CHARACTERIZING THE PERFORMANCE OF
PROCESS FLEXIBILITY STRUCTURES

O. Zeynep Akşin*
and
Fikri Karaesmen†

* Graduate School of Business
Koç University
34450, Istanbul TURKEY

† Department of Industrial Engineering
Koç University
34450, Istanbul TURKEY

zaksin@ku.edu.tr, fkaraesmen@ku.edu.tr

July 2004
Characterizing the Performance of Process Flexibility Structures

O. Zeynep Akşin ∗ Fikri Karaesmen †

July 2004

Abstract

Most service systems consist of multi-departmental structures corresponding to multiple types of service requests, with possibly multi-skill agents that can deal with several types of service requests. The design of flexibility in terms of agents’ skill sets and assignments of requests is a critical issue for such systems. A similar problem exists in the manufacturing of multiple products in multiple plants. For such systems, our objective is to identify preferred flexibility structures when demand is random and capacity is finite. Considering a network flow type model as the basis of our analysis, we identify general structural properties of flexibility design pertaining to the marginal values of flexibility and capacity.

1 Introduction

This paper considers service systems with multi-departmental structures having possibly multi-skill servers that treat several types of service requests. In any such system, it is possible to have a different mix of skill sets with a different number of servers belonging to each skill set. It is well known that more flexibility leads to better operational performance. However given that there are costs associated with creating and maintaining this flexibility, and difficulties managing the resulting more complex system, it is desirable to understand the value of this flexibility in more depth. This paper will focus on providing a better understanding of the relationship between different flexibility structures and value. The following questions are relevant in this setting: How many skills should servers have (how much flexibility)? What are the ideal skill-sets for

∗Graduate School of Business, Koç University
†Dept. of Ind. Eng., Koç University
those that are cross-trained \textit{(what type of flexibility)}? How should these skill-sets be formed in a multi-departmental structure, where each server has a primary skill and some secondary skills \textit{(where)}? This set of questions motivate our research and will be labeled as the flexibility design problem in the ensuing analysis. We provide guidelines that will be useful in addressing such a flexibility design problem.

A well known application of this flexibility design problem in a manufacturing setting is the problem studied by Jordan and Graves [8]. In this setting, the departments are different plants or production lines, while the customer types represent different products to be produced in these production facilities. Process flexibility constitutes the ability of producing a product in multiple plants or production lines. The model that we study in Section 3 is identical to the one in Jordan and Graves [8]. Using this model as a basis, we formalize some results pertaining to the performance of different flexible structures that were observed numerically in [8].

The remaining parts of this paper are organized as follows. Related literature is reviewed in the next section. Section 3 introduces the model and the problem. Section 4 presents the results on the diminishing returns property of flexibility. The results on flexibility/capacity interactions are presented in Section 5. Finally, the concluding remarks can be found in Section 6.

\section{Literature Review}

The importance of flexibility in service delivery is well known. A significant source of service delivery process flexibility comes from the use of cross-trained servers. While the practice itself is widespread, there is little formal evaluation of the value of this type of practice from an operations standpoint. Pinker and Shumsky [11] consider trade-offs between capacity and quality for cross-trained workers in service systems. Numerous studies in manufacturing have looked at the case of flexible workers and their impact on performance in terms of operational measures like throughput. Most of these studies analyze specific work-sharing schemes in queueing network models (Van Oyen et al. [14], and references therein). Karaesmen et al. [9] investigate flexibility in the context of field service design. These papers assess the value of certain workforce flexibility practices in given settings, however do not tackle the broader question of designing the type of flexibility in these systems.

More generally, the benefits and design of flexibility in operations have been studied ex-
tensively (DeGroote, [2]; Sethi and Sethi, [12]). An important stream consists of papers that address the capacity investment problem in the presence of flexible resources (Fine and Freund [3]; Van Mieghem [15]; Netessine et al. [10]). These papers assume a certain form of flexibility and then explore the question of the ideal level of this flexibility and how it relates to value under uncertain demand.

In this paper, we consider capacity to be fixed and explore the relationship between different flexibility structures and value, without explicitly addressing the optimal capacity issue. In this regard, our analysis parallels that in Jordan and Graves [8]. Focusing on process flexibility, Jordan and Graves explore the problem of assigning multiple products to multiple plants, where the flexibility of the plants determines which products they can handle. The authors illustrate that well designed limited flexibility is almost as good as full flexibility. To address the question of where flexibility should be added in a system, the authors define a \textit{chain} structure as a group of directly or indirectly connected group of products and plants. It is shown that a structure that enables the formation of fewer long chains is superior to one with multiple short chains. The principles are illustrated with simulations. Graves and Tomlin [5] perform a similar analysis for multi-stage systems and develop a flexibility measure for such systems. Hopp et al. [7] explore the benefits of chaining in the context of cross-training for production lines. Finally, Gurumurthi and Benjaafar [6] present a numerical investigation of the benefits of chaining based on a queueing model.

The objective of this paper is to represent flexibility structures through a network flow model as in Jordan and Graves [8] and formalize earlier observations on flexibility using this model. By doing this, we will establish certain flexibility design principles for service and manufacturing systems. In particular, we focus on two sets of design issues: diminishing returns to increased flexibility, and flexibility/capacity interactions.

3 Modeling Process Flexibility

Consider a service system with multiple customer types. Customer types differ in terms of their service requirements. Servers specialize by customer type, but can be flexible with overlapping skill sets, allowing them to treat customer requests from different types. The service system can be represented as a directed graph $G = (N, A)$ with a set of nodes $N$, of which one is a source node and one a sink node, and a set of arcs $A$ whose elements are ordered pairs of distinct
nodes. Some standard definitions are useful to formalize the description of this network. A directed arc \((i, j)\) emanating from node \(i\) is said to have tail \(i\), terminating in node \(j\) known as the head of the arc. For an arc \((i, j) \in A\), the node \(j\) is said to be adjacent to node \(i\). The node adjacency list \(A(i)\) is the set of adjacent nodes, \(A(i) = \{j \in N : (i, j) \in A\}\). The indegree of a node is the number of incoming arcs of that node and its outdegree is the number of outgoing arcs.

An instance of a network that represents the service system is depicted in Figure 3. This graph illustrates a system with \(n\) customer types given by the set of nodes \(I = \{1, 2, \ldots, n\}\), served by servers in \(n\) departments, given by the set of nodes \(J = \{n + 1, n + 2, \ldots, 2n\}\). Note that since servers are assumed to be organized by their primary skills, the number of customer types is equal to the number of departments. The case where the number of customer types is larger than the number of departments can also be treated within this framework, where the additional classes can be served by dummy departments with no servers in them. The arcs emanating from the source node \(s\) and terminating in nodes \(i \in I\) represent the service demand, and have capacity given by the demand vector \(\lambda = (\lambda_1, \ldots, \lambda_n)\). This vector represents the realization of demand for a given period. The arcs emanating from nodes \(j \in J\) and terminating in the sink node \(z\) represent the capacity of each department. These arcs have a capacity given by the vector \(C = (C_1, \ldots, C_n)\). The arcs \((i, j)\) with \(i \in I\) and \(j \in J\) represent the flexibility of the system. Whenever a customer of type \(i \in I\) can be served by a server of

Figure 1: An \(n\) class service system with full flexibility
type $j \in J$, an arc $(i, j)$ with infinite capacity is added to the network. The network in Figure 3 illustrates a case where all customers can be treated by all servers, i.e. where the system has full flexibility. In a system with $n$ departments, full flexibility implies that each node $i \in I$ has outdegree equal to $n$. In general, the outdegree of node $i \in I$ represents the number of possible routings for customers of type $i$, and the indegree of a node $j \in J$ represents the number of skills a server of type $j$ has. Assuming that each customer request of type $i \in I$ is worth $r$ to the system, the problem of maximizing the value generated by a given configuration for a demand realization $\lambda$ is equivalent to the maximum flow problem for this network. We refer to the maximal flow as the throughput and denote it by $T(\lambda, C)$. A closely related problem, first studied by Jordan and Graves [8], considers a random demand vector $\lambda = (\lambda_1, ..., \lambda_n)$ where the performance measure of interest is the expected throughput $E[T(\lambda, C)]$. In this case, a maximum flow problem is solved for each realization of the random demand vector $\lambda$ and the expectation is taken over all realizations of $\lambda$. Jordan and Graves also discuss the relevance of the expected throughput maximization objective and relate it to a number of other possible objectives.

As in Jordan and Graves [8], the emphasis will be on a special class of processing network which is defined below:

**Definition 1** The flexibility of a network is defined as the set of arcs $F$ between the node set $I$ and the node set $J$.

**Definition 2** A symmetric network is defined as a network where i) every customer type can be processed by the same number of server types (departments), i.e. each node $i \in I$ has the same outdegree and ii) every department treats the same number of customer types, i.e. every node $j \in J$ has the same indegree. The flexibility of each symmetric network is denoted by $F_k$ with $k = 1, 2, ..., n$, where $k$ indexes the number of server types that a customer type can be served by (which is equal to the outdegree of the nodes $i \in I$), or equivalently the number of customer types that a server type can treat (which is equal to the indegree of the nodes $j \in J$). A network with flexibility $F_k$ is constructed such that for each node $i \in I$, there is an arc joining it to the nodes $j = n+i, n+i+1, ..., n+i+k-1$, where the labeling of the nodes is such that whenever $j > 2n$ we take $j - n$. Note that, for a fixed value of $n$, a higher index $k$ denotes more flexibility (i.e. $F_{k-1} \subset F_k$), with $F_1$ representing the case of specialized servers and $F_n$ the case of fully flexible servers. For $1 < k < n$, $F_k$s are constructed such that the network is fully connected and contains a cycle.
Throughout, the maximum flow of a network $G$ is denoted by $T_G(\mathcal{F}, \lambda, C)$. For notational compactness, these are replaced by $T_G$ and $T(\mathcal{F})$ whenever possible.

4 Diminishing Returns to Increased Flexibility

It is well known that the performance of a service system with limited flexibility rapidly approaches that of a system with full flexibility. Numerical examples in a number of articles (e.g. Jordan and Graves [8], Karaesmen et al. [9]) support this point. This issue will be explored first by investigating structural properties of the maximum flow as a function of $F$.

Let $u_{ij}$ be the capacity of arc $(i, j)$ (where $i \in I$ and $j \in J$). Note that $u_{ij} = \infty$ if $(i, j) \in \mathcal{F}$ and $u_{ij} = 0$ otherwise. Let $\mathcal{U}$ be the corresponding set of arc capacities for all $i \in I$ and $j \in J$: $\mathcal{U} = \{u_{ij}\}$. The following proposition establishes a property of the throughput as a function of the capacity set $\mathcal{U}$. Note that, for any flexible arc set $\mathcal{F}$, $(\mathcal{F}, \lambda, C)$ and $(\mathcal{F}_n, \lambda, C, \mathcal{U})$ are equivalent representations in terms of the maximum flow problem.

**Theorem 1**: The throughput (maximum flow) of the network, $T_G(\mathcal{F}_n, \lambda, C, \mathcal{U})$ is submodular in $\mathcal{U}$.

**Proof**: $\mathcal{U}$ consists of the capacities of all arcs in $\mathcal{F}_n$. All arcs in $\mathcal{F}_n$ are parallel (i.e. there is no simple undirected cycle in the graph in which two of these arcs have the same direction). Gale and Politoff [4] show that for any pair of parallel arcs $((i, j), (i', j'))$ the throughput has decreasing differences in $(u_{ij}, u_{i'j'})$. By Theorem 2.6.2 and Corollary 2.6.1 in Topkis [13], the throughput is then submodular in $\mathcal{U}$. \qed

The theorem can also be stated in terms of the set $\mathcal{F}$ using the fact that the addition of a new arc $(i, j)$ to the existing set of arcs $\mathcal{F}$ is equivalent to setting $u_{ij} = \infty$. The equivalent statement to that of Theorem 1 would then be: the throughput (maximum flow) of the network is submodular in the arc set $\mathcal{F}$.

Finally, let us consider the case where $\lambda$ is a random vector.

**Corollary 1**: The expected throughput of the network is submodular in the flexible capacity set $\mathcal{U}$ (or equivalently is submodular in $\mathcal{F}$).
Proof: For any realization of the random vector $\lambda$, the throughput is submodular in $U$ by Theorem 1. Because submodularity is preserved under the expected value operation (Topkis, [13]), the expected throughput is also submodular in $U$. 

Submodularity of the expected throughput in terms of the arc set $F$ reflects one side of the diminishing returns property: marginal additions of flexibility (in terms of arcs) are relatively more beneficial in terms of throughput than larger (joint) additions. In general, however, the diminishing returns property is not interpreted in terms of submodularity, but in terms of concavity of the throughput as flexibility increases. Next, we investigate the concavity properties.

Proposition 1: If $C = (C, ..., C)$ and $\lambda_i (i = 1, 2, ..., n)$ are independent and identically distributed random variables, then the expected throughput of the network as a function of $F_k$ is concave in $k$ (for $k = 2, ..., n - 1$).

Proof: Consider the network $F_{k-1}$, with the flexible arc set $\{(1, n + k + 1), (2, n + 1), ..., (n, n + k + 2)\}$. $F_k = F_{k-1} \cup \{(1, n + k + 2), (2, n + k + 2), ..., (n, n + k + 2)\}$. Now consider the following set: $F'_k = F_{k-1} \cup \{(1, 2n), (2, n + 1), ..., (n, 1)\}$. By Theorem 1 and Corollary 1 we obtain:

$$E[T(F_k \cup F'_k)] + E[T(F_k \cap F'_k)] \leq E[T(F_k)] + E[T(F'_k)].$$  \hspace{1cm} (1)

Now note that, by construction: $F_k \cap F'_k = F_{k-1}$ and $F_{\|} \cup F'_\| = F'_{\|+\infty}$ where $F'_{k+1}$ has identical structure to $F_{k+1}$ by a relabeling of the nodes. Similarly $F'_k$ and $F_k$ have identical structures which implies that: $E[T(F'_k) = E[T(F_k)]$ and $E[T(F'_{k+1}) = E[T(F_{k+1})]$. Using these equalities, inequality (1) can be expressed as:

$$E[T(F_{k+1})] + E[T(F_k)] \leq 2E[T(F_k)]$$

which is the desired result.

Proposition 1 proves the concavity property that is observed in the numerical examples of Jordan and Graves [8] for independent and identically distributed demands. On the other hand, all numerical results seem to indicate that the above concavity holds under even weaker assumptions on the demand distributions. Unfortunately, a general proof under weaker assumptions...
eludes us. Nevertheless, the next proposition establishes that, in the special case of a 3 by 3 network, concavity of the expected throughput (in terms of the flexibility index) holds for any joint demand distribution.

**Proposition 2**: If $C = (C, C, C)$ and $\lambda_i$ ($i = 1, 2, 3$) are jointly distributed random variables, then the expected throughput of the network has the following diminishing returns property:

$$E[T(F_3)] - E[T(F_2)] \leq E[T(F_2)] - E[T(F_1)]$$

**Proof**: We only provide a sketch of the proof which is straightforward but tedious. For any realization of demand, it can be shown that the throughput has the desired concavity property by performing a case by case comparison of the corresponding minimum cuts in the network. A stochastic coupling argument then ensures that the property also holds under the expected value operator. The complete proof can be found in Aksin and Karaesmen [1].

5 Some Properties on Flexibility/Capacity Interactions

A different way of improving performance in such a system is by adding capacity. For example, one may choose to invest in additional capacity rather than investing in additional flexibility. The close interactions between flexibility and capacity are evident. The following result formalizes this relationship:

**Theorem 2** Consider a symmetric service network with demand vector $\lambda$ and symmetric capacity vector $C$. If $\sum_{i=1}^{n} \lambda_i > \sum_{i=1}^{n} C_i$ then $T(F, \lambda, C + \Delta C) - T(F_k, \lambda, C) \geq T(F_{k-1}, \lambda, C + \Delta C) - T(F_{k-1}, \lambda, C)$ for small enough $\Delta C$ such that one still has $\sum_{i=1}^{n} \lambda_i \geq \sum_{i=1}^{n} C_i + \Delta_i C$.

**Proof**: For the case when $\sum_{i=1}^{n} \lambda_i \geq \sum_{i=1}^{n} C_i$, first note the following equivalence. For any network $G$ with demand vector $\lambda$, flexibility $F_i$ and capacity vector $C$, a corresponding network $G'$ with demand vector $\lambda' > \lambda$, flexibility $F_{i-1}$ and capacity vector $C$ can be constructed such that $T_G(\lambda, C) = T_{G'}(\lambda', C)$ or equivalently $T(F_k, \lambda, C) = T(F_{k-1}, \lambda', C)$. More precisely, this construction can be performed by adding a vector $\Delta^\lambda$ to the original demand vector $\lambda$, that ensures the equivalence in the maximum flows, with $\Delta^\lambda_i > 0$ only if $i$ represents a department with excess capacity (i.e. $C_i > \lambda_i$). Note that if there is no department with excess capacity, the maximum throughput will be equal to the sum of the capacities and cannot be improved via a
change in flexibility only. Making this transformation, the desired condition in the Theorem can be restated as \( T(F_{k-1}, \lambda' + \Delta C, C + \Delta C) - T(F_{k-1}, \lambda', C) \geq T(F_{k-1}, \lambda, C + \Delta C) - T(F_{k-1}, \lambda, C) \).

Consider the following inequality, which is obtained by replacing \( \lambda' + \Delta C \) with \( \lambda' \).

\[
T(F_{k-1}, \lambda', C + \Delta C) - T(F_{k-1}, \lambda', C) \geq T(F_{k-1}, \lambda, C + \Delta C) - T(F_{k-1}, \lambda, C).
\] (2)

Since \( T \) is non-decreasing in \( \lambda \) showing that this inequality holds is sufficient to show the desired result. It is known that the optimal value of the objective function in the minimum cut problem \( T(F_i, \lambda, C) \) is submodular in \( (\lambda, -C) \) (Theorem 3.7.1 in Topkis, [13]). Combining this with the fact that for a function \( f(x) \) that is submodular in \( x \), \( -f(x) \) is supermodular in \( x \), one can state that \( T(F_i, \lambda, C) \) is supermodular in \( (-\lambda, C) \). Then the desired relationship in the inequality (2) holds by the definition of supermodularity, which completes the proof.

Theorem 2 demonstrates that flexibility and capacity are complements up to a certain threshold, parameterized by the demand and capacity vectors. Note that this threshold is equal to \( T(F_n, \lambda, C) \), the throughput of the fully flexible system (recalling that \( T(F_n, \lambda, C) = \min(\sum_{i=1}^{n} \lambda_i, \sum_{i=1}^{n} C_i) \)). Beyond the threshold, capacity may act as a substitute to flexibility.

For the case with random demand, this result cannot be stated except in the special case of 'chronically overloaded' systems where total demand always exceeds total capacity for all possible demand realizations. Informally, for heavily utilized systems, which are also those systems where system flexibility is sought most, one will mostly be in the former region with capacity and flexibility acting as complements. For these types of systems, the result suggests that an additional server is more valuable in the system with superior flexibility. Thus flexibility and capacity should be jointly designed. For systems with lower utilization, this result may be reversed, and an additional person can be worth more in a system with less flexibility. The high flexibility systems are already able to respond to most of their demand, thus the marginal value of capacity is low, becoming zero for the full flexibility system. On the other hand, the low flexibility systems can still improve performance with additional capacity, leading to a positive marginal value for capacity.

So far, we have shown basic features of flexibility that would be useful in answering the how much flexibility type of question. Next, the theory of majorization (Marshall and Olkin 1979) is used to explore the type of flexibility, thereby further refining the notion of smart limited flexibility.

**Definition 3** For a vector \( x \in \mathbb{R}^n \), let \([i] \) denote a permutation of the indices \( \{1, 2, ..., n\} \) such
that $x[1] \geq x[2] \geq \ldots \geq x[n]$. Then, for $x, y \in \mathbb{R}^n$, $x$ is said to be majorized by $y$, $x \prec y$, if \( \sum_{i=1}^{n} x[i] = \sum_{i=1}^{n} y[i] \) and for all $k = 1, \ldots, n-1$, \( \sum_{i=1}^{k} x[i] \leq \sum_{i=1}^{k} y[i] \).

Let $i_d(J) \in \mathbb{R}^n$ be the vector of indegrees for $j \in J$, and $o_d(I) \in \mathbb{R}^n$ be the vector of outdegrees for $i \in I$. Recall that the former represents a vector with the number of skills of the servers in each department, and the latter the number of possible routings for each customer class. The following definitions are proposed for the graph $G = (N, A)$.

\textbf{Definition 4} If $i_d(J)$ has $i_d(n+1) = i_d(n+2) = \ldots = i_d(2n)$, the service system is said to have balanced skill diversity. Symmetrically, if $o_d(I)$ has $o_d(1) = o_d(2) = \ldots = o_d(n)$, the system is said to have balanced routings. For two networks $G = (N, A)$ and $G'(N, A')$, and skill diversity vectors $i_d(J)$ and $i'_d(J)$ (routing vectors $o_d(I)$ and $o'_d(I)$), whenever $i_d(J) < i'_d(J)$ ($o_d(I) < o'_d(I)$) the system $G(N, A)$ is said to have more balanced skill diversity (more balanced routings) than $G'(N, A')$.

\textbf{Theorem 3} Consider a network $G(N, A)$ with demand vector $\lambda = (\lambda_1, \ldots, \lambda_n)$, symmetric capacity vector $C = (C, \ldots, C)$, and routing vector $o_d(I)$. Whenever $G(N, A)$ does not have balanced skill diversity, one can find $G'(N', A')$ with demand vector $\lambda = (\lambda_1, \ldots, \lambda_n)$, symmetric capacity vector $C = (C, \ldots, C)$, routing vector $o'_d(I') = o_d(I)$, and $i'_d(J') < i_d(J)$ such that $T_{G'} \geq T_{G}$.

\textbf{Proof:} The proof requires some additional definitions and notation and is provided in the Appendix.

\textbf{Corollary 2} Consider a network $G(N, A)$ with symmetric demand vector $\lambda = (\lambda, \ldots, \lambda)$, capacity vector $C = (C_1, \ldots, C_n)$, and skill diversity vector $i_d(J)$. Whenever $G(N, A)$ does not have balanced routing, one can find $G'(N', A')$ with symmetric demand vector $\lambda = (\lambda, \ldots, \lambda)$, capacity vector $C = (C_1, \ldots, C_n)$, skill diversity vector $i'_d(J') = i_d(J)$, and $o'_d(I') < o_d(I)$ such that $T_{G'} \geq T_{G}$.

\textbf{Proof:} The proof can be found in the Appendix.

The results in Theorem 3 and Corollary 2 continue to hold for the expected throughput when the demand is random. They formalize similar guidelines suggested by Jordan and Graves, that
recommend “equalizing the number of plants (measured in total units of capacity) to which each product in the chain is connected” and “equalizing the number of products (measured in total units of expected demand) to which each plant in the chain is directly connected”.

6 Concluding Remarks

We investigated the flexibility design problem through using a network flow framework and provided formal proofs of a number of properties that were previously suggested by numerical experiments. In addition, the formalization also enabled us to obtain some new properties on the effects of increased flexibility and its interactions with capacity.

On the effects of increased flexibility, a submodularity property was established. This implies that all marginal flexibility improvements are substitutes of each other; thereby suggesting the effectiveness of limited smart flexibility. We also showed that, for a particular class of networks frequently investigated in the operations literature, the expected throughput of the network manifests the diminishing returns property as flexibility increases.

On flexibility/capacity interactions, our results suggest that flexibility and capacity may be complements or substitutes depending on the demand and capacity. In particular, if capacity is scarce, more flexible systems benefit more from capacity increases than less flexible systems. In addition, as sometimes suggested in the literature, balanced skill or routing diversity improves performance.

In parallel work, we investigate the relevance of the guidelines obtained on stochastic processing networks taking into account the dynamic aspects of the problem.

References


A Proofs

The following additional notation and definitions are used in the proof Theorem 3: For a set of nodes $I$ and $J$, $(I, J) = \{(i, j) : i \in I, j \in J\}$ represents the set of all arcs between $I$ and $J$. The capacity of an arc $(i, j)$ is given by the capacity function $c(i, j)$. For subsets $I$ and $J$ of $N$, denote the sum of all capacities on all arcs $(I, J)$ by $c(I, J) = \sum_{i \in I, j \in J} c(i, j)$. Let $X$ be a subset of $N$. Then $X$ is a cut if it contains the source but not the sink of the network. The cut capacity function is given by $f(X) = c(X, N \setminus X)$. For $X$ being a cut, the minimum of $f(X)$ over all $X$ is known as a minimum cut.

The following lemma is used in the proof of Theorem 3.

Lemma 1 For the graph $G = (N, A)$ representing the flexible service system, any cut that satisfies at least one of the conditions below cannot be a minimum cut: i) There exists an $i \in X$ with $j \in J$ and $j \not\in X$ ii) $f(X) > \min(\sum_{i=1}^{n} \lambda_i, \sum_{i=1}^{n} C_i)$.

Proof: Any cut $X$ that satisfies i) will have $f(X) = \infty$. Since there exists cuts with finite capacity, $X$ cannot be the minimum cut of this network. To show that any cut $X$ that satisfies ii) cannot be a minimum cut, note that both $\sum_{i=1}^{n} \lambda_i$ and $\sum_{i=1}^{n} C_i$ are cuts of this network. □

A.1 Proof of Theorem 3

Take any network, characterized by the graph $G(N, A)$. Let $NA_G$ denote the set of cuts of this network, which cannot be eliminated by one of the rules in Lemma 1. This set will be called the set of uneliminated cuts. By Lemma 1, the minimum cut of the network $G$ is a cut in $NA_G$. Now consider a second network $G'(N, A')$, where an arc $(a, b)$ has been replaced by arc $(a, b')$ and everything else is the same. The nodes $b \in J$ and $b' \in J'$ are chosen such that $i_d(J') < i_d(J)$. Then according to Definition 4, the network $G'$ is said to have more balanced skill diversity. It is next shown that $G'$ thus obtained has maximum flow $T_{G'} \geq T_G$.

For any cut $X \in NA_G$ let $X_I$ denote the nodes of $X$ that are in the set $I$, i.e. $X_I = \{x \in X : x \in I\}$. Recall that $A(X_I)$ denotes the set of adjacent nodes to the nodes in $X_I$. By Lemma 1, all $y \in A(X_I)$ are also in $X$, i.e. $y \in X$. $NA_{G'}$ can be obtained from $NA_G$, by noting that the cuts in $NA_G$ fall into three distinct sets: 1) The set of cuts that do not change as a result
of the arc replacement being considered. 2) The set of cuts that change, however have the same value as before. 3) The set of cuts that are eliminated by Lemma 1 in the new network. In addition, there may be some new cuts.

More precisely, the cuts in \( NA_G \) can be grouped as follows. 1) All cuts \( X \in NA_G \) such that \( a \notin X \) belong to the first group, since the proposed change in the network only impacts \( A(a) \). All such cuts will also be in \( NA_{G'} \). 2) The cuts in the second group are those that are obtained from a cut \( X \in NA_G \) by replacing the node \( b \) in the cut by \( b' \), i.e. \( X' = X - b + b' \), such that for all \( x \in X'_I \), \( A(X'_I) \subseteq X' \). Thus, these cuts cannot be eliminated by Lemma 1. Note that for this group of cuts, \( f(X) = f(X') \) by symmetry of the capacity vector \( C \). 3) Consider a cut \( X \) with nodes \( x_1 \in X_I \) and \( x_2 \in X_I \), and \( b \in \{A(x_1) \cap A(x_2)\} \). If for this cut \( b' \notin \{A(x_1) \cup A(x_2)\} \), then one will have \( A'(X'_I) \supset A(X_I) \) in the new network \( G' \). But then all of these cuts will be eliminated by condition i) of Lemma 1. 4) The argument for the third group of cuts shows that there may be some new cuts \( X' \) in \( NA_{G'} \) with \( x_1, x_2 \in X' \) and \( A'(X'_I) \subseteq X' \). In other words, these cuts \( X' \) contain all the nodes of cuts \( X \) that are in group three above, and in addition also contain node \( b' \). Note that for these additional cuts \( f(X) + C = f(X') \).

Characterizing the cuts of the new network, using the set of uneliminated cuts for the initial network, one observes that some cuts of \( G \) are eliminated in \( G' \), while those that are not eliminated preserve the same value \( f(X) = f(X') \). All additional cuts that can be added to \( G' \) without being eliminated by Lemma 1 are shown to have \( f(X') > f(X) \) for some \( X \in NA_G, \notin NA_{G'} \). Thus one has that the minimum cut of \( G' \) is greater than or equal to the minimum cut of \( G \), which implies that \( T_{G'} \geq T_G \). The same argument can be repeated for another arc change that induces more balanced skill sets. Thus any network \( G' \) can be obtained from a network \( G \) through a finite number of arc changes, each improving throughput. This proves the result. □

A.2 Proof of Corollary 2

Note that the service networks represented by graphs \( G(N, A) \) are fully symmetric in \( \lambda \) and \( C \). In other words, a network where the flow is from the sink node towards the source node, and where the capacity vector \( C \) has been replaced by the demand vector \( \lambda \) and vice versa, has identical maximum flow with the original network. These latter types of networks can be labeled as the reversed networks. Observe, furthermore, that in the reversed network, all results previously shown for skill sets hold, and these are equivalent to results in terms of routings in
the original network. Using this equivalence, and noting that the \( \lambda \) and \( C \) vectors in the corollary ensure that the reversed network has the same characteristics as the original network (any demand vector, symmetric capacity vector), the result stated in the Corollary follows by the proof for Theorem 3. \( \Box \)