On the Team Selection Problem

Milan Vojnović and Se-Young Yun[1]

February 2016

Technical Report
MSR-TR-2016-7

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
http://www.research.microsoft.com

[1]Milan Vojnović (milanv@microsoft.com) and Se-Young Yun (t-seyun@microsoft.com) are with Microsoft Research, Cambridge, United Kingdom.

Abstract – We consider a team selection problem that requires to hire a team of individuals that maximizes a profit function defined as difference of the utility of production and the cost of hiring. We show that for any monotone submodular utility of production and any increasing cost function of the team size with increasing marginal costs, a natural greedy algorithm guarantees a $1 - \log(a)/(a-1)$–approximation when $a \leq e$ and a $1 - a/e(a-1)$–approximation when $a \geq e$, where a is the ratio of the utility of production and the hiring cost of a profit-maximizing team selection. We also consider the class of test-score algorithms for maximizing a utility of production subject to a cardinality constraint, where the goal is to hire a team of given size based on greedy choices using individual test scores. We show that the existence of test scores that guarantee a constant-factor approximation is equivalent to the existence of special type of test scores – so called replication test scores. A set of sufficient conditions is identified that implies the existence of replication test scores that guarantee a constant-factor approximation. These sufficient conditions are shown to hold for a large number of classic models of team production, including a monotone concave function of total production, best-shot, and constant elasticity of substitution production function. We also present some results on the performance of different kinds of test scores for different models of team production, and report empirical results using data from a popular online labour platform for software development.

Keywords: Online Services, Online Marketplaces, Social Sciences, Industrial Organization, Team Performance, Submodular Functions

1 Introduction

The performance of teams has been one the most central topics of study in areas such as organization science, industrial organization, theory of firms, management sciences, social psychology, and has recently received much attention in the context of online labour platforms; for example, in the context of competition-based crowdsourcing platforms where solutions to tasks are derived by aggregating inputs from multiple online workers, or in the context of paid-labour online marketplaces for matchmaking between tasks and independent contractors.

A standard model of team performance defines actual productivity of a team of individuals as difference of a *potential productivity* and a *process loss* [Ste72]. Here, the potential productivity is the highest level of performance attainable by a team and the process losses arise due to various factors including *motivational loss* and *coordination loss*. Motivational losses can arise if individual objectives are not aligned with that of the team objective. Coordination losses occur when individuals fail to organize their efforts optimally as a team. The decrease in individual effort that occurs when an individual works within a group is often referred to as *social loafing* in psychology, e.g., [KM86], [LWH79], [Mue12] and [SMF12]. One key question studied in literature is about optimal team size [LN81]. A review of the literature on team performance in organizations is provided in [GS92]. Several books provide valuable insights on the team performance, e.g., [Pag07] and [KM15].

A key issue is the problem of team selection for solving a given task. In this paper, we study a formulation of a *team selection problem* defined as follows. Suppose that given a set of individuals $N = \{1, 2, \ldots, n\}$, the actual productivity of a team of individuals $S \subseteq N$ is given by a function $p(S) = u(S) - c(S)$, where $u(S)$ denotes the utility of production and $c(S)$ denotes a cost function. We may interpret $p(S)$ as a profit to a principal realized by hiring a team S, defined as difference of the utility of production and the team hiring cost. The team selection problem asks to select a set of individuals S^* that maximizes the profit function, i.e. finding a set of individuals S^* such that

$p(S^*) \geq p(S)$, for every $S \subseteq N$. We will sometimes also refer to this problem as a *profit maximization problem*. We shall consider instances where the cost function is an increasing function of the team size; this accommodates many interesting cases, e.g., the case of a *linear cost* where a constant marginal cost is incurred per each team member, or the case of a *cardinality constraint* where the cost of hiring any given team is equal to zero as long as the team size is smaller than or equal to given cardinality constraint, and is infinite otherwise.

We shall consider the class of utility functions that are non-negative, monotone submodular set functions, and the class of cost functions that are functions of the team size with increasing marginal costs. The class of non-negative, monotone submodular utility functions accommodates diminishing-returns production systems, where the marginal gain of increasing a team size diminishes with the team size. For some of our results, we shall consider a utility of production according to a *stochastic model of team production*, defined as the expected value of a given mapping of individual performances to a team performance output, where the individual performances are independent random variables with given cumulative distribution functions; this model of team production was originally introduced by [KR15] and is in the spirit of team performance according to a generalized Thurstone model [Thu27], e.g., used by popular rating systems such as TrueSkill [GMH07]. Under the given assumptions, the team selection problem asks to maximize a submodular function that, in general, is an NP-hard problem.

In this paper, we consider two types of approximation algorithms for the team selection problem. We consider a natural greedy algorithm that sequentially hires individuals based on greedy choices with largest marginal profit as long as this is beneficial. For the team selection problem with a cardinality constraint, we consider test-score algorithms that select a team of a given size that consist of individuals with largest individual test scores. The individual test scores are computed for each individual by performing a test of some kind, e.g., this could be an interview for a job applicant, a screening survey in an online labour platform such as Upwork or TopCoder, or an admission test such as SAT or GRE used for college or graduate school admissions.

Summary of Main Contributions We characterize the approximation ratio of the greedy algorithm for the team selection problem for arbitrary cost function of the team size with increasing marginal costs. This approximation ratio is parametrized with the parameter $a > 1$, which is equal to the ratio of the utility of production and the cost of a profit-maximizing solution, and is as given here

$$1 - \frac{\log(a)}{a-1} \quad \text{when} \quad a \leq e \quad \text{and} \quad 1 - \frac{a}{e(a-1)} \quad \text{when} \quad a \geq e.$$

The special case of the team selection problem with a cardinality constraint is a limit case as the value of parameter a goes to infinity: in this limit case, our approximation ratio coincides with known approximation ratio of value $1 - 1/e$ for the problem of maximizing a non-negative monotone submodular set function. Our result extends the previously best-known approximation guarantee of the greedy algorithm by [FIMN13], which is restricted to the special case of linear cost functions. The result is obtained using a novel proof, which allows to study the case of arbitrary increasing cost functions with increasing marginal costs.

For the team selection problem with a cardinality constraint, we show several new results on the approximation guarantees of team-score algorithms. We show that the existence of test scores that guarantee a constant-factor approximation is equivalent to the existence of special type of test scores – we refer to as *replication test scores*. For a given team production function, the replication test score of an individual is defined as the expected team production output of a team consisting of independent

replicas of the given individual. The result implies that when searching for good test scores that guarantee a constant-factor approximation for the team selection problem, it suffices to restrict attention to replication test scores. We identify a set of sufficient conditions for replication test scores to guarantee a constant-factor approximation for a given team selection problem; specifically, we show that these sufficient conditions guarantee a $1/9$ approximation. These sufficient conditions are shown to be verified for a large set of special instances of stochastic models of team production, including best-shot and constant elasticity of substitution production functions; defined in Section 3. We evaluate performance of team-score algorithms using data about individual performances as observed in a popular platform for software development.

The paper is structured as follows. Section 2 provides a discussion of related work. Section 3 introduces notation, problem definition, and a catalogue of examples of production functions used as running examples throughout the paper. Section 4 presents the approximation guarantee of the greedy algorithm. Section 5 introduces test-score algorithms and presents results on their approximation guarantees. Section 6 contains our experimental results. Finally, we conclude in Section 7.

2 Related Work

The celebrated result by [NWF78] established that for maximizing a non-negative monotone submodular set function subject to a cardinality constraint, the greedy algorithm guarantees a $1 - 1/e$-approximation of the optimal solution. This factor has been shown to be optimal for the value oracle model where an algorithm only has access to the value of the function for each given subset of the ground set and if only a polynomial number of queries is allowed [NW78], [Fei98] and [KLMM08].

The problem of maximizing a non-negative monotone submodular set function has been subsequently studied for different types of constraints. [Von08] showed that for a submodular welfare problem, defined as maximizing a sum of monotone submodular utility functions subject to a matroid constraint, a greedy algorithm guarantees a $1 - 1/e - o(1)$-approximation of the optimal solution. [AG12] showed that for linear packing constraints $Ax(S) \leq b$, where $A \in [0,1]^{m \times n}$, $b \in [1, \infty)^m$, and $x_i(S) = 1$ if $i \in S$ and $x_i(S) = 0$ otherwise, there exists a $\Omega(1/m^{1/W})$-approximation algorithm, where $W = \min\{b_i/A_{i,j} : A_{i,j} > 0\}$ is the width of the packing constraints; this implies a constant-factor approximation when the number of constraints is constant, or when the width of the constraints is sufficiently large.

More recent work studied efficient algorithms for maximizing a non-negative monotone submodular set function subject to different types of constraints. [BV14] found fast algorithms for maximizing a non-negative monotone submodular set function $f : 2^{[n]} \to \mathbf{R}_+$ subject to different types of constraints. In particular, for the problem with a cardinality constraint, they found an $1 - 1/e - \epsilon$-approximation algorithm that uses $O(\frac{n}{\epsilon} \log \frac{n}{\epsilon})$ queries; note that standard greedy algorithm requires instead $O(nk)$ queries, for the cardinality of value k. Further results in this direction were established by [BFS15]. [BMKK14] have found a one-pass streaming algorithm for maximizing a monotone submodular set function subject to a cardinality constraint that guarantees a $1/2 - \epsilon$-approximation using a memory of size $O(k \log(k)/\epsilon)$ and a running time of value $O(n \log(k)/\epsilon)$, for an arbitrary constant $\epsilon > 0$.

The problem of maximizing a non-negative *non-monotone* submodular set function subject to a cardinality constraint has been studied by several authors. [FNS11] have found an $1/e - o(1)$-approximation algorithm when the number of items in the solution is within given cardinality constraint. [Von09] have found a $1/4 - o(1)$-approximation algorithm for the case when the number of items in the solution is required to be exactly equal to the given cardinality constraint. [BFNS14]

Table 1: Approximation results for the profit maximization problem.

Utility function	Cost function	Algorithm	Approximation ratio
monotone submodular	cardinality constraint	greedy	$1 - \frac{1}{e}$ [NW78]
monotone submodular	linear	greedy	$1 - \frac{\log(a)}{a-1}$ [FIMN13]
monotone submodular	convex	greedy	$1 - \frac{\log(a)}{a-1}$ if $a \leq e$; $1 - \frac{a}{e(a-1)}$ if $a \geq e$
linear	convex	greedy*	$\frac{1}{3}$ [BUCM12]
top-m	cardinality constraint	test score	$\frac{1}{30}$ [KR15]
a class of submodular	cardinality constraint	test score	$\frac{1}{9}$

derived several improved approximation guarantees. [FMV07] and [LMNS09] studied the problem subject to matroid or knapsack constraints; in particular, they have found an $1/5 - \epsilon$-approximation algorithm for any number of knapsack constraints, where $\epsilon > 0$ is any constant. These results do not apply to our problem as our objective function is not necessarily non-negative.

The problem of maximizing a profit function defined as difference of a non-negative monotone submodular set function and a non-negative monotone cost function have also been studied. [FIMN13] studied the special case of a linear cost function, and showed that in a worst-case, the value of the solution of the greedy algorithm can be an arbitrarily small fraction of the optimum solution. As a way to circumvent the negative results of the worst-case analysis, they introduced a framework for designing and analysing algorithms that is suited to problems that are inapproximable according to the standard worst-case analysis. This amounts to designing guarantees for classes of instances, parametrized according to the properties of the optimal solution. In particular, for the problem of maximizing a profit function with a non-negative monotone submodular set function and a linear cost function, they showed that the greedy algorithm guarantees a $1-\log(a)/(a-1)$-approximation of the optimal solution, where a is the ratio of the utility and the cost of the optimal solution, and they showed that this is optimal. We extend this result for a more general class of convex cost function, which includes linear cost functions as a special case.

Constant-factor approximation algorithms are known for special classes of utility and cost functions; for example, for the problem with the utility function defined as a sum of the values of items and a cost function that is a convex function of the sum of weights of items, taking the best of the following two outputs yields a $1/3$-approximation: (i) the output of a greedy algorithm and (ii) a single item that maximizes the profit [BUCM12]; referred to as greedy* in Table 1.

The problem of maximizing a set function subject to a cardinality constraint using a test-score algorithm was first introduced by [KR15]. They showed that for maximizing a particular submodular set function (top-m function) subject to a cardinality constraint, there exists a test-score algorithm that guarantees a constant-factor approximation. We obtained several new results for the test-score algorithms. We found that the existence of test scores that guarantee a constant-factor approximation is equivalent to the existence of special type of test scores – replication test scores. We identified a set of sufficient conditions for the existence of replication test scores that guarantee a constant-factor approximation; this conditions are shown to hold for most of production functions from our catalogue. We obtained new results for the performance of different types of test scores for the family of CES production functions.

The approximation results for the profit maximization problem are summarized in Table 1.

3 Team Selection Problem, Production and Cost Functions

In this section, we first provide a formal definition of the team selection problem and then introduce a number of classic models of team production.

Team Selection Problem Suppose given is a set of individuals $N = \{1, 2, \ldots, n\}$, a *utility of production function* $u : 2^N \to \mathbf{R}_+$ that returns a positive real-value for any subset $S \subseteq N$, and a *cost function* $c : \{0, 1, \ldots, n\} \to \mathbf{R}_+$ that returns a positive real-value for any team size $|S|$. We define the *profit function* $p : 2^N \to \mathbf{R}$ to be quasi-linear in the utility of production and the cost function, i.e. for every $S \subseteq N$ it is defined by $p(S) = u(S) - c(|S|)$. The goal is to find a set of individuals $S^* \subseteq N$ that maximizes the profit function, i.e.

$$S^* \subseteq \text{argmax}_{S \subseteq N} p(S).$$

We will use the abbreviating notation $p^* = p(S^*)$, $u^* = u(S^*)$, and $c^* = c(|S^*|)$. We also define $a = u^*/c^*$, which we will show to play an important role in characterizing the performance of the greedy algorithm.

The utility of production function u is assumed to be non-negative, monotonically increasing, and submodular set function, and the cost function is assumed to be non-negative and with monotonically increasing increments. Under these assumptions, the goal is to solve a combinatorial optimization problem of maximizing a submodular function, which is known to be NP-hard. Hence, we have to settle to consider approximation algorithms for the given problem. An algorithm $\mathcal{A}$ is said to be a c-approximation algorithm if it outputs a set $S^{\mathcal{A}}$ with a profit of value $p^{\mathcal{A}} = p(S^{\mathcal{A}})$ such that $p^{\mathcal{A}} \geq cp^*$.

Stochastic Model of Team Production A stochastic model of team production assumes additional structure that is used to define the utility of production function u. Suppose that the individuals are associated with respective performances $X_1, X_2, \ldots, X_n$ that are assumed to be independent random variables with cumulative distribution functions $F_1, F_2, \ldots, F_n$. Suppose that for every given set of individuals $S \subseteq N$, a function $f : \mathbf{R}^{|S|} \to \mathbf{R}_+$ is given, which returns a positive real-value for every given vector of individual performances $X_S = (X_i, i \in S)$. We assume that function f is permutation-invariant, meaning that it assumes the same value for any permutation of its arguments. For every given $S \subseteq N$, the utility of production is defined as

$$u(S) = \mathbf{E}[f(X_S)]. \tag{1}$$

We shall refer to $(f, \mathcal{F})$ as a *stochastic model of team production*, where $\mathcal{F} = (F_1, F_2, \ldots, F_n)$.

A Catalogue of Production Functions We introduce some classic models of production functions that are defined for every given non-empty set of individuals $S \subseteq N$ and values of individual production inputs $x = (x_i, i \in S)$ as follows:

1. *Total production*:

$$f(x_S) = g\left(\sum_{i \in S} x_i\right) \tag{2}$$

where $g : \mathbf{R} \to \mathbf{R}_+$ is a non-negative monotone increasing function.

2. *Best-shot*:
$$f(x_S) = \max_{i \in S} x_i. \tag{3}$$

3. *Top-m*: given an integer m such that $1 \leq m \leq |S|$,
$$f(x_S) = \sum_{i=1}^{m} x_{(S,i)} \tag{4}$$
where $x_{(S,1)}, x_{(S,2)}, \ldots, x_{(S,|S|)}$ are the values of x_S rearranged in decreasing order.

4. *Constant Elasticity of Substitution (CES)*: for given parameter $p > 0$,
$$f(x_S) = \left(\sum_{i \in S} x_i^p \right)^{1/p}. \tag{5}$$

5. *Success-Probability*:
$$f(x_S) = 1 - \prod_{i \in S} (1 - g(x_i)), \tag{6}$$
where $g : \mathbf{R} \to [0, 1]$ is an increasing function.

A production function defined as an increasing function of the total individual investment in a production activity, as given in (2), is a natural model of production. For a concave function g, this production function exhibits a diminishing returns increase property, where the marginal increase of the production output becomes smaller or remains constant, the larger the total investment in the production.

The best-shot production function in (3) defines the production output to be the largest production input invested in the production activity. This type of production is common in online crowdsourcing systems where often multiple solutions are solicited for a task, and eventually only the best solution is selected.

The top-m production function in (4) is a natural generalization of the best-shot production function, where instead of restricting to selecting only the best solution for a task, a given number of best solutions is selected.

The constant elasticity of substitution or CES production function, defined in (5), was considered in [Sol56] and has been much popularized following [ACMS61]. The CES production function is a textbook example of a production function, e.g., see Chapter 1 in [Var92]. [ACMS61] used a CES production function to describe how capital and labour map to a value of production. [Arm69] used a CES production function as a model of demand for products distinguished by place of production. [DS77] used a CES production function as a model of demand for commodities that are substitutes among themselves in a monopolistic market to study optimum product diversity. Several properties of the CES production functions were studied by [Uza62] and [McF63]. The CES production function exhibits constant returns to scale, meaning that scaling the production inputs by a factor results in scaling the production output by the same factor. This is equivalent to saying that the production function is homogenous of degree 1, i.e. $f(tx_1, tx_2, \ldots, tx_n) = tf(x_1, x_2, \ldots, x_n)$ for all $t \geq 0$. The CES production function corresponds to a weighted mean defined in [HLP52] as follows: for given values $x = (x_1, x_2, \ldots, x_n) \in \mathbf{R}^n$, and fixed parameters $q_i > 0$ for $i = 1, 2, \ldots, n$, a weighted mean of x is given by $\mathcal{M}_p(x) = (\sum_{i=1}^n q_i x_i^p / \sum_{i=1}^n q_i)^{1/p}$, where p is any real value except for (i) $p = 0$

and (ii) $p < 0$ and $x_i = 0$ for some $i \in \{1, 2, \ldots, n\}$. The family of CES production functions accommodates different types of productions by a suitable choice of parameter p. The CES production function models a production that is linear in the total production input for the value of parameter $p = 1$, and it corresponds to the best-shot production in the limit as the value of parameter p goes to infinity. The success-probability production function, defined in (6), is often used as a model of tasks for which each individual solution is either good or bad, and it suffices to have at least one good solution for the task to be successfully solved.

The utility of production function is a submodular set function under the following conditions on the production function f for our given examples. For the total production model, it suffices to assume in addition that function g is a concave function, i.e. it exhibits a diminishing increase with the value of the total production. The utility of production under either the best-shot, the top-m or the success-probability production function is a submodular function without making any additional assumptions. The utility of production under the CES production function is a submodular function if and only if $p \geq 1$.

Cost Functions The class of cost functions with increasing marginal cost accommodates several special cases of interest. For example, it accommodates a linear cost function $c(x) = ax$, for a constant marginal cost $a > 0$, or a quadratic cost function $c(x) = \binom{x}{2}$ that can be interpreted as the number of potential ties between individual team members. Another example of interest is that of a *cardinality constraint*: given an integer $k \geq 1$, the cost function is defined as follows

$$c(x) = \begin{cases} 0 & \text{if} \quad x \leq k \\ \infty & \text{if} \quad x > k. \end{cases} \tag{7}$$

4 Greedy Algorithm and its Approximation Guarantee

We consider a natural *greedy algorithm* for the team selection problem that selects team members sequentially with each hiring decision being a greedy choice that maximizes marginal profit gain, until either a hiring decision yields a non-positive marginal profit or all the team members are hired. Formally, the greedy algorithm initializes $S_0 \leftarrow \emptyset$ at step $t = 0$, and then performs the following steps:

At each step t, do:
  Find $i^* = \text{argmax}_{i \in N \setminus S_t} p(S_t \cup \{i\}) - p(S_t)$
    If $p(S_t \cup \{i^*\}) - p(S_t) > 0$
      $S_{t+1} \leftarrow S_t \cup \{i^*\}$
    Else terminate.

The greedy algorithm is known to be optimal for some particular models of stochastic team production. For example, from the work [KR15], for the best-shot production function and individual production inputs according to weighted Bernoulli random variables, for any cost function with increasing marginal cost, the greedy algorithm is optimal.

From the classic work of [NW78], the greedy algorithm is known to have the following approximation guarantee for the special case of the budget constraint: $p^{\mathcal{G}} \geq (1 - 1/e)p^*$. A more detailed statement is asserted in the following proposition:

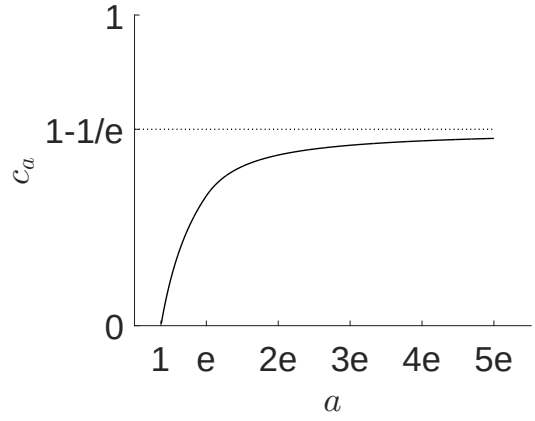

Figure 1: Constant c_a of the greedy algorithm approximation versus a.

Proposition 1 ([NW78]) *For every given k, and every step of* greedy algorithm*, it holds*

$$u(S_t) \geq \left(1 - \left(1 - \frac{1}{k}\right)^t\right) \max_{S:|S|=k} u(S) \geq \left(1 - e^{-t/k}\right) \max_{S:|S|=k} u(S).$$

For the special case of the cost function restricted to be linear, from the work by [KPR98], it is known that the greedy algorithm can guarantee a constant-factor approximation when c^*/u^* is bounded away from 1, and [FIMN13] have established the following approximation guarantee:

$$p^{\mathcal{G}} \geq \left(1 - \frac{\log(a)}{a-1}\right) p^*. \tag{8}$$

In the following theorem, we establish an approximation guarantee of the greedy algorithm for the team selection problem with an arbitrary increasing cost function with increasing marginal cost.

Theorem 1 *For the profit maximization problem with a non-negative monotone submodular utility function and a cost function with increasing marginal cost, the greedy algorithm guarantees a solution that is a c_a-approximation of the optimum solution, i.e. $p^{\mathcal{G}} \geq c_a p^*$ where*

$$c_a = \begin{cases} 1 - \frac{\log(a)}{a-1}, & \text{if } a \leq e \\ 1 - \frac{1}{e}\frac{a}{a-1}, & \text{if } a \geq e. \end{cases}$$

See Figure 1 for a graph of c_a versus a.

Proof Let $k^* = |S^*|$. From Proposition 1, for every $t \in \{0, 1, \ldots, k^*\}$,

$$\begin{aligned} u(S_t) - c(t) &\geq \left(1 - \left(1 - \frac{1}{k^*}\right)^t\right) u^* - c(t) \\ &\geq \left(1 - \left(1 - \frac{1}{k^*}\right)^t\right) u^* - \frac{t}{k^*} c^*, \end{aligned} \tag{9}$$

where the last inequality holds by the assumption that c is a function with increasing increments.

Let g be a decreasing piecewise linear function defined as follows:

$$g(x) = \begin{cases} \left(1 - \frac{1}{k^*}\right)^x, & \text{for} \quad x \in \{0, 1, \dots\} \\ (\lceil x \rceil - x)\left(1 - \frac{1}{k^*}\right)^{\lfloor x \rfloor} + (x - \lfloor x \rfloor)\left(1 - \frac{1}{k^*}\right)^{\lceil x \rceil}, & \text{otherwise.} \end{cases}$$

Combining with (9), we have

$$p^{\mathcal{G}} \geq \max_{x \in [0, k^*]} \left\{ (1 - g(x))\, u^* - \frac{x}{k^*} c^* \right\}. \tag{10}$$

We now establish the following inequality:

$$g(x) \leq e^{-x/k^*}, \text{ for every } x \in \mathbf{R}_+. \tag{11}$$

It suffices to show that for every positive integer t, we have

$$\max_{x \in [t, t+1]} g(x) \leq e^{-x/k^*}. \tag{12}$$

Suppose that $g(t) = e^{-\bar{x}/k^*}$ for some $\bar{x} \in [t, t+1]$. Otherwise, the condition $g(t) \neq e^{-x/k^*}$ for all $x \in [t, t+1]$ and the fact $g(t) = (1 - 1/k^*)^t \leq e^{-t/k^*}$ imply that $g(t) < e^{-x/k^*}$ for all $x \in [t, t+1]$, which is because g is a decreasing function implies (12).

Note that $g(x) - e^{-x/k^*}$ is a concave function on $[t, t+1]$, which is maximized at a unique point $\bar{x}$ since $\frac{d}{dx} g(x) = -\frac{1}{k^*} g(t)$ and $g(t) = e^{-\bar{x}/k^*}$. It is readily checked that

$$\max_{x \in [t, t+1]} \left\{ g(x) - e^{-x/k^*} \right\} = (\bar{x} - t) \left(\left(1 - \frac{1}{k^*}\right)^{t+1} - \left(1 - \frac{1}{k^*}\right)^{t} \right) \leq 0$$

which establishes inequality (11).

Combining (10) and (11), we obtain

$$p^{\mathcal{G}} \quad \geq \quad \max_{x \in [0, k^*]} \left\{ \left(1 - e^{-x/k^*}\right) u^* - \frac{x}{k^*} c^* \right\}.$$

Therefore, if $c^*/u^* \geq 1/e$, we have

$$\frac{p^{\mathcal{G}}}{p^*} \geq \frac{u^* - \left(1 - \log\left(\frac{c^*}{u^*}\right)\right) c^*}{u^* - c^*} = 1 - \frac{\log(a)}{a - 1},$$

and, otherwise, we have

$$\frac{p^{\mathcal{G}}}{p^*} \geq \frac{(1 - e^{-1}) u^* - c^*}{u^* - c^*} = 1 - \frac{a/e}{a - 1}.$$

■

The approximation ratio c_a in Theorem 1 increases with parameter a, from zero value at $a = 1$ to value $1 - 1/e$ as a goes to infinity. The limit value of the approximation ratio c_a as a goes to infinity coincides to that for the team selection problem with a cardinality constraint. This is intuitive as for the team selection with a cardinality constraint, the value of the utility of production is strictly positive and the value of the cost in any optimal solution is equal to zero, hence the value of parameter a is infinite. The approximation ratio c_a is indeed smaller than or equal to the value in (8), which is an approximation ratio for the restricted case of the team selection problem with a linear cost function. More specifically, c_a coincides to that in (8) for the case $a \leq e$, and is strictly smaller, otherwise.

The result in Theorem 1 is established using a proof that uses the known approximation guarantee of the greedy algorithm for the problem of maximizing a non-negative monotone submodular set function subject to a cardinality constraint and the increasing marginal cost property. The proof is different from that in [FIMN13], which is for the special case of a linear cost function; we provide a more detailed discussion in Appendix A.

5 Test-Score Algorithms and Their Approximation Guarantees

In this section, we consider approximation guarantees of test-score algorithms for the team selection problem with a cardinality constraint. A *test-score algorithm* is defined for given values of individual test scores $s_1, s_2, \ldots, s_n$ and a cardinality constraint $k \geq 1$ by hiring a set of k individuals with the largest values of test scores. We consider the utility of production according to a $(f, \mathcal{F})$ stochastic model of team production, which is introduced in Section 3. The test scores can be defined in different ways for a given choice of the production function f and cumulative distribution functions of individual performances $\mathcal{F} = (F_1, F_2, \ldots, F_n)$.

Several examples of test scores are given in the following list of examples:

1. *Mean test scores*: $s_i = \mathbf{E}[X_i]$, for $i \in N$.
2. *Quantile test scores*: $s_i = \mathbf{E}\left[X_i \mid F_i(X_i) \geq q\right]$, for $i \in N$, where q is a positive-valued parameter.
3. *Replication test scores*: $s_i = \mathbf{E}[f(X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(k)})]$, for $i \in N$, where $X_i^{(1)}, \ldots, X_i^{(k)}$ are independent and identically distributed random variables with cumulative distribution F_i.

The mean test scores represent a natural definition of test scores, defining a score of an individual to be equal to his or her expected performance. The quantile test score of an individual is defined as the conditional expected value of his or her individual performance, conditional on that it is at least of value as large as the q-quantile of his or her distribution of performance. The quantile test scores have been considered by [KR15] for the stochastic model of production with top-m production function. In particular, for the best-shot production function, $q = 1 - \theta/k$ for a positive constant θ. The replication test score for an individual is defined for given choice of the production function f as the expected value of production of a team of size k that consists of independent replicas of the given individual.

5.1 Good Test Scores

We introduce a condition that defines a subset of test scores, we refer to as *good test scores*, and then show that any test-score algorithm that uses good test scores guarantees a constant-factor approximation for the team selection problem.

Definition 1 (good test scores) *For given utility of production function u, any given test scores $s_1, s_2, \ldots, s_n$ are said to be* good test scores, *if there exists a monotone increasing function h and constants $c_1, c_2 > 0$ such that*

$$c_2 \min_{i \in S} h(s_i) \leq u(S) \leq c_1 \max_{i \in S} h(s_i), \text{ for every } S \subseteq \tilde{N} \tag{13}$$

where $\tilde{N}$ denotes a multi-set that consists of all elements in N, and each of these elements having at least $n-1$ duplicates.

The good test scores are shown to imply the following guarantee.

Theorem 2 *Whenever for a given utility of production there exist good test scores, then the test-score algorithm that uses good test scores yields a solution with the following approximation guarantee:*

$$p^{\mathcal{S}} \geq \frac{c_2}{c_1 + c_2} p^*.$$

Proof Let S be the set output by the test-score algorithm. By the monotone increasing property of the utility of production function u, we have

$$\begin{aligned} u(S^*) &\leq u(S^* \cup S) \leq u(S) + u(S^* \setminus S) \\ &\leq u(S) + c_1 \min_{i \in S} h(s_i) \leq u(S) + \frac{c_1}{c_2} u(S) \end{aligned}$$

which establishes the theorem. ■

The result in Theorem 2 tells us that if for a given utility of production function there exist good test scores, then the test-score algorithm using good test scores guarantees a constant-factor approximation for the team selection problem. It remains to understand when for particular choice of a utility of production there exist good test scores, and given that there exist good test scores, how in particular one would define them, and what exactly is the approximation gurantee of a given choice of good test scores.

5.2 A Necessary and Sufficient Condition for the Existence of Good Test Scores

The following theorem shows that the existence of good test scores is equivalent to the existence of a special type of test scores – replication test scores.

Theorem 3 *If for a given utility of production function there exist good test scores, then replication test scores are good test scores.*

Proof Suppose that for a utility of production function u, there exist good test scores $\bar{s}_1, \ldots, \bar{s}_n$, so that we have

$$c_2 \min_{i \in S} \bar{s}_i \leq u(S) \leq c_1 \max_{i \in S} \bar{s}_i, \text{ for every } S \subseteq \tilde{N}. \tag{14}$$

Consider an arbitrary $i \in N$. Let $i^{(1)}, \ldots, i^{(k)}$ be a sequence of individuals with performances $X_i^{(1)}, \ldots, X_i^{(k)}$, which are independent and identically distributed random variables with cumulative distribution function F_i. Let s_i denote the replication test score defined by

$$s_i = \mathbf{E}[f(X_i^{(1)}, \ldots, X_i^{(k)})] = u(\{i^{(1)}, \ldots, i^{(k)}\}). \tag{15}$$

Since $\bar{s}_1, \ldots, \bar{s}_n$ are good test scores, we have

$$c_2 \bar{s}_i \leq u(\{i^{(1)}, \ldots, i^{(k)}\}) \leq c_1 \bar{s}_i. \tag{16}$$

From (14), (15), and (16), we have

$$\frac{c_2}{c_1} \min_{i \in S} s_i \leq c_2 \min_{i \in S} \bar{s}_i \leq u(S) \leq c_1 \max_{i \in S} \bar{s}_i \leq \frac{c_1}{c_2} \max_{i \in S} s_i$$

which implies that replication test scores are good test scores. ■

From Theorem 3, we observe that for every given production function, we can check whether there exist good test scores by just checking whether replication test scores are good test scores. If for a given production function, replication test scores are not good test scores, then there exist no good test scores.

5.3 A Sufficient Condition for Replication Test Scores to be Good

We present a set of sufficient conditions for replication test scores to be good test scores. This set of conditions holds for all production functions from our catalogue of examples except top-m function; see Appendix B. Note that for top-m function there exist good test scores by the result in [KR15], which implies that replication test scores are good test scores for top-m function from Theorem 3.

We introduce the following conditions:

(S) $u(S) = \mathbf{E}[f(X_S)]$ is a non-negative, monotone submodular set function;

(M) $f(x, y) - f(x)$ is decreasing in x for every $y \in \mathbf{R}_+$;

(B) $f(f^{-1}(f(x_1, x_2, \ldots, x_{l-1})), x_l) \leq f(x_1, x_2, \ldots, x_l)$

where $f^{-1}(x) = \max\{y \in \mathbf{R}_+ \mid f(y) \leq x\}$.

The next theorem tells us when replication test scores are good test scores, and identifies the values of constants in the definition of good test scores.

Theorem 4 *The following two claims hold:*

1. *If condition (S) holds, then replication test scores satisfy the lower bound in (14) with $c_2 = 1/2$.*
2. *If conditions (S), (M), and (B) hold, then replication test scores satisfy the upper bound in (14) with $c_1 = 4$.*

Proof We prove the two asserted claims as follows.

Proof of Claim 1 Without loss of generality, consider the set $S = \{1, 2, \ldots, k\}$ and assume that $s_1 = \min_{i \in S} s_i$. We show that for every $j \in \{1, 2, \ldots, k\}$, $u(\{1, 2, \ldots, j\}) \geq \frac{j}{2k} s_1$. From this, it then follows that $u(S) \geq s_1/2$. The proof is by mathematical induction. Base case $j = 1$: since u is a non-negative, monotone submodular set function, we have

$$u(\{1\}) = \frac{1}{k} \sum_{t=1}^{k} u(\{1^{(t)}\}) \geq \frac{1}{k} u(\{1^{(1)}, \ldots, 1^{(k)}\}) = \frac{s_1}{k}. \tag{17}$$

Induction step: suppose that $u(\{1, \ldots, j\}) \geq \frac{j}{2k} s_1$ holds for $1 \leq j < k$ and we need to show that it holds that $u(\{1, \ldots, j+1\}) \geq \frac{j+1}{2k} s_1$. Note that

$$\begin{aligned}
& u(\{1, \ldots, j+1\}) = u(\{1, \ldots, j\}) + [u(\{1, \ldots, j+1\}) - u(\{1, \ldots, j\})] \\
\overset{(a)}{\geq}\ & u(\{1, \ldots, j\}) + \frac{1}{k}[u(\{1, \ldots, j, (j+1)^{(1)}, \ldots, (j+1)^{(k)}\}) - u(\{1, \ldots, j\})] \\
\overset{(b)}{\geq}\ & u(\{1, \ldots, j\}) + \frac{1}{k}[u(\{(j+1)^{(1)}, \ldots, (j+1)^{(k)}\}) - u(\{1, \ldots, j\})] \\
\geq\ & \left(1 - \frac{1}{k}\right) u(\{1, \ldots, j\}) + \frac{s_{j+1}}{k} \\
\geq\ & \frac{j+1}{2k} s_1,
\end{aligned} \tag{18}$$

where (a) and (b) hold by the assumption that u is a non-negative, montonically increasing, and submodular function, and the last inequality follows by the induction hypothesis.

Proof of Claim 2 Without loss of generality, assume that $S = \{1, 2, \ldots, k\}$ and $s_1 \leq s_2 \leq \cdots \leq s_k$. Let i^* be an individual such that $X_{i^*} = x^*$ with probability 1, for x^* such that $f(x^*) = cs_k$, for a constant $c \geq 1$. Since u is a non-negative, monotone increasing, and submodular function, we have

$$\begin{aligned} u(S) &\leq u(\{i^*\} \cup S) \\ &\leq u(\{i^*\}) + \sum_{i=1}^{k} \left(u(\{i^*\} \cup \{i\}) - u(\{i^*\})\right) \\ &= cs_k + \sum_{i=1}^{k} \left(u(\{i^*\} \cup \{i\}) - u(\{i^*\})\right). \end{aligned} \tag{19}$$

Now, note that

$$\begin{aligned} &u(\{i^*\} \cup \{i\}) - u(\{i^*\}) = \mathbf{E}[f(x^*, X_i) - cs_k] \\ &\overset{(a)}{\leq} \mathbf{E}\left[f(f^{-1}(f(X_i^{(1)}, \ldots, X_i^{(k-1)})), X_i^{(k)}) - f(X_i^{(1)}, \ldots, X_i^{(k-1)}) | f(X_i^{(1)}, \ldots, X_i^{(k-1)}) \leq cs_k\right] \\ &\overset{(b)}{\leq} \mathbf{E}\left[f(X_i^{(1)}, \ldots, X_i^{(k)}) - f(X_i^{(1)}, \ldots, X_i^{(k-1)}) | f(X_i^{(1)}, \ldots, X_i^{(k-1)}) \leq cs_k\right] \\ &\leq \frac{u\left(\{i^{(1)}, \ldots, i^{(k)}\}\right) - u\left(\{i^{(1)}, \ldots, i^{(k-1)}\}\right)}{\mathbf{Pr}[f(X_i^{(1)}, \ldots, X_i^{(k-1)}) \leq cs_k]} \\ &\overset{(c)}{\leq} \frac{s_i/k}{\mathbf{Pr}[f(X_i^{(1)}, \ldots, X_i^{(k-1)}) \leq cs_k]} \\ &\overset{(d)}{\leq} \left(1 - \frac{1}{c}\right)^{-1} \frac{s_k}{k}, \end{aligned} \tag{20}$$

where (a), (b), and (c) hold by conditions (M), (B), and (S), respectively, and (d) holds because by Markov's inequality and condition (S)

$$\mathbf{Pr}[f(X_i^{(1)}, \ldots, X_i^{(k-1)}) \geq cs_k] \leq \frac{\mathbf{E}[f(X_i^{(1)}, \ldots, X_i^{(k-1)})]}{cs_k} \leq \frac{\mathbf{E}[f(X_i^{(1)}, \ldots, X_i^{(k)})]}{cs_k} \leq \frac{1}{c}.$$

From (19) and (20), we obtain $u(S) \leq \frac{c^2}{c-1} s_k$, which implies Claim 2 when $c = 2$. ■

The result of Theorem 4 has the following corollary.

Corollary 1 *Under conditions (S), (M), and (B), the test-score algorithm that uses the replication test scores yields a solution that guarantees a* $1/9$ *approximation of the optimum solution.*

5.4 CES Production Function

In this section we characterize the approximation guarantees of the team selection by using either the mean test scores or the quantile test scores for a stochastic model of production according to a CES production function with parameter $p \geq 1$.

5.4.1 Mean Test Scores

The following theorem characterizes the approximation guarantee of the team selection using the mean test scores.

Theorem 5 *Suppose that the utility of production is according to a stochastic model of production with a CES production function with parameter $p \geq 1$. For every given team size $k \geq 1$, let M be a team of size k that consists of individuals of highest mean test scores, and let S^* be a team of size k that maximizes the expected utility of production. Then, we have*

$$u(M) \geq \frac{1}{k^{1-1/p}} u(S^*).$$

Moreover, this bound is tight.

Proof of the theorem is provided in Appendix C.

Note that for the value of parameter $p = 1$, selecting a team of individuals with the largest mean test scores is optimal. Intuitively, one would expect that for small enough values of parameter $p > 1$, the mean test scores would be good test scores. The result of Theorem 5 tells us that this is so if and only if $p = 1 + O(1/\log(k))$. In the limit as p goes to infinity, in which the CES production function corresponds to the best-shot production function, we have that the expected utility of a team with the largest mean test scores is guaranteed to be at least $1/k$ of the optimum expected utility, and this is a tight bound; this conforms to the result in [KR15].

5.4.2 Quantile Test Scores

Since the CES production function corresponds to the best-shot function in the limit of large values of parameter p, and we know from the result in [KR15] that quantile test scores are good test scores for the best-shot function, one would expect that quantile test scores are good test scores for the CES production function provided that the value of parameter p is large enough. In the next theorem, we characterize a tight threshold for the parameter p below which the quantile test scores are not good test scores for the CES production function.

Theorem 6 *The following claims hold for quantile test scores with $q = 1 - \theta/k$:*

1. *If $p = o(\log(k))$ and $p > 1$, the quantile test scores are not good test scores for any value of parameter $\theta > 0$;*
2. *If $p = \Omega(\log(k))$, the test-score algorithm with quantile test scores with $\theta = 1$ yields a solution that is a constant-factor approximation of the optimum solution.*

Proof of the theorem is given in Appendix D.

6 Experimental Results

In this section, we present results of empirical study using data crawled from TopCoder, a popular online platform for software development. Our goal is to evaluate performance of different test-score algorithms for the team selection problem and compare their performance with optimum team selection. Overall, the empirical results show that the average performance of a team selected using replication test scores is typically near optimal, and that for some stochastic models of team production, they have significantly better worst-case performance than some other test scores.

Figure 2: (Left) Number of tasks per worker versus the worker rank with respect to the number of tasks; (Middle) Mean score per worker versus the worker rank with respect to the mean scores; (Right) Mean score per worker conditional on the number of tasks per worker.

6.1 Dataset and Basic Statistics

Our dataset contains information about solutions to web software development tasks submitted by coders (we interchangeably refer to as workers) over a period from November 2003 to January 2013. In our dataset, each solution is associated with the identity of the coder, identity of the task, and the value of the score associated to the solution. Each such score is a real number in the interval from value 0 to value 100. These scores are assigned to solutions by a rating procedure that is part of TopCoder online platform. We use these scores as indicators of individual performances. Table 2 presents some basic statistics.

Table 2: TopCoder dataset summary statistics.

# of workers	# of tasks	# of solutions	Tasks per worker	Workers per task	Mean score
658	2,924	7,127	10.83	2.44	87.54

Some additional statistics is shown in Figure 2. The number of tasks per worker covers a range from 1 to about 500 tasks. Out of total of 658 workers, 75 of them have submitted solutions to 20 or more tasks and 124 of them have submitted solutions to 10 or more tasks. The mean scores of solutions submitted by workers cover a wide range from around 40 to 100 point scores. According to intuition, the workers seem to improve upon their performance as they submit more solutions to tasks.

Figure 3 shows statistics for the scores of submitted solutions. The values of scores cover a wide range of values from around 25 to 100. Conditioning on the scores of solutions submitted by workers who submitted at least θ solutions skews the distribution towards larger values. The cumulative distribution functions of scores of solutions submitted by individual workers in general differ from one worker to another as shown in Figure 3-right, for a set of workers who made the largest number of solutions.

6.2 Performance of Test-Score Algorithms

We evaluate performance of test-score algorithms for the team selection problem by the following method. We consider the set of workers who submitted at least θ solutions. We report the results for $\theta = 20$; we have also experimented with other values and observed qualitatively the same results. We fix a stochastic model of team production, the team size k, a test-score algorithm, and the distributions of individual performances to the empirical distributions of the scores observed in the data. Specifically,

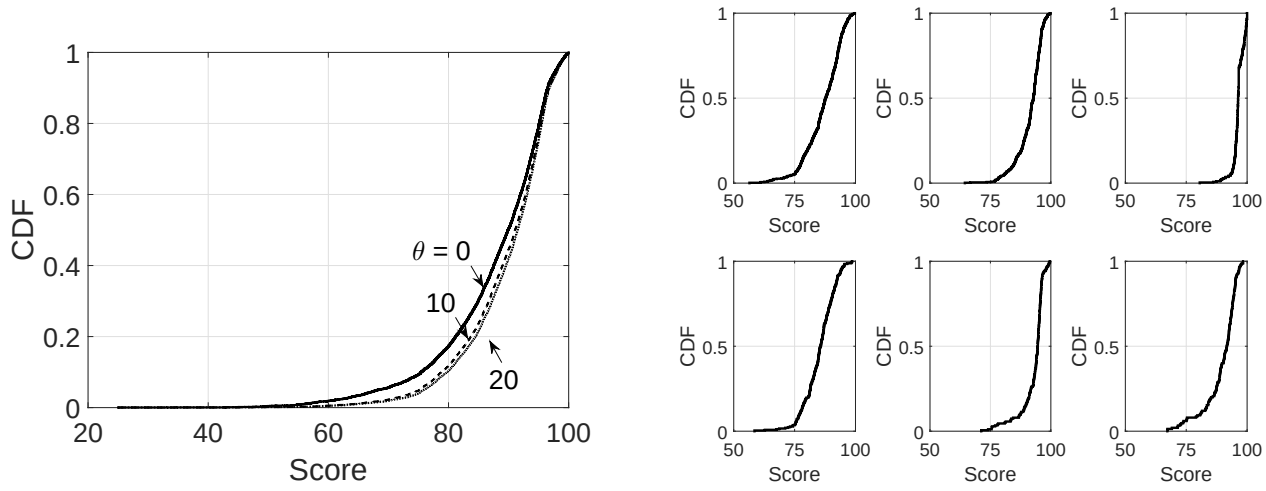

Figure 3: Cumulative distribution function of individual performance score: (left) aggregate over all submissions conditional on workers with at least θ submissions, (right) top 6 workers with respect to the number of submissions.

for a given worker i, we denote with $\hat{F}_i$ the empirical cumulative distribution function of his or her performance scores. For a worker i with n_i observed performance scores of values $x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n_i)}$, we define $\hat{F}_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{1}(x_i^{(j)} \leq x)$.

We estimate the expected utility of production by sampling a set of n workers uniformly at random without replacement from the input set of workers, and then apply given algorithm for the team selection problem for each sample of the set of workers. In all our experiments, we draw 10,000 such samples. We conducted experiments for different values of parameter n, including the values 5, 10, and 15. For space reasons, we report the results only for the case $n = 10$; the results for other values of this parameter were observed to be qualitatively similar. For every given team size, we compute the optimum value of the expected utility by a brute force search, examining all possible teams of given size.

6.2.1 Best-Shot Production Function

We compare the performance of the replication test scores and the mean test scores for the best-shot production function. The replication test scores are estimated by $\hat{s}_i = \int_{\mathbf{R}_+} (1 - \hat{F}_i(x)^k) dx$ and the mean test scores are computed by using the same formula but with $k = 1$.

Figure 4 shows results of our experiments. We observe that the expected utility indeed exhibits a diminishing returns increase with the team size. We observe that the replication test scores provide nearly optimal performance. The mean test scores provide worse performance, which on average is still near to the optimum performance, but has worst-case performance that can be significantly worse than the optimal performance.

6.2.2 Success-Probability Production Function

We conducted a similar analysis for the success-probability production function that is defined as follows. We assume that a solution submitted by a worker is successful if it achieves a point score larger than or equal to a threshold value δ. We have experimented with different values of this parameter, but for space reasons report the results only for $\delta = 90$. The replication test scores are estimated by

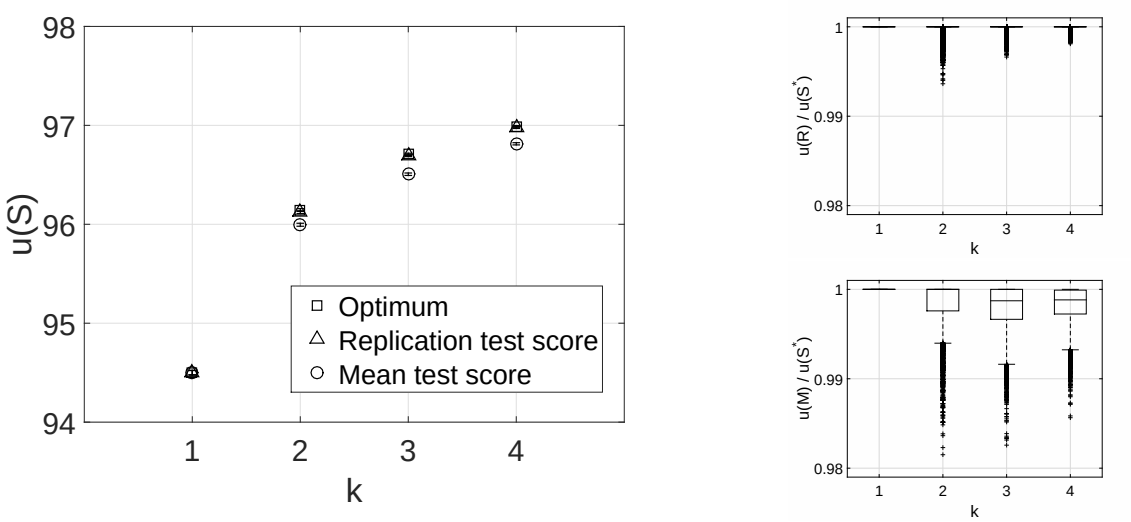

Figure 4: Team selection using test scores for the stochastic model of team production according to the best-shot function: (left) expected utility versus team size; (right-top) approximation ratio for replication test scores; (right-bottom) approximation ratio for mean test scores.

$\hat{s}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{1}(x_i^{(j)} \geq \delta)$. Figure 5 show the results of our experiments. We observe that there exist cases when the mean test scores provide significantly worse performance than using the replication test scores.

7 Conclusion

In this paper, we established the approximation guarantee of the greedy algorithm for the team selection problem with arbitrary increasing cost functions in the team size that has increasing marginal costs. For the team selection problem with a cardinality constraint, we showed that the existence of a test-score algorithm that guarantees a constant-factor approximation of the optimum solution is equivalent to the existence of good replication test scores. Sufficient conditions are identified for the existence of good replication test scores, which are shown to hold for several special instances of stochastic models of team production, and are shown to guarantee a $1/9$-approximation guarantee. For the constant elasticity of substitution production functions, we characterized the approximation guarantees of mean test scores and quantile test scores.

An open problem for future work is to further study the tightness of approximation guarantees of test-score algorithms.

References

[ACMS61] K. J. Arrow, H. B. Chenery, B. S. Minhas, and R. M. Solow. Capital-labor substitution and economic efficiency. *Review of Economics and Statistics (The MIT Press)*, 43(3):225–250, 1961.

[AG12] Yossi Azar and Iftah Gamzu. Efficient submodular function maximization under linear packing constraints. In *Proc. of ICALP '12*, pages 38–50, 2012.

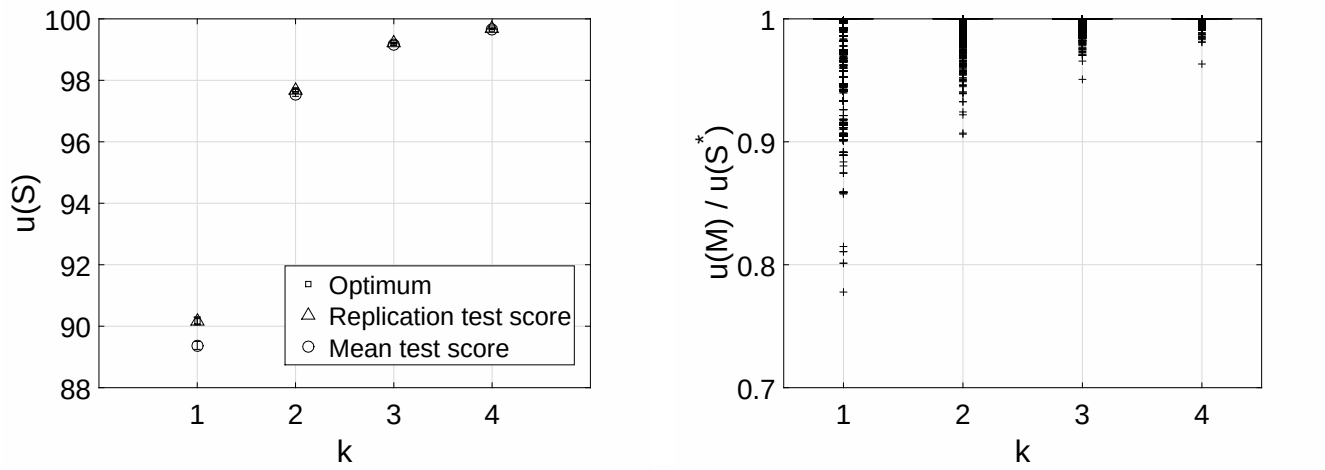

Figure 5: Team selection using test scores for the stochastic model of team production according to the success-probability function: (left) expected utility versus team size; (right) approximation ratio for mean test scores.

[Arm69] Paul S. Armington. A theory of demand for products distinguished by place of production. *Staff Papers (International Monetary Fund)*, 16(1):159–178, 1969.

[BFNS14] Niv Buchbinder, Moran Feldman, Joseph (Seffi) Naor, and Roy Schwartz. Submodular maximization with cardinality constraints. In *Proc. of SODA '14*, pages 1433–1452. SIAM, 2014.

[BFS15] Niv Buchbinder, Moran Feldman, and Roy Schwartz. Comparing apples and oranges: Query tradeoff in submodular maximization. In *Proc. of ACM SODA '15*, pages 1149–1168. SIAM, 2015.

[BMKK14] Ashwinkumar Badanidiyuru, Baharan Mirzasoleiman, Amin Karbasi, and Andreas Krause. Streaming submodular maximization: Massive data summarization on the fly. In *Proc. of ACM KDD '14*, pages 671–680, New York, NY, USA, 2014. ACM.

[BUCM12] Siddharth Barman, Seeun Umboh, Shuchi Chawla, and David Malec. Secretary problems with convex costs. In *Proc. of ICALP '12*, pages 75–87, 2012.

[BV14] Ashwinkumar Badanidiyuru and Jan Vondrák. Fast algorithms for maximizing submodular functions. In *Proc. of SODA '14*, pages 1497–1514. SIAM, 2014.

[DS77] Avinash K Dixit and Joseph E Stiglitz. Monopolistic Competition and Optimum Product Diversity. *American Economic Review*, 67(3):297–308, June 1977.

[Fei98] Uriel Feige. A threshold of ln n for approximating set cover. *J. ACM*, 45(4):634–652, July 1998.

[FIMN13] Uriel Feige, Nicole Immorlica, Vahab S Mirrokni, and Hamid Nazerzadeh. Pass approximation: A framework for analyzing and designing heuristics. *Algorithmica*, 66(2):450–478, 2013.

[FMV07] Uriel Feige, Vahab S. Mirrokni, and Jan Vondrák. Maximizing non-monotone submodular functions. In *Proc. of FOCS '07*, pages 461–471, Oct 2007.

[FNS11] M. Feldman, J. Naor, and R. Schwartz. A unified continuous greedy algorithm for submodular maximization. In *Proc. of FOCS '11*, pages 570–579, Oct 2011.

[GMH07] Thore Graepel, Tom Minka, and Ralf Herbrich. Trueskill(tm): A bayesian skill rating system. *Proc. of NIPS '07*, 19:569–576, 2007.

[GS92] Richard A. Guzzo and Gregory P. Shea. Group performance and intergroup relations in organizations. In Marvin D. Dunnette and Leaetta M. Hough, editors, *Handbook of industrial and organizational psychology*, chapter 5, pages 269–313. Consulting Psychologists Press, Palo Alto, CA, US, 1992.

[HLP52] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 2 edition, 1952.

[KLMM08] Subhash Khot, RichardJ. Lipton, Evangelos Markakis, and Aranyak Mehta. Inapproximability results for combinatorial auctions with submodular utility functions. *Algorithmica*, 52(1):3–18, 2008.

[KM86] David A. Kravitz and Barbara Martin. Ringelmann rediscovered: The original article. *Journal of Personality and Social Psychology*, 50(5):936–941, 1986.

[KM15] Rich Karlgaard and Michael S. Malone. *Team Genius: The New Science of High-Performing Organizations*. Harper Business, New York, NY, 2015.

[KPR98] Jon Kleinberg, Christos Papadimitriou, and Prabhakar Raghavan. Segmentation problems. In *Proc. of ACM STOC '98*, pages 473–482. ACM, 1998.

[KR15] Jon Kleinberg and Maithra Raghu. Team performance with test scores. In *Proc. of ACM EC '15*, pages 511–528. ACM, 2015.

[LMNS09] Jon Lee, Vahab S. Mirrokni, Viswanath Nagarajan, and Maxim Sviridenko. Non-monotone submodular maximization under matroid and knapsack constraints. In *Proc. of ACM STOC '09*, pages 323–332, New York, NY, USA, 2009. ACM.

[LN81] Bibb Latané and Steve Nida. Ten years of research on group size and helping. *Psychological Bulletin*, 89(2):308 – 324, 1981.

[LWH79] Bibb Latané, Kipling Williams, and Stephen Harkins. Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, 37(6):822–832, 1979.

[McF63] Daniel McFadden. Constant elasticity of substitution production functions. *The Review of Economic Studies*, 30(2):73–83, 1963.

[Mue12] Jennifer S. Mueller. Why individuals in larger teams perform worse. *Organizational Behavior and Human Decision Processes*, 117(1):111 – 124, 2012.

[NW78] George L Nemhauser and Leonard A Wolsey. Best algorithms for approximating the maximum of a submodular set function. *Math. of operations research*, 3(3):177–188, 1978.

[NWF78] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functionsi. *Math. Programming*, 14(1):265–294, 1978.

[Pag07] Scott E. Page. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Society*. Princeton University Press, 2007.

[SMF12] Bradley R. Staats, Katherine L. Milkman, and Craig R. Fox. The team scaling fallacy: Underestimating the declining efficiency of larger teams. *Organizational Behavior and Human Decision Processes*, 118(2):132 – 142, 2012.

[Sol56] R. M. Solow. A contribution to the theory of economic growth. *The Quaterly Journal of Economics*, 70:65–94, 1956.

[Ste72] I. Steiner. *Group process and productivity*. Academic Press, New York, NY, 1972.

[Thu27] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(2):273–286, 1927.

[Uza62] Hirofumi Uzawa. Production functions with constant elasticities of substitution. *The Review of Economic Studies*, 29(4):291–299, 1962.

[Var92] Hal R. Varian. *Microeconomic Analysis*. W. W. Norton & Company, 3 edition, 1992.

[Von08] Jan Vondrak. Optimal approximation for the submodular welfare problem in the value oracle model. In *Proc. of ACM STOC '08*, pages 67–74, New York, NY, USA, 2008. ACM.

[Von09] J. Vondrak. Symmetry and approximability of submodular maximization problems. In *Proc. of FOCS '09*, pages 651–670, Oct 2009.

A Discussion of the Proof of Theorem 1

We now discuss the approximation guarantee established in Theorem 1 and that in (8) that was established by [FIMN13] for the special case of linear cost functions. For the case of linear cost functions, it was shown in [FIMN13] that for every $\varepsilon > 0$, it is NP-hard to find a set $S \subset N$ such that

$$p(S) \geq \left(1 - \frac{\log(a)}{a-1} + \varepsilon\right) p^*.$$

and that the greedy algorithm guarantees

$$p^{\mathcal{G}} \geq \left(1 - \frac{\log(a)}{a-1}\right) p^*.$$

Their proof exploits a property of the linearity of the cost function, which we discuss as follows. When the greedy algorithm has utility of value x for set S_t at round t and $x < p^*$, $S^* \setminus S_t$ is non empty. From the non-decreasing property of u, we have

$$u(S_t \cup S^*) - u(S_t) \geq u^* - x.$$

Since u is a submodular function and c is a linear function, we have

$$\begin{aligned}\frac{p(S_{t+1})-p(S_t)}{u(S_{t+1})-u(S_t)} &\geq \frac{p(S_t\cup S^*)-p(S_t)}{u(S_t\cup S^*)-u(S_t)}\\ &\overset{(a)}{\geq} \frac{u(S_t\cup S^*)-u(S_t)-c^*}{u(S_t\cup S^*)-u(S_t)}\\ &\geq \frac{u^*-x-c^*}{u^*-x},\end{aligned}$$

where (a) uses the linearity of the cost function c.

Note that when the cost function has strictly increasing increments, then $c(|S_t\cup S^*|)-c(t)$ can be much larger than c^*.

It follows that for the case of linear cost functions, we have

$$\begin{aligned}p^{\mathcal{G}} &\geq \sum_{t=0}^{\infty}\max\left\{0,\frac{u^*-u(S_t)-c^*}{u^*-u(S_t)}(u(S_{t+1})-u(S_t))\right\}\\ &\geq \int_{x=0}^{u^*-c^*}\frac{u^*-x-c^*}{u^*-x}dx\\ &= \left(1-\frac{\log(a)}{a-1}\right)p^*.\end{aligned}$$

We prove that the same bound holds for every cost function with increasing increments whenever $u^*/c^*\leq e$, and, otherwise, establish that the following bound holds

$$p^{\mathcal{G}}\geq\left(1-\frac{a}{a-1}\frac{1}{e}\right)p^*.$$

Note that it is not possible to guarantee that for every cost function with increasing increments,

$$p(S)\geq\left(1-\frac{\log(a)}{a-1}+\varepsilon\right)p^*.$$

For instance, for the case of a budget constraint in (7), the greedy algorithm can guarantee at most $1-1/e$ whereas

$$\lim_{a\to\infty}\left\{1-\frac{\log(a)}{a-1}\right\}=1$$

which is a contradiction.

B Checking Conditions for Some Production Functions

One can easily check that all production functions from our catalogue examples are non-negative, monotone submodular set function. In this section, therefore, we show that conditions (M) and (B) hold for all production functions in the catalogue of examples in Section 3.

Total production: $f(x_S)=g(\sum_{i\in S}x_i)$ Condition (M) holds because $f(x,y)-f(x)=g(x+y)-g(x)$ and g has decreasing increments being a concave function. Condition (B) holds because $f^{-1}(x)=g^{-1}(x)$ where g is the inverse function of g, and hence $f(f^{-1}(f(x_1,\ldots,x_{l-1})),x_l)=g(g^{-1}(g(x_1+\cdots+x_{l-1}))+x_l)=g(x_1+\cdots+x_l)=f(x_1,\ldots,x_l)$.

Best-shot: $f(x_S) = \max_{i \in S} x_i$ Note that $f^{-1}(x) = x$. Condition (M) holds because $f(x,y) - x = \max\{x,y\} - x$, which is indeed decreasing in x, for every fixed value $y \in \mathbf{R}_+$. Condition (B) holds because $f(f^{-1}(f(x_1,\ldots,x_{l-1})),x_l) = \max\{\max\{x_1,\ldots,x_{l-1}\},x_l\} = \max\{x_1,\ldots,x_l\} = f(x_1,x_2,\ldots,x_l)$.

CES: $f(x_S) = (\sum_{i \in S} x_i^p)^{1/p}$**, for parameter** $p \geq 1$ Note that $f^{-1}(x) = x$. Condition (M) holds because $f(x,y) - x = (x^p + y^p)^{1/p} - x$, and hence,

$$\frac{\partial}{\partial x}(f(x,y) - x) = \left(\frac{x}{(x^p + y^p)^{1/p}}\right)^{p-1} - 1 \leq 0.$$

Condition (B) holds because $f(f^{-1}(f(x_1,\ldots,x_{l-1})),x_l) = (((x_1^p + \ldots + x_{l-1}^p)^{1/p})^p + x_l^p)^{1/p} = (x_1^p + \cdots + x_l^p)^{1/p} = f(x_1,\ldots,x_l)$.

Success-Probability: $f(x_S) = 1 - \prod_{i \in S}(1 - g(x_i))$ Condition (M) holds because $f(x,y) - f(x) = g(y)(1-g(x))$ and g is an increasing function. Condition (B) holds because $f(f^{-1}(f(x_1,\ldots,x_{l-1})),x_l) = 1 - \prod_{i=1}^{l}(1 - g(x_i)) = f(x_1,\ldots,x_l)$.

C Proof of Theorem 5

Without loss of generality, assume that $\mathbf{E}[X_1] \geq \mathbf{E}[X_2] \geq \cdots \geq \mathbf{E}[X_n]$. Let $S = \{i_1, i_2, \ldots, i_k\}$ be an arbitrary team. Then, we have

$$\begin{aligned}
u(S) &= \mathbf{E}[f(X_S)] \\
&= \mathbf{E}[(f(X_S) - f(X_{S\setminus\{i_k\}})) + (f(X_{S\setminus\{i_k\}} - f(X_{S\setminus\{i_{k-1},i_k\}})) + \cdots + (f(X_{\{i_1\}}) - f(0))] \\
&\leq \mathbf{E}[f(X_{\{i_k\}}) + f(X_{\{i_{k-1}\}}) + \cdots + f(X_{\{i_1\}})] \\
&= \sum_{i \in S} \mathbf{E}[X_i] \\
&\leq \sum_{i=1}^{k} \mathbf{E}[X_i] \qquad (21)
\end{aligned}$$

where the first inequality follows by the submodularity of function $u(S)$, the second inequality is by the assumption that individuals are enumerated in decreasing order of the mean test scores.

Now, observe that for every $(x_1, x_2, \ldots, x_k) \in \mathbf{R}_+^k$,

$$\frac{1}{k}\sum_{i=1}^{k} x_i = \frac{1}{k}\sum_{i=1}^{k}(x_i^p)^{1/p} \leq \left(\frac{1}{k}\sum_{i=1}^{k} x_i^p\right)^{1/p}$$

which is by Jensen's inquality. Hence,

$$\sum_{i=1}^{k} \mathbf{E}[X_i] \leq k^{1-1/p}\mathbf{E}\left[\left(\sum_{i=1}^{k} X_i^p\right)^{1/p}\right].$$

Combining with (21), it follows that for every $S \subseteq N$ such that $|S| = k$,

$$\mathbf{E}\left[\left(\sum_{i=1}^{k} X_i^p\right)^{1/p}\right] \geq \frac{1}{k^{1-1/p}}\mathbf{E}\left[\left(\sum_{i\in S} X_i^p\right)^{1/p}\right].$$

The tightness can be established as follows. Let N consists of two subsets of individuals M and R, where M consists of k individuals whose each individual performance is of value $1+\epsilon$ with probability 1, for a parameter $\epsilon > 0$, and R consists of k individuals whose each individual performance is of value a with probability $1/a$ and of value 0 otherwise, for parameter $a \geq 1$. Then, we note that

$$u(M) = k^{1/p}(1+\epsilon)$$

and

$$\begin{aligned} u(S^*) \geq u(R) &= \mathbf{E}\left[\left(\sum_{i\in R} X_i^p\right)^{1/p}\right] \\ &\geq a\mathbf{Pr}\left[\sum_{i\in R} X_i > 0\right] \\ &= a\left(1-\left(1-\frac{1}{a}\right)^k\right) \\ &\geq a\left(1-e^{-k/a}\right). \end{aligned}$$

Hence, it follows that

$$\frac{u(M)}{u(S^*)} \leq (1+\epsilon)\frac{1}{k^{1-1/p}}\frac{k/a}{1-e^{-k/a}}.$$

The tightness claim follows by taking a such that $k = o(a)$, so that $(k/a)/(1-e^{-k/a}) = 1 + o(1)$.

D Proof of Theorem 6

D.1 Proof of Claim 1

If k is a constant, there is no p satisfying both conditions $p = o(1)$ and $p > 1$. Hence, it suffices to consider $k = \omega(1)$ and show that the following statement holds: for any given $\theta > 0$, there exists an instance for which the quantile test-score based team selection cannot give a constant-factor approximation.

Consider the following distributions for X_i:

1. Let each X_i be equal to a with probability 1 for $1 \leq i \leq k$. Then, each quantile test-score is equal to a and each replication test-score is equal to $ak^{1/p}$.

2. Let each X_i be equal to 0 with probability $1-1/n$, and equal to $b\theta n/k$ with probability $1/n$ for $k+1 \leq i \leq 2k$. Then, in the limit as n grows large, each quantile test-score is equal to b and each replication test score is equal to $b\theta$.

3. Let each X_i be equal to 0 with probability $1 - \theta/k$ and equal to c with probability θ/k for $2k + 1 \leq i \leq 3k$. Then, each quantile test-score is equal to c and each replication test-score is less than or equal to $c\theta^{1/p}$.

4. Let X_i be equal to 0 for $3k + 1 \leq i \leq n$.

If θ is a constant (i.e., $\theta = O(1)$), we can easily check that the quantile test-score algorithm cannot give a constant-factor approximation with $a = b = 1$ and $c = 2$. Under this condition, the set of individuals $\{2k + 1, \ldots, 3k\}$ is selected by the quantile test-score algorithm. However,

$$\begin{aligned}
\frac{\mathbf{E}\left[\left(\sum_{i=2k+1}^{3k} X_i^p\right)^{1/p}\right]}{\mathbf{E}\left[\left(\sum_{i=1}^{k} X_i^p\right)^{1/p}\right]} &= \frac{\mathbf{E}\left[\left(\sum_{i=2k+1}^{3k} X_i^p\right)^{1/p}\right]}{k^{1/p}} \\
&\leq \frac{\left(\sum_{i=2k+1}^{3k} \mathbf{E}\left[X_i^p\right]\right)^{1/p}}{k^{1/p}} \\
&= 2\left(\frac{\theta}{k}\right)^{1/p} = o(1),
\end{aligned}$$

since $k = \omega(1)$, $\theta = O(1)$, and $p = o(\log(k))$.

If θ goes to infinity as n goes to infinity (i.e., $\theta = \omega(1)$), we have

$$\begin{aligned}
\frac{\mathbf{E}\left[\left(\sum_{i=2k+1}^{3k} X_i^p\right)^{1/p}\right]}{\mathbf{E}\left[\left(\sum_{i=k+1}^{2k} X_i^p\right)^{1/p}\right]} &\leq \frac{\left(\sum_{i=2k+1}^{3k} \mathbf{E}\left[X_i^p\right]\right)^{1/p}}{\theta} \\
&= 2\theta^{(1-p)/p} = o(1),
\end{aligned}$$

because $p > 1$. Therefore, the quantile test-score based team selection has a negligible utility compared to the optimal utility.

D.2 Proof of Claim 2

Let $T(X_S)$ be a subset of S such that $i \in T(X_S)$ if, and only if, $X_i \geq F_i^{-1}(1 - h/k)$, for $i \in S$. Let $s_{\max} = \max_{i \in S} s_i$ and $s_{\min} = \min_{i \in S} s_i$. In this proof, we will show that there exist constants c_1 and c_2 such that

$$c_1 s_{\min} \leq \mathbf{E}\left[\left(\sum_{i \in S} X_i^p\right)^{1/p}\right] \leq c_2 s_{\max}.$$

Since $(x + y)^{1/p} \leq x^{1/p} + y^{1/p}$ when $x, y \geq 0$ and $p > 1$,

$$\begin{aligned}
\mathbf{E}\left[\left(\sum_{i \in S} X_i^p\right)^{1/p}\right] &= \mathbf{E}\left[\left(\sum_{i \in T(X_S)} X_i^p + \sum_{i \in S \setminus T(X_S)} X_i^p\right)^{1/p}\right] \\
&\leq \mathbf{E}\left[\left(\sum_{i \in T(X_S)} X_i^p\right)^{1/p} + \left(\sum_{i \in S \setminus T(X_S)} X_i^p\right)^{1/p}\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbf{E}\left[\sum_{i\in T(X_S)} X_i + \left(\sum_{i\in S\setminus T(X_S)} X_i^p\right)^{1/p}\right] \\
&\leq \mathbf{E}\left[\sum_{i\in T(X_S)} X_i + \left(\sum_{i\in S\setminus T(X_S)} s_{\max}^p\right)^{1/p}\right] \\
&\leq \left(\mathbf{E}\left[|T(X_S)|\right] + k^{1/p}\right) s_{\max} \;=\; (1+k^{1/p}) s_{\max}.
\end{aligned}
$$

By the Minkowski inequality, $\left(\sum_{i\in A} \mathbf{E}\left[X_i\right]^p\right)^{1/p} \leq \mathbf{E}\left[\left(\sum_{i\in A} X_i^p\right)^{1/p}\right]$ for all $A \subseteq S$. Thus, we have

$$
\begin{aligned}
\mathbf{E}\left[\left(\sum_{i\in S} X_i^p\right)^{1/p}\right] &= \mathbf{E}\left[\left(\sum_{i\in T(X_S)} X_i^p + \sum_{i\in S\setminus T(X_S)} X_i^p\right)^{1/p}\right] \\
&\geq \mathbf{E}\left[\left(\sum_{i\in T(X_S)} X_i^p\right)^{1/p}\right] \\
&= \sum_{A\subset S} \mathbf{Pr}\{T(X_S)=A\}\mathbf{E}\left[\left(\sum_{i\in A} X_i^p\right)^{1/p} \middle| T(X_S)=A\right] \\
&\geq \sum_{A\subset S} \mathbf{Pr}\{T(X_S)=A\}\left(\sum_{i\in A} \mathbf{E}[X_i | i\in T(X_S)]^p\right)^{1/p} \\
&\geq \sum_{A\subset S} \mathbf{Pr}\{T(X_S)=A\}|A|^{1/p} s_{\min} \\
&\geq \left(1-(1-1/k)^k\right) s_{\min} \;\geq\; (1-1/e) s_{\min}.
\end{aligned}
$$

Therefore, the quantile test-score team selection is a constant-factor approximation algorithm.