requesters.

In a separate direction, there has been a rise of *supply-driven* marketplaces, where the workers (the "suppliers") are the main drivers of the types of tasks that may be accomplished. In a supply-driven marketplaces, workers advertise their skills and special talents with the hopes of differentiation from each other, in sharp contrast to the demand-driven marketplaces in which participants are essentially interchangeable. Popular freelancing services like Elance, Guru.com, oDesk and Freelancer can be viewed as examples of supply-driven marketplaces in which workers typically advertise an hourly rate for their services (like web programming, creative writing, and so on). Services like TaskRabbit focus on smaller jobs like running errands or house cleaning often at a fixed fee. While there have been many studies investigating the workers, incentives, and tasks in demand-driven marketplaces like AMT, there has been little scholarly research on these emerging supply-driven marketplaces.

Hence in this paper, we present a comprehensive data-driven measurement study of one prominent supply-driven marketplace – Fiverr, a rapidly growing global microtask marketplace and one of the 100-most popular sites in the US and one of the 200-most popular sites in the world (Fiv 2014). Fiverr participants can offer services and tasks (called "gigs") for a fixed base price of $5. Compared to freelancing sites in which workers advertise their talents for an hourly rate, Fiverr is distinguished by these fixed-price gigs. These gigs range over a variety of outcomes from video production (e.g., "I will create a simple, but VERY professional whiteboard video for $5") to odd jobs (e.g., "I will do odd and random jobs in London zones 1 and 2 for $5").

But who are these sellers? And what strategies do they adopt? What kind of gigs are being offered? And how are these gigs received by the community in terms of ratings and feedback? By systematically investigating Fiverr, we aim to provide a first comprehensive study of one popular emerging supply-driven crowd marketplace. Such an investigation can provide deeper insights into the evolution of crowd-based marketplaces and how to be successful on Fiverr. Concretely, we conduct a multi-part investigation into sellers, gigs, and and ratings of Fiverr:

- Sellers: First, we describe behaviors of sellers from three perspectives including their geographic distribution, longevity, and ratings. We identify the key features distinguishing "super sellers" from regular participants, which is important for understanding how sellers can best position themselves in these emerging marketplaces.
- Gigs: Second, we investigate what kinds of services are

Item	Count	
Sellers	5,941	
Only Seller	4,217	(71%)
Seller & Buyer	1,724	(29%)
Gigs	35,003	
Reviews	547,602	
Votes	592,556	
Positive Votes	562,214	(95%)
Negative Votes	30,342	(5%)
Purchases	2,082,905	

Table 1: Fiverr dataset

being offered on Fiverr and their distinguishing features including ratings and buyers' feedback.

- Ratings: Finally, since many gigs and sellers have little feedback, we develop a machine learning based approach for inferring the quality of gigs, which is especially important for the vast majority of gigs with little feedback.

Data and Setup

Since Fiverr does not provide an API for programmatically querying for information regarding users and gigs, we developed a crawler to collect both users and gigs for this study. We ran the crawler for seven months – from 19 April 2013 to 19 November 2013.

Dataset Summary. Overall, we collected data on 10,032 distinct users, of which 5,941 were still selling gigs with an active status. In total, these sellers have created 35,003 unique gigs, accounting for around 40% of all active gigs on Fiverr (Lee, Webb, and Ge 2014). The high-level results from the crawler are illustrated in Table 1. Based on our observation and investigation, some sellers not only are selling services, but also are buying gigs from others. Hence, about 29% of all active sellers we collected can be considered as being both a seller and buyer. Additionally, we found 547,602 reviews from customers on 35,003 gigs collected in this study. Each gig has an average of 15.6 reviews. All customers can vote on gigs they purchased: we find a total of 592,566 votes, of which 95% are positive. Since the user pages reveal purchase information, we find 2,082,905 purchases on 35,003 gigs, for an average of 59.5 purchases per gig. Since each gig is worth at minimum $5, we can see that sellers earn roughly $300 on average. As a lower bound, the overall market size of Fiverr from the sampled data is over $10M. Based on these (potentially dubious) statements from Fiverr and our analysis of the sampled dataset, this suggests that at least $700 million has been transacted on Fiverr since it launched. By taking a 20% commission from every gig, Fiverr is estimated to have revenues of $120 million at minimum.

The Investigation. In the following sections we present our four-part investigation. In Section 3, we examine who is selling gigs on Fiverr by analyzing the characteristics of sellers. In Section 4, we examine what distinguishes "super sellers" from other sellers. In Section 5, we explore the gigs sold on Fiverr. We then propose a machine learning approach to infer the quality of gigs in Section 6, before concluding with related work and some final thoughts.

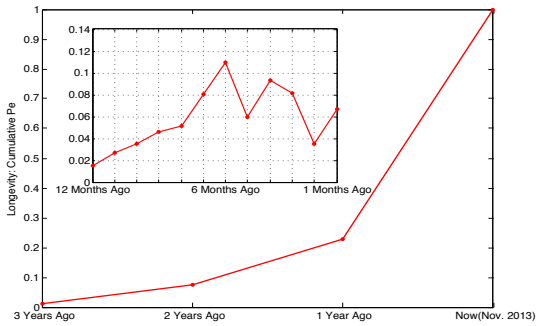

Figure 1: Longevity of sellers: the cumulative percentage of the longevity of active sellers over time.

Characterizing Sellers

In this section, we begin our study by examining the sellers on Fiverr. Where are they from? How active are they? How responsive? How many gigs do they manage? And how are they viewed by the community of buyers on Fiverr?

Geographic Origin of Sellers. We begin by inspecting the geographic origins for each of 5,941 sellers. Table 2 shows the top locations where sellers come from. Most sellers are from the United States, followed by India (IN), Great Britain (GB), and Canada (CA). We also observe a large number of sellers from south Asia including India (IN), Pakistan (PK), Sri Lanka (LK), and Bangladesh (BD). Similar findings have been made of participants in other crowdsourcing marketplaces like Amazon Mechanical Turk (AMT), Microworkers, and ShortTask (Ross et al. 2010).

Country	US	IN	GB	CA	Others
Count	2,467	540	356	200	2,378
Percentage	42%	9%	6%	3%	40%

Table 2: Where are sellers from? Ranked by gigs.

Longevity of Sellers. We next examine the longevity of sellers on Fiverr as shown in Figure 1. The number of active sellers on Fiverr has been increasing, especially for the past two years. Such an increase appears faster than linear as over 80% of the sampled sellers started to sell their services within the most recent one year. In order to better demonstrate the growth of active sellers on Fiverr, we zoom in on the distribution of longevity of sellers within the most recent 1 year, as shown in the sub-figure of Figure 1. As we can see, the number of sellers is still significantly increasing with a roughly linear growth rate.

Average Response Time. The average response time is commonly used to evaluate the quality of sellers for customers on Fiverr, which is defined as the time that it takes sellers to respond to inquiries from each customer. We plot

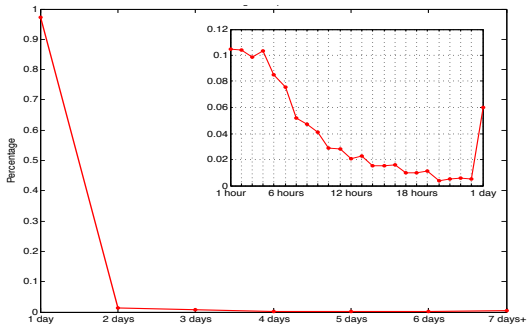

Figure 2: Average response time of sellers: the distribution of average response time of sellers associated with the time from 1 day to over 1 week.

the distribution of average response time of sellers in Figure 2. Most sellers respond to inquiries in less than 10 hours, which implies that most sellers on Fiverr are very active. On the other hand, as illustrated in the sub-figure of Figure 2, it is interesting that there is a jump at the average response time of one day. Two possibilities that can be used to explain this observation are: (i) One is that there might be some sellers who only consider selling services on Fiverr as a part-time job since they have to spend most of their time on an outside job rather than addressing inquiries from customers. Therefore, users only check the system once per day at some specific time, e.g. morning. (ii) Another possibility is that some sellers pre-set one day as the longest time to respond to customers' inquiries, and usually do not make any responses until the last minute (between 22 hours and 24 hours) within a day.

Interaction	Correlation	*p*-value
Longevity vs. # of Gigs	0.1627	0.0074
Longevity vs. # of Sales	0.1611	0.0085
# of Gigs vs. # of Sales	0.1438	0.0011

Table 3: Correlation coefficients between longevity, number of gigs, and number of sales.

Workload of Sellers. On average, each seller is selling 5.89 gigs on Fiverr, and few sellers possess a very large number of gigs as the distribution of number of gigs per seller is exponential. For further analysis, we investigate the correlations between the longevity of a seller, the number of gigs a seller created, and the number of sales a seller conducted in order to understand how the workload of a seller correlates with longevity and profit (number of sales). The correlation coefficients can be found in Table 3. As it can be seen, interactions between each two of three variables are very weak with the correlation coefficients less than 0.17.

Bias of Sales. Are gigs for a seller equally popular? Or are there some breakout hits? To address this, we adopt an entropy-like measure σ to measure the bias of sales for a seller as:

$$\sigma = -\sum_{i=1}^{N} \frac{s_i}{\sum_{i=1}^{N} s_i} \log\left(\frac{s_i + \varepsilon}{\sum_{i=1}^{N} s_i}\right)$$

where N is the number of gigs a seller holds, s_i is the number of sales for the gig i, and ε is a constant (set to 0.1 to overcome sellers with zero sales for some gigs). A larger bias indicates these seller relies on a few gigs for most of their sales. In Figure 3, the x-axis is the number of gigs for a seller, and the y-axis is the log number of sales for a seller. Each circle represents a seller, in which the diameter stands for the bias; a larger diameter indicates a larger bias. As can be seen in Figure 3, the bias decreases with the number of gigs. However, for focused sellers with fewer than five gigs, we see a strong bias indicating dependence on a few breakout hits. On the other hand, we also observe that most top sellers with a large number of sales do not hold significant biases, implying that gigs produced by top sellers are more evenly popular.

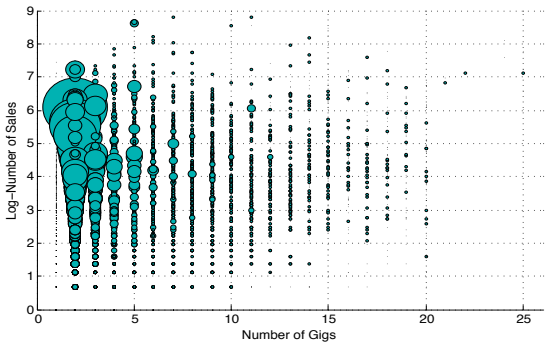

Figure 3: Bias of number of sales in gigs for a seller.

Seller Ratings

Fiverr has its own rating system used to evaluate and measure the quality of sellers. Similar to eBay, each gig on Fiverr can be rated by customers with a binary rating of a positive vote (*thumbs up*) or a negative vote (*thumbs down*). The rating of a seller is calculated based upon the ratio of positive votes which is then transformed to the ratings on a scale of 0 to 5 stars with an interval of 0.5 (e.g., a seller with 100% positive votes is mapped to a 5 star rating; a seller with a ratio of positive votes between 99% and 95% is mapped to 4.5 stars; a seller with a ratio of positive votes between 95% and 90% is mapped to 4 stars). It should be noted that the default rating of a seller is 0 stars if no customer purchases the service or customers do not rate the seller after the purchase. Figure 4 illustrates the distribution of ratings on sampled sellers from Fiverr. Over 80% of all sellers are rated by customers with 4 stars and above. In contrast, about 5% of all sellers receive low ratings between 0.5 and 3.5 star, and another 10% of all sellers have a 0 star rating. On inspection, we found that most of these 0-star sellers are new sellers (with a longevity of less than one month) and so have yet to accumulate purchases or ratings-based feedback.

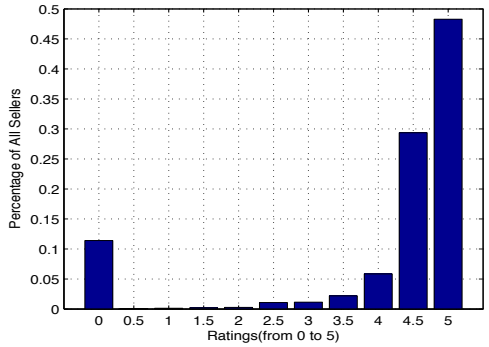

Figure 4: Distribution of seller ratings.

Super Sellers

Based upon our investigation, close to 60% of sellers on Fiverr earn more than $100 for their efforts. The remaining 40% of sellers earn between $5 and $100. However, there is an elite of Fiverr – the top 2% of sellers earn more than $10,000. For them, selling gigs can be considered as a full-time job. We list the top-5 sellers in our dataset in Table 4. We see that the top-ranked seller – *crorkservice* – has sold 131,338 gigs and earned at least $656,690 over two years. What makes these "super sellers" successful on Fiverr? Are there particular attributes of these sellers that drive their success? And perhaps these attributes can be adopted by other less successful sellers.

User Name	Gigs	Transactions	Earned(Min.)	Gig Category
crorkservice	30	131,338	656,690	OM
dino_stark	3	61,048	305,240	OM
alanletsgo	29	36,728	183,640	OM, Bu & Ad
bestoftwitter	7	26,525	132,625	OM & Ad
amitbt	9	18,574	92,870	OM & WT

Table 4: The Top-5 Sellers. Note: *OM* stands for Online Marketing, *Bu* stands for Business, *Ad* stands for Advertising, *WT* stands for Writing&Translation.

Super Sellers vs. Regular Sellers

To answer these questions, we examine the behaviors of both super and regular sellers from different perspectives. In this study, we propose a *UserScore* denoted as ϕ to measure the *selling efficiency* of all users we collected on Fiverr. For a user i, we let o_i be the number of sold gigs (transactions), and d_i be the number of days since this user registered on Fiverr. The rate of transaction γ_i for user i can be defined as $\gamma_i = \frac{o_i}{d_i}$. Then the *UserScore* ϕ can be defined as $\phi_i = log(1 + o_i) * \gamma_i$.

Through this *UserScore*, the rate of transaction is taken into account since a seller may have sold many gigs simply because he has been in business for a long time, not because he is particularly efficient at selling gigs. Due to the difference between magnitudes of the number of sold gigs o_i and the rate of transaction γ_i, we use the logarithm of the number of sold gigs to balance its impact on *UserScore*. Based on this selling efficiency, we define a *Super Seller* as a user in the top 5% of all users whose *UserScore* is larger than a threshold. We adopt a threshold of 11.09 that is determined by the 95% percentile of *UserScore* for all collected users. In total, we identify 300 super sellers out of 5,941 sellers. How do these 300 super sellers differ from the other sellers?

Longevity. In Figures 5 and 6, we can see that most popular gigs and super sellers have already been on Fiverr for a relatively long time comparing with regular sellers and normal gigs. This indicates that it is challenging for sellers to become popular on Fiverr; the entrenched super sellers appear to mainly be early adopters.

Responsiveness. We consider two perspectives on how responsive a seller is:

Average Response Time. The first is the average response time for super sellers and for regular sellers, as shown in Figure 7. Surprisingly, we find that super sellers do not respond to customers' inquiries within a short period (less than 4 hours) as compared to regular sellers, but instead give responses with an average time of one day. This may be since super sellers already have many orders in the queue, and so do not have sufficient time to respond to customers in time. However, simply optimizing on response time is not necessarily a good predictor of super seller status.

Ratio of Leaving Feedback to Customer Reviews. An alternate measure of seller responsiveness is how often they reply to customer reviews (by leaving feedback). In Figure 8, we show the ratio of leaving feedback for customer reviews for super and regular sellers. As can be seen, super sellers perform more actively than regular sellers with a much higher ratio of leaving feedback. This indicates that super sellers are more active and professional in engaging with their customers.

Presentation. What about how super sellers present their gigs? We consider two factors:

Length of Title and Description. As a crude approximation for gig quality, Figure 9 highlights the length of a gig's description for both super sellers and regular ones (a qualitatively similar relationship holds for length of titles as well). Note that "CDF" along the y-axis stands for the cumulative percentage in these figures. We can observe that super sellers typically employ longer titles and descriptions which might embed more detailed information in terms of their gigs in order to make buyers better understand what they are buying and what their options are.

Ratio of Leaving Work Samples in Reviews. Sellers can additionally provide work samples associated with the reviews left by their customers. In this way, a seller can give more concrete evidence of the quality of the gig. Work samples are especially prominent for photo and video related gigs. We observe that 46% of super sellers do leave work samples, and all other sellers do so at a rate of 40%. So mere evidence of a work sample does not appear to be strongly indicative of super sellers; presumably gig quality (as of the work sample) is a more important factor.

Figure 5: Longevity of gigs

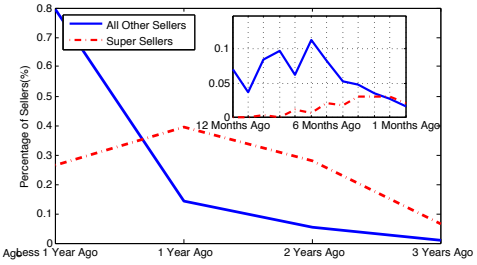

Figure 6: Longevity of users

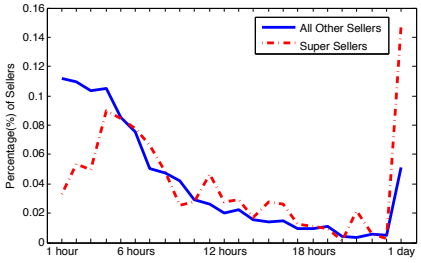

Figure 7: Average response time

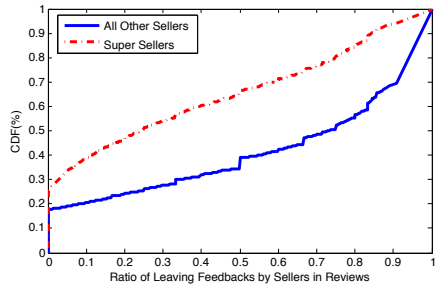

Figure 8: Ratio of leaving feedback

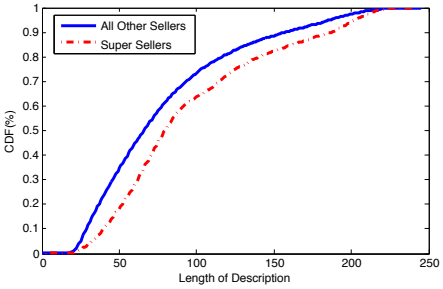

Figure 9: Length of description

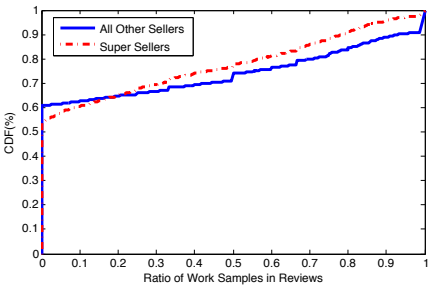

Figure 10: Ratio of leaving work samples

Discoverability. How easy it is to discover a seller's gigs?

Number of Tags. Seller can tag their gigs as an additional opportunity for discoverability through tag-based search and browsing. Overall, the number of tags used in gigs for super sellers is larger than that of tags for regular sellers. Super sellers associate at least three tags in their gigs, and about 70% of them label no more than 5 tags in gigs while less than 2% of sellers put more than 13 tags.

Featured Sellers and Gigs. An alternate method of highlighting gigs is to acquire a *featured* label from Fiverr. These *featured gigs* are considered of high-quality by Fiverr (though the exact determinants of what gigs are featured is not publicly available) and are often featured on the front-page of Fiverr. We find that super sellers are much more likely to offer featured gigs – around 35% of super sellers have at least one featured gig compared to only 2.1% of all other sellers.

Strategies of Super Sellers. Based on these findings, we can see that super sellers do engage in distinct behaviors relative to normal sellers: they leave work samples, focus on title and descriptions, and increase exposure via tags. On the other hand, according to our investigations, we can see that 67.3% of super sellers hold gigs related to the category of Online Marketing, e.g. adding more likes on Facebook, tweeting to tons of people, promoting twitters, etc. We suspect that most super sellers could be likely crowdturfers or at least less organic users.

Identifying Future Super Sellers

Given these observations, we can see that in some aspects super sellers do behave differently from regular sellers. Is it possible to distinguish a super seller from a regular seller based only on the seller's profile and the gigs it advertises? That is, are there intrinsic characteristics of how super sellers present themselves on Fiverr that are correlated with their success? Identifying such presentational cues is important for identifying hot new sellers on Fiverr and providing enhanced search and browsing over gigs and sellers.

Concretely, we investigate this question by building several machine-learning binary classification models over the super sellers and regular sellers. We consider features related to the seller's gigs and profile only since we only have these information when sellers initially create their profiles and gigs. We exclude all behavioral and sales-related information like the number of sales, average response time, the number of ratings, and so on. Our features consist of country of the seller, ratio of gigs with video introductions, average length of title, average number of sentences in the description, average number of words per sentence in the description, average length of description, average number of tag used in gig, and average entropy of tag used in the gig. Additionally, we also include bag-of-words features such as "professional", "extra", "video", "check", and "script". In order to overcome the imbalance issue in our data, the resampling method of Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. 2002) is applied to balance the number of samples in the training data set. We tried several different machine learning models such as Naive Bayes, Random Forest, Random Tree, SVM, and Logistic Regression with 10-fold cross validation. Overall, Logistic Regression performs the best, but achieves only a 34.1% precision and 9.2% recall, indicating the difficulty of identifying super sellers only based on the characteristics of their profiles and gigs. It should be noticed here that precision and recall employed here are only for super sellers, not for all sellers

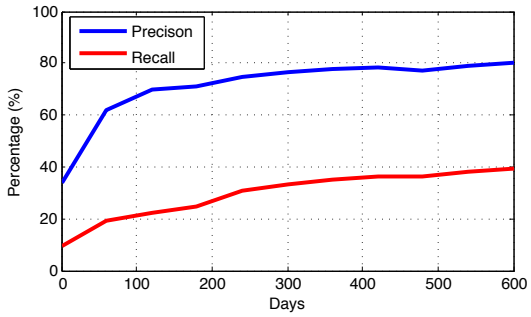

Figure 11: Performance of logistic regression model with initial and snapshot features.

including regular ones.

But perhaps there are clues in how a super seller operates in its initial time on Fiverr that can provide evidence of future growth. Hence, from time t_0 (when a seller initially creates their profiles and initial gigs) to time t_i (the ith snapshot of sellers' profiles and gigs), we extract three additional features at each snapshot: the number of reviews associated with the seller, the ratio of leaving feedback to customer reviews (suggesting seller responsiveness), and the ratio of leaving work samples (suggesting a maturity of the seller). With these additional features, we find a much improved precision and recall as shown in Figure 11. After a user creates their profile and gigs to about 3 months, we see a very sharp increase in precision and recall, indicating that these temporal features provide additional evidence for identifying future super sellers. Initial activity is strongly indicative of later success.

Characterizing Gigs

In this section, we turn our attention from sellers to gigs.

The Types of Gigs. There are twelve main categories on Fiverr ranging from promoting products to designing various kinds of goods, each of which contains several subcategories with a total of 110 sub-categories. In Table 5, we summarized the number of gigs in each of the twelve main categories on Fiverr with the number of gigs in that category and its corresponding percentage. As we can see, most gigs are created in the categories of *Graphics&Design*, *Online Marketing*, *Writing&Translation*, *Video&Animation*, and *Programming&Tech*. Nearly 70% of gigs on Fiverr are created in these five main categories.

Gig Ratings. Each gig on Fiverr can be rated by customers after the purchase. There are two different ratings: a *positive rating* and a *rating star*. A positive rating is calculated based on how many votes from customers are positive (*thumbs up*), expressed as a percentage. The rating star is then derived from the positive rating on a scale of 0 to 5 stars. Figure 12 illustrates the distribution of ratings of gigs. Apart from gigs without ratings, most gigs come with very high ratings, as we observed for sellers. This highly skewed rating system is similar to eBay's rating system; Resnick and Zeckhauser

Category	Gigs	Pct
Graphics&Design	7,856	22.4%
Online Marketing	4,861	13.9%
Writing&Translation	3,643	10.4%
Video&Animation	3,216	9.2%
Programming&Tech	2,970	8.5%
Life Style	2,155	6.2%
Advertising	1,958	5.6%
Music&Audio	1,860	5.3%
Fun&Bizarre	1,781	5.1%
Business	1,766	5.0%
Gifts	1,662	4.7%
Other	1,275	3.6%

Table 5: Distribution of Gigs per Category

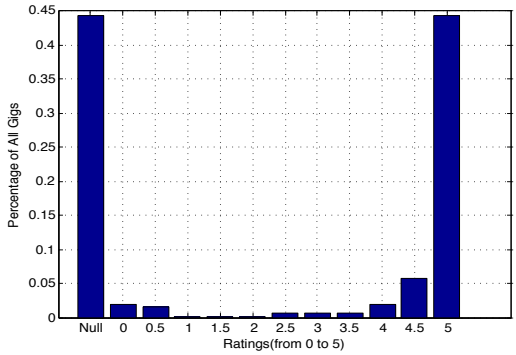

Figure 12: Distribution of gig ratings.

(Resnick and Zeckhauser 2002) observed that 99% of feedback on eBay is positive.

Feedback. We explore reviews left by customers in gigs through investigating the frequency of words. In order to understand the behaviors of customers, it is necessary to examine how feedback from customers are distributed in gigs. The scatter plot in Figure 13 (right) shows the relationship between the number of customers and the number of votes (reviews) for all gigs of each seller. It can be observed that not all customers vote for gigs they bought and 92% of points are located in the grid area defined by 5000 customers and 2000 votes. In addition, we plot the curve for the boundary of ratio between the number of customers and the number of votes in corresponding gigs. The curve is given by the linear function $y = 0.715x$ where y is the number of votes and x is the number of customers, which means that up to 71.5% of customers on Fiverr leave reviews for gigs they purchased. Meanwhile, we also examine the feedback from sellers corresponding to customers' reviews. Figure 13 (left) illustrates that there is a trend for sellers to respond more to reviews from customers though the underlying skew is not very clear due to the use of log for the frequency. As we can see, all points in Figure 13 (left) are distributed towards the high percentage of leaving feedback from sellers to customer reviews, indicating that some sellers on Fiverr are trying to respond to reviews more actively. Looking back to Figure 8, where we observe that super sellers are more active in leav-

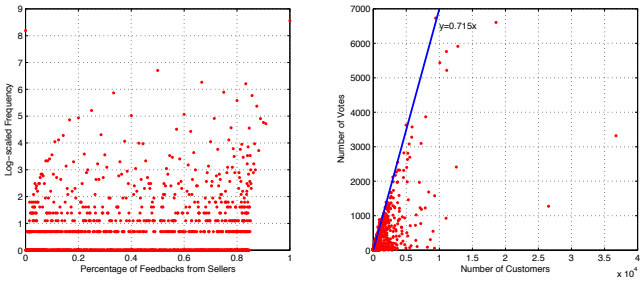

Figure 13: Distribution of the number of feedbacks from sellers corresponding to customers' reviews in gigs (left). Relationship between the number of customers and the number of votes in gigs (right).

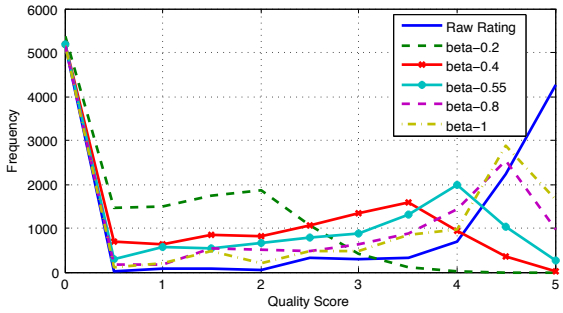

Figure 14: Sensitivity analysis for the parameter β in terms of the quality of the gig.

ing feedback to customer reviews. These two findings are inter-related. Therefore, it can be seen that actively leaving feedback to customer reviews could be one of the important factors towards the success of sellers and their gigs.

Predicting Gig Quality

So far we have measured and analyzed characteristics of gigs and sellers on Fiverr, and seen that there are key differences in how super sellers engage with the service compared to regular sellers. Next, we turn our attention in this study to the popularity of gigs created by sellers and look to determine salient features which statistically contribute to the popularity of gigs. Can we identify key factors that contribute to a gig's popularity?

Estimating Gig Quality

As we have seen, the rating of the gig on Fiverr is heavily biased as most gigs are positively rated by customers with high ratings. In addition, a 5-star gig with 10 purchases, a 5-star gig with 100 purchases, and a 5-star gig with 1,000 purchases are all equally treated in the current reputation system on Fiverr. Hence, we first propose to estimate gig quality by considering both star rating and number of purchases. Our intuition is to reshape the distribution of the ratings of a gig on Fiverr by employing the sigmoid function which has an "S" shape bounded in the range from 0 to 1: $f(t) = \frac{1}{1+e^{-t}}$.

The total number of purchases for a seller could be an important indicator for the quality of the gig since more sales can be equally interpreted as the popularity of the gig. We assume that the quality of the gig without sales is 0. Taking this into account, the quality Q of the gig can be defined as a function of the total number of purchases and the corresponding rating of the gig as follows:

$$Q(n, r) = \left(\frac{2}{1 + e^{-\beta \log(n+1)}} - 1 \right) * r$$

where n is the number of sales for a seller, r is the rating of the gig and β is used to adjust the shape of the sigmoid function. Since the total number of sales for a seller never be negative, we apply a simple transformation to map the sigmoid function $f(t), t \geq 0$ from the original range 0.5 to 1 to the range 0 to 1 as the quality should be equal to 0 when there are no purchases. This new quality score ranges from 0 to 5, just as in the existing rating system.

In order to determine the value of the parameter β in the quality formula, we are inspired by the way of determining the speed limit in transportation engineering. The speed limit is commonly set to the 85th percentile operating speed so that no more than 15% of traffic is excess of this value in order to make majority of drivers in the safe speed (Spe 2014). Hence, we first find the number of purchases at the 15th percentile for all 5-star gigs, and then calculate the value of the parameter β corresponding to the mapping from the rating of 5 to the quality of 4. Finally, the best value of the parameter β is equal to 0.55.

After transferring to the quality schema we proposed, the distribution of the rating of gigs is re-shaped with the parameter β of 0.55 as shown in Figure 14. As we can see in Figure 14, the quality of gigs is pushed towards the range of the middle and upper ratings. For instance, a 5 star gig with 10 sales will be projected to 1.5 star in the quality scale; a 5 star gig with 100 sales will be projected to 5 star in the quality scale; a 4 star with 100 sales will be projected to 3.5 star in the quality scale; a 4 star with 1,000 sales will be projected to 4 star in the quality scale.

Factors Impacting Gig Quality

Given this view of gig quality, we next turn to the challenge of predicting what gigs will be well received and which will not. We created a wide variety of features which are User Related (***UR***) and Gig Related (***GR***). User Related features contain ones extracted from profiles of users on Fiverr such as the length of user ID, the number of sales since joining Fiverr, and so on. Gig Related features include information extracted from the gig webpage such as the number of votes, the number of reviews, and so on. For the content of gigs, we also consider the Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) extracted from gig titles and descriptions. Though the dimensionality of both BOW and TF-IDF is large, we consider them as two features only. In total, there are 32 features.

To understand which features could contribute to the popularity of gigs, we measure the chi-square statistic of the features, which is widely used to test the lack of independence of two events. The larger the chi-square value is, the

Feature Selection for Ratings			
Features	***Chi-square***	***Features***	***Chi-square***
number of positive votes	41832.029	is featured by Fiverr	416.977
ratio of leaving reviews for customers	34188.594	has video in the introduction	892.918
average number of words in review	31060.049	length of the user ID	3144.911
ratio of replying buyers' reviews by the seller	28932.912	average length of sentences in the description	3281.121
average number of words in replies by the seller	28612.703	number of words in the title	3456.817
number of sales per seller	21442.459	number of gigs created by the user	3470.393
entropy of geographic distribution of buyers	10028.213	number of sentence in the description	3708.199
locality	9371.381	number of tags used in the gig	3846.641
ratio of work samples showing in reviews	9272.681	entropy of tags used in the gig	3878.399
average response time by the user	8875.246	number of words in the description	4091.190
Feature Selection for Quality Scores			
Features	***Chi-square***	***Features***	***Chi-square***
number of sales per seller	77727.404	is featured by Fiverr	301.807
locality	46020.506	is featured by Fiverr	1093.815
number of positive votes	43751.142	has video in the introduction	3235.276
rate of sales	43146.495	length of the user ID	3309.503
ratio of leaving reviews for customers	39740.427	average length of sentences in the description	3802.834
average number of words in review	35270.833	number of sentence in the description	4013.094
level marked by Fiverr	33816.894	number of words in the title	4225.884
ratio of replying buyers' reviews by the seller	30343.432	number of words in the description	4225.964
ratio of work samples showing in reviews	29570.923	number of tags used in the gig	4268.227
average number of words in replies by the seller	25606.661	entropy of tags used in the gig	4589.807

Table 6: Top-10 features with the most contributions and bottom-10 features with the least contributions for ratings and quality scores.

higher contribution to the rating of the gig the corresponding feature could have. Through 10-fold cross validation, we select the top-10 features with the most contributions and the bottom-10 features with the least contributions for both the original (raw) ratings and the proposed quality score, respectively, as shown in Table 6. As we can see, it is interesting that, there are 7 common features between the top 10 features for ratings and quality scores, which are *number of positive votes*, *locality*, *number of sales*, *ratio of replying buyers' reviews by the seller*, *ratio of leaving reviews for customers*, *average number of words in replies by the seller*, *ratio of work samples showing in reviews*. We observe that the review in the gig plays an important role to promote the popularity of the gig. It seem that the gig will be popular while buyers more actively leave feedback or sellers more actively reply to customers' reviews and provide more samples. Being active and responsive is crucial for selling services on Fiverr.

On the other hand, we also observe some features that have little impact on the popularity of gigs from the bottom-10 features as demonstrated in Table 6. First of all, it is surprising that the bottom-10 features for ratings and quality scores are exactly the same. It can be seen that tags do not contribute to the high rating of gigs. After all, the main goal of tags labeled by the user in the gig is to help visitors on Fiverr to better retrieve gigs corresponding to their queries. Moreover, we can also observe that the length of title, the number of gigs the user created, and the average length of sentences in the description might not promote the rating of the gig with sufficient contribution. However, it is surprising that whether the gig is associated with a video introduction does not significantly affect the rating of the gig, as well as whether the gig is featured by Fiverr though featured gig manually selected by Fiverr agents means more exposure.

Predicting Gig Quality

Finally, we turn to the challenge of predicting gig quality based on the features identified in the previous sections. We consider both the original (raw) rating of a gig, as well as the proposed gig quality. Predicting gig quality is helpful for identifying potential super sellers and their gigs, as well as providing insight to sellers to better shape their gigs. Of the features identified in the previous section, we focus solely on those that are under the control of the seller: *ratio of replying buyers' reviews by the seller*, *average number of words in replies by the seller*, *ratio of work samples showing in reviews*, *length of description*, *length of title*, *number of tags used in the gig*, *entropy of tags used in the gig*, *average fabrication time of the gig*, *average number of words per sentence in description* and *average response time by the user*. It should be noticed that when predicting the gig quality, in order to treat temporal information appropriately all features mentioned above for a gig are calculated only from other gigs which are produced by the same seller before the generation of this gig. All features related to reviews are only considered as ones generated before the generation of this gig. For the feature of average response time, it can be considered as a constant in the prediction since it rarely has significant changes based upon our observations.

In order to find the best model for the prediction of the population of the gig, we test six different machine learning algorithms such as Naive Bayes, Random Forest, Random Tree, SVM, Logistic Regression and Linear Regression as the baseline via the WEKA machine learning toolkit (Hall et

al. 2009). The prediction of gig quality can be considered as a classification problem for 11 classes from 5 to 0 stars with an interval of 0.5. Note that the results from Linear Regression will be rounded to the closest one of the 11 classes. The standard 10-fold cross-validation is employed, which splits our data into 10 sub-samples. For a given classifier, each of 10 sub-samples will be considered as the test data, and the rest of the 9 sub-samples are used to train the classifier. The 10 classification results are macro-averaged as the final result for a given classifier in order to avoid results being dominated by the performance on the larger class. Additionally, we compute precision, recall and F-measure as metrics to evaluate each classifier.

The result of the six machine learning algorithms we tested is shown in Tables 7 and 8. First, the tree-based machine learning algorithms outperform the rest. In particular, random forest produced the highest accuracy of 97.3% for the rating and 98.3% for the quality. Moreover, we can see that in general, the accuracy of the classifier decreases with 10 selected features comparing with one with all features, which illustrates that the rest of features except for 10 selected ones are valuable and capable of providing useful information in terms of predicting the popularity of the gig. In all, based on the results demonstrated above, we can successfully predict the popularity of the gig by the classifiers with features we generated for both original (raw) ratings and for the proposed quality score which has better performance in the prediction.

Classifier	**All Features**		**Selected Features**	
	Accuracy	*F1*	*Accuracy*	*F1*
Random Forest	97.3%	0.973	79.4%	0.804
Random Tree	91.9%	0.918	74.0%	0.740
Navie Bayes	75.8%	0.703	77.1%	0.778
Linear Regression	57.8%	0.515	42.6%	0.428
SVM	87.2%	0.874	74.6%	0.750
Logistic Regression	83.3%	0.831	72.1%	0.722

Table 7: Predicting Original (Raw) Ratings.

Classifier	**All Features**		**Selected Features**	
	Accuracy	*F1*	*Accuracy*	*F1*
Random Forest	98.3%	0.989	82.8%	0.833
Random Tree	93.7%	0.935	76.1%	0.759
Navie Bayes	80.5%	0.809	67.1%	0.658
Linear Regression	62.2%	0.632	53.9%	0.531
SVM	91.3%	0.912	78.6%	0.787
Logistic Regression	87.3%	0.871	76.9%	0.772

Table 8: Predicting Proposed Quality Score.

Related Work

There is a vast research literature on measuring and understanding new web, social, and crowdsourcing systems. In this section, we focus on existing research with respect to measurement analysis of these systems, as well as the problem of ratings prediction.

In terms of crowdsourcing marketplaces, Ross et al. (Ross et al. 2010) presented a demographic analysis for microworkers over a 20-month period on Amazon Mechanical Turk (AMT). Similarly, Ipeirotis (Ipeirotis 2010) analyzed AMT with a comprehensive measurement study. In addition, there are many related studies of these emerging crowdsourcing platforms (Berinsky, Huber, and Lenz 2012; Mason and Suri 2012; Anderson et al. 2012; Liu et al. 2014; Kosinski et al. 2012).

In a separate direction, Christin (Christin 2013) produced a comprehensive measurement analysis of Silk Road which is an anonymous online market place for exchanging "black market" goods. Wang et al. (Wang et al. 2012) analyzed the two largest crowdturfing sites in China. Teodoro et al. (Teodoro et al. 2014) studied the motivations of the on-demand mobile workforce services through examining the socio-technical factors. Lee et al. (Lee, Webb, and Ge 2014) studied Fiverr from the perspective of detecting crowdturfing.

In terms of online social networks, there have been many comprehensive studies. Mislove et al. (Mislove et al. 2007) measured YouTube, Flickr, LiveJournal, and Orkut. The work produced by Kwak (Kwak et al. 2010) studied characteristics and properties of Twitter with a comprehensive measurement and analysis. Gilbert et al. (Gilbert et al. 2013) conducted a global statistical analysis of the Pinterest network. Wang and his colleagues (Wang et al. 2013) studied Quora which is a popular question and answer (Q&A) site. Additionally, researchers have also examined Facebook (Lampe, Ellison, and Steinfield 2007), Google+ (Gong et al. 2012), and Foursquare (Vasconcelos et al. 2012).

Predicting ratings online is one of the core problems in recommendation systems. The most popular method in the prediction of ratings is collaborative filtering developed by Resnick et al. (Resnick and Varian 1997). Goldberg with his colleagues (Goldberg and Zhu 2006) also developed a graph-based semi-supervised learning algorithm in order to successfully predict ratings. Koenigstein et al. (Koenigstein, Dror, and Koren 2011) modeled the shift in users' interests to improve the prediction accuracy of ratings on Yahoo! Music. Moghaddam et al. (Moghaddam, Jamali, and Ester 2012) studied the prediction of review helpfulness by a probabilistic graphical model with a high-dimensional tensor factorization.

Conclusion

In this paper, we have performed the first comprehensive measurement analysis of the supply-driven marketplace Fiverr. Compared to existing ecommerce sites like Zappos and eBay, as well as crowdsourcing marketplaces like Amazon Mechanical Turk and CrowdFlower, Fiverr is distinguished by its task-orientation and seller-driven offerings. Through analysis of 30,000 gigs and 10,000 users over a six-month period, we have initiated a three-part investigation into the gigs, sellers, and ratings of Fiverr. We have examined these gigs and their distinguishing features, and seen how "super sellers" distinguish themselves from regular participants, which is especially important for understanding how sellers can best position themselves in these emerging marketplaces. We ended with a dive into popularity of gigs

on Fiverr, where we studied what factors are important for inferring the quality of gigs. Moving forward, we are interested to revisit Fiverr gigs and sellers to study the dynamics of popularity as the service continues to evolve. We are also eager to explore how to estimate a gig's ultimate rating (and popularity) at the earliest moments of its listing on Fiverr.

References

Anderson, A.; Huttenlocher, D.; Kleinberg, J.; and Leskovec, J. 2012. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD*.

Berinsky, A. J.; Huber, G. A.; and Lenz, G. S. 2012. Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. *Political Analysis*.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*.

Christin, N. 2013. Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22nd WWW*.

2014. *Fiverr.com Site Information*. `http://www.alexa.com/siteinfo/fiverr.com`.

Gao, H.; Barbier, G.; Goolsby, R.; and Zeng, D. 2011. Harnessing the crowdsourcing power of social media for disaster relief. Technical report, DTIC Document.

Gilbert, E.; Bakhshi, S.; Chang, S.; and Terveen, L. 2013. I need to try this?: a statistical overview of pinterest. In *Proceedings of the SIGCHI Conference*.

Goldberg, A. B., and Zhu, X. 2006. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*.

Gong, N. Z.; Xu, W.; Huang, L.; Mittal, P.; Stefanov, E.; Sekar, V.; and Song, D. 2012. Evolution of social-attribute networks: measurements, modeling, and implications using google+. In *Proceedings of the 2012 ACM IM*.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: an update. *ACM SIGKDD*.

Ipeirotis, P. G. 2010. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*.

Khatib, F.; DiMaio, F.; Cooper, S.; Kazmierczyk, M.; Gilski, M.; Krzywda, S.; Zabranska, H.; Pichova, I.; Thompson, J.; Popović, Z.; et al. 2011. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*.

Koenigstein, N.; Dror, G.; and Koren, Y. 2011. Yahoo! music recommendations: modeling music ratings with temporal dynamics and item taxonomy. In *Proceedings of the 5th ACM conference on Recommender systems*.

Kosinski, M.; Bachrach, Y.; Kasneci, G.; Van-Gael, J.; and Graepel, T. 2012. Crowd iq: Measuring the intelligence of crowdsourcing platforms. In *Proceedings of the 3rd ACM WebSci*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th WWW*.

Lampe, C. A.; Ellison, N.; and Steinfield, C. 2007. A familiar face (book): profile elements as signals in an online social network. In *Proceedings of the SIGCHI conference*.

Lee, K.; Webb, S.; and Ge, H. 2014. The dark side of micro-task marketplaces: Characterizing fiverr and automatically detecting crowdturfing. In *Proceedings of ICWSM*.

Liu, T. X.; Yang, J.; Adamic, L. A.; and Chen, Y. 2014. Crowdsourcing with all-pay auctions: a field experiment on taskcn. *Management Science*.

Mason, W., and Suri, S. 2012. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM*.

Moghaddam, S.; Jamali, M.; and Ester, M. 2012. Etf: extended tensor factorization model for personalizing prediction of review helpfulness. In *Proceedings of the 5th ACM WSDM*.

Resnick, P., and Varian, H. R. 1997. Recommender systems. *Communications of the ACM*.

Resnick, P., and Zeckhauser, R. 2002. *Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system*.

Ross, J.; Irani, L.; Silberman, M.; Zaldivar, A.; and Tomlinson, B. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*.

2014. *Speed Limit*. `http://http://en.wikipedia.org/wiki/Speed_limit`.

Teodoro, R.; Ozturk, P.; Naaman, M.; Mason, W.; and Lindqvist, J. 2014. The motivations and experiences of the on-demand mobile workforce. In *Proceedings of the 17th ACM CSCW*.

Vasconcelos, M. A.; Ricci, S.; Almeida, J.; Benevenuto, F.; and Almeida, V. 2012. Tips, dones and todos: uncovering user profiles in foursquare. In *Proceedings of the 5th ACM WSDM*.

Wang, G.; Wilson, C.; Zhao, X.; Zhu, Y.; Mohanlal, M.; Zheng, H.; and Zhao, B. Y. 2012. Serf and turf: crowdturfing for fun and profit. In *Proceedings of the 21st WWW*. ACM.

Wang, G.; Gill, K.; Mohanlal, M.; Zheng, H.; and Zhao, B. Y. 2013. Wisdom in the social crowd: an analysis of quora. In *Proceedings of the 22nd WWW*.

Zaidan, O. F., and Callison-Burch, C. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*.