European Journal of Operational Research 207 (2010) 711–724

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Production, Manufacturing and Logistics

On range and response: Dimensions of process flexibility

Mabel C. Chou [a], Geoffrey A. Chua [b], Chung-Piaw Teo [a,*]

[a] Department of Decision Sciences, NUS Business School, Singapore
[b] Nanyang Business School, Nanyang Techological University, Singapore

ARTICLE INFO

Article history:
Received 24 March 2009
Accepted 24 May 2010
Available online 2 June 2010

Keywords:
Probability: renewal processes
Production: process flexibility
Facility planning: design

ABSTRACT

There are two dimensions to process flexibility: range versus response. Range is the extent to which a system can adapt, while response is the rate at which the system can adapt. Although both dimensions are important, the existing literature does not analytically examine the response dimension vis-a-vis the range dimension.

In this paper, we model the response dimension in terms of uniformity of production cost. We distinguish between primary and secondary production where the latter is more expensive. We examine how the range and response dimension interact to affect the performance of the process flexible structure. We provide analytical lower bounds to show that under all scenarios on response flexibility, moderate form of range flexibility (via chaining structure) still manages to accrue non-negligible benefits vis-a-vis the fully flexible structure (the bound is 29.29% when demand is normally distributed).

We show further that given limited resources, upgrading system response dimension outperforms upgrading system range dimension in most cases. This confirms what most managers believe in intuitively. We observe also that improving system response can provide even more benefits when coupled with initiatives to reduce demand variability. This is in direct contrast with range flexibility, which is more valuable when the system has higher variability.

1. Introduction

Trends in consumer markets have shown a shift towards more customized products and faster upgrades in technology. This has resulted in more product lines, shorter product life cycles, and higher demand variability. Facing increased demand uncertainty as well as heightened market competition, businesses can no longer rely on capacity, pricing, quality, or timeliness alone as competitive strategies. In particular, firms are turning to process flexibility to improve their ability to match supply with uncertain demand. This can be observed in manufacturing industries such as the automotive industry (Wall, 2003; Van Biesebroeck, 2007), the textile/apparel industry (DesMarteau, 1999) and the semiconductor/electronics industry (McCutcheon, 2004), as well as service industries such as call centers (Wallace and Whitt, 2005).

Over the past three decades, the academic literature on flexibility has grown extensively. Not surprisingly, most of the early works were reviews and taxonomies for manufacturing flexibility (Mandelbaum, 1978; Buzacott, 1982; Browne et al., 1984; Kusiak, 1986; Gupta and Goyal, 1989; Sethi and Sethi, 1990; Parker and Wirth, 1999; Beach et al., 2000). During this period, efforts were focused on understanding the very nature of flexibility and on developing measures and evaluation criteria for flexible manufacturing systems (FMS). Because flexibility is a broad and abstract concept, Browne et al.'s (1984) taxonomy breaks flexibility down into eight categories. The list was subsequently expanded by Sethi and Sethi (1990) into eleven classes, summarized in Table 1.

Various measures have been developed for the different types of flexibility. Gupta and Goyal (1989) present a classification of flexibility measures into six types of approaches: economic consequence based approaches, performance criteria approaches, multi-dimensional approaches, petri-net approaches, information theoretic approaches, and decision theoretic approaches. For example, financial losses due to failure to cope with demand fluctuations or machine breakdowns is a measure based on economic consequences, while the ratio of the number of capabilities in a particular FMS to the same number for an ideal FMS with the same number of facilities (Primrose and Leonard, 1984) is a performance-based measure.

* Corresponding author. Tel.: +65 6516 5223.
E-mail addresses: bizchoum@nus.edu.sg (M.C. Chou), gbachua@ntu.edu.sg (G.A. Chua), bizteocp@nus.edu.sg (C.-P. Teo).

0377-2217/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2010.05.038

Table 1
Definition of flexibility types.

Flexibility type	Definition
Machine	Ability of a machine to perform various types of operations
Material handling	Ability to move different parts efficiently through a manufacturing facility
Operation	Ability to produce a part in different ways
Process	Ability to make different parts without a major setup
Product	Ease of adding or substituting new products in a manufacturing facility
Routing	Ability to produce a part by alternate routes through a system
Volume	Ability to operate profitably at different overall output levels
Expansion	Ease by which a manufacturing system can increase capacity and capability
Program	Ability of a system to run unattended for a period of time
Production	Ability to produce different parts without adding major capital equipment
Market	Ease with which a manufacturing system can adapt to a changing market

Most works in the FMS literature focus on machine flexibility and routing flexibility (Chandra and Tombak, 1992; Wahab et al., 2008). Such emphasis on operational details is understandable because managers in the past faced the urgency of effective implementation of the FMS already existing in their companies. Equally if not more important, though, is the strategic issue of design (i.e. what kind of flexibility to employ and how much). However, when it comes to investing in flexibility, managers have already formed a negative impression because of the enormous cost it entails and the little success it has achieved in the past (Jaikumar, 1986). To address this issue, researchers in the management science community (Fine and Freund, 1990; Jordan and Graves, 1995; Van Mieghem, 1998; Bish and Wang, 2004; Akşin and Karaesmen, 2007; Chou et al., 2008, 2009, 2010; Bassamboo et al., 2009) began to examine process flexibility, which Sethi and Sethi (1990) define as "the ability to make different parts without a major setup".

The theoretical justification for the effectiveness of process flexibility can be traced back to the early work of Eppen (1979). For a multi-location newsvendor problem, he showed that the mismatch cost for a decentralized system exceed those in a centralized system, and that the gap between these two systems depends on the demand correlation. Indeed, a decentralized system is analogous to a dedicated production system, while the centralized system corresponds to flexible production. Likewise, it makes sense that process flexibility is most effective when product demands are negatively correlated and least effective when demand correlation is positive.

It should be noted, however, that Eppen's (1979) result on the benefits of consolidation or risk pooling is predicated on the assumption of full consolidation or complete pooling. In the context of process flexibility, we must have a fully flexible production system where all facilities can produce all products for the said theory to hold. In addition, most of the early works on process flexibility examine the appropriate mix of dedicated versus flexible resources, thus focusing only on fully flexible resources (Fine and Freund, 1990; Van Mieghem, 1998; Bish and Wang, 2004). Since companies realize that full flexibility typically comes at great expense, they can only make limited use of these theories on full flexibility, hence the need for an extended theory of partial flexibility.

With most facilities capable of producing most products, one may overinvest in process flexibility. On the other hand, when one has too little or no flexibility at all, this may result in a high level of lost sales. This becomes a question of whether one can achieve the benefits of full flexibility at an acceptable cost level. Jordan and Graves (1995) show via simulation studies that this is possible using the concept of a simple "chaining" strategy. Here, a facility capable of producing a small number of products, but with proper choice of the *process structure* (i.e. product-facility linkages), can achieve nearly as much benefit as the full flexibility system. In the language of Gupta and Goyal's (1989) flexibility measurement classification, the chaining structure may score very poorly based on a performance criterion such as inherent level of flexibility (score of only 20% compared to the ideal fully flexible structure). However, according to an economic consequence-based approach (e.g. expected financial benefits), the chaining structure fares just as well as full flexibility.

While the idea of chaining has been extended in various directions (Graves and Tomlin, 2003; Gurumurthi and Benjaafar, 2004; Hopp et al., 2004; Iravani et al., 2005), efforts were also expended to strengthen the analytical aspect of the chaining theory (Akşin and Karaesmen, 2007; Chou et al., 2009, 2010; Bassamboo et al., 2009). We briefly discuss some of the findings in Chou et al.'s (2010) paper. The approach taken is asymptotic analysis, i.e. their objective is twofold. First, they examine the effectiveness of chaining as system size grows very large. While Jordan and Graves (1995) used a 2-chain (i.e. each node has degree 2), when system size is large, say $n = 100$, should a firm use a 20-chain? How would a 2-chain perform? Secondly, Chou et al. (2010) are able to prove their results analytically.

To illustrate asymptotic analysis, we consider the following n-facility, n-product example. Suppose each plant has a capacity of $C_j = 100$ units for each j, and each product consumes one unit of capacity and has an expected demand of $D_i = 100$ units for each i.[1] We assume further that the demand is normally distributed with a standard deviation of 33 units (so that the probability of negative demand is negligible), and let $\mathbf{D}$ be the vector of demand realization. Process flexibility can be represented using a bipartite graph. A set $\mathcal{A}(n)$ of n product nodes lies on one side of the graph while a set $\mathcal{B}(n)$ of n facility nodes lies on the other side. A link connecting product node i with facility node j means that facility j has the capability to produce product i. Let $\mathcal{G}(n) \subseteq \mathcal{A}(n) \times \mathcal{B}(n)$ denote the set of all such links, i.e. the edge set of the bipartite graph. Hence, each process flexible structure can be uniquely represented by the edge set $\mathcal{G}(n)$. Below are the three most common structures. See Fig. 1(a)–(c) for an illustration of these structures.

1. The dedicated structure: $\mathcal{D}(n) = \{(i,i) | i \in \{1,2,\ldots,n\}\}$.
2. The chaining structure: $\mathcal{C}(n) = \mathcal{D}(n) \cup \{(1,2),(2,3),\ldots,(n-1,n),(n,1)\}$.
3. The fully flexible structure: $\mathcal{F}(n) = \mathcal{A}(n) \times \mathcal{B}(n)$.

[1] Note that the (mean) demand and supply are balanced and identical in this case.

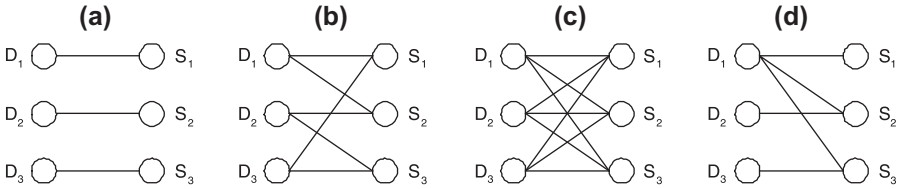

Fig. 1. Bipartite graph representation of 3×3 process flexible structures.

Table 2
Expected sales ratio and expected benefits ratio as system size increases.

System size n	Expected sales			Ratios	
	Dedicated	Chaining	Fully flexible	Expected sales (%)	Expected improvement (%)
10	864.47	949.36	955.14	99.39	93.62
15	1297.51	1434.44	1447.00	99.13	91.59
20	1728.52	1915.78	1938.93	98.81	89.00
25	2179.81	2401.94	2441.73	98.37	84.81
30	2601.84	2871.06	2929.84	97.99	82.08
35	3044.48	3352.66	3430.70	97.73	79.79
40	3469.06	3807.16	3905.48	97.48	77.47

To obtain the *expected system sales* of structure $\mathcal{G}(n)$, we simulate several scenarios of the demand vector $\mathbf{D}$ and take the expected value of the optimal sales over all demand scenarios. In each scenario, we solve the following maximum flow problem, where x_{ij} is the amount of product i produced by facility j and $Z^*(\mathcal{G}(n), \mathbf{D})$ denotes the optimal sales.

$$
\begin{aligned}
Z^*(\mathcal{G}(n), \mathbf{D}) = \quad & \max \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} \\
\text{s.t.} \quad & \sum_{j=1}^{n} x_{ij} \leqslant D_i, \quad \sum_{i=1}^{n} x_{ij} \leqslant C_j \\
& x_{ij} \geqslant 0 \quad \forall\, i,j = 1, \ldots, n, \\
& x_{ij} = 0 \quad \forall (i,j) \notin \mathcal{G}(n).
\end{aligned}
$$

Table 2 shows the expected sales $\mathrm{E}[Z^*(\mathcal{G}(n), \mathbf{D})]$ of the different structures as n varies. The expected sales ratio is the ratio of expected sales in a chaining structure vis-a-viz the fully flexible structure. The expected improvement ratio measures the benefits of a structure (whether chaining or full flexibility) in terms of improvement over the dedicated structure, i.e. expected sales of the given structure minus expected sales of the dedicated structure, vis-a-viz the improvement obtained by the fully flexible structure.

For small n (say $n = 10$), our simulation shows that the expected sales in the three structures are 864.47, 949.36, and 955.14, respectively. This demonstrates that chaining already achieves most (93.62%) of the benefits of full flexibility. However, Table 2 also shows that as the system expands, chaining efficiency[2] deteriorates to as low as 77.47% for $n = 40$. Chou et al. (2010) prove that chaining efficiency for the above example converges to approximately 70% as n approaches infinity. Suitable for any general demand distribution, the approach they developed provides an exact method to capture the asymptotic performance of the chaining structure. Because chaining efficiency decreases in system size, the asymptotic chaining efficiency obtained using this method serves as an analytical lower bound on the performance of any finite chaining system.

Another classification of flexibility is by Slack (1987), who suggested that flexibility has two dimensions: range and response. Range is the extent to which a system can adapt, whereas response is the rate at which the system can adapt. Although both dimensions are important, most papers in the process flexibility literature only consider range flexibility (e.g. partial flexibility versus full flexibility). To the best of our knowledge, there has been no paper in the literature that analytically examines both range flexibility and response flexibility. This paper is an attempt to bridge this gap by extending the theory of partial (range) flexibility (Jordan and Graves, 1995; Chou et al., 2010) to include the response dimension.

To further understand the response dimension, we observe that Slack's (1987) work seems to have influenced Upton (1994) who defined flexibility as "the ability to change or react with little penalty in time, effort, cost or performance". Like Slack (1987), Upton (1994) refers to the extent to which the system can change or react as range. Unlike Slack (1987), he further breaks down the response dimension into mobility and uniformity. Mobility is measured by the penalty incurred as the system switches from state to state. On the other hand, uniformity is measured by how the system can maintain the same performance level as it changes its state.

To model mobility, we can consider the setup time or the setup cost incurred when switching from producing product A to product B. Both are undesirable as setup time effectively reduces capacity whereas setup cost reduces total profits. Moreover, a fully flexible system can be expected to exhibit more production switching than a less flexible system like chaining. Modeling response this way, chaining

[2] We also refer to the expected improvement ratio of chaining to full flexibility as "chaining efficiency".

efficiency or the performance of sparse structures can only improve as the response level deteriorates. This implies that the core model in Chou et al. (2010) as well as their results are already robust against such setup effects. Hence, modeling response as mobility may not be a productive endeavor.

In this paper, we model response as uniformity of production cost. To this end, we distinguish between primary and secondary production. Suppose that the facilities are primarily designed to produce certain products and can only serve as less efficient (or more costly) secondary (or back-up) production options for other products. We then model the response dimension in terms of production cost, whereby secondary production is at least as costly as primary production.[3] If secondary production cost is high, we say that *the response is low*. If secondary production cost is low (comparable to primary production), we say that *the response is high*. In the special case when secondary production cost equals primary production cost, we say that *the response is perfect*. To incorporate these production costs, we use an expected profit criterion for evaluating process flexibility. This criterion generalizes the expected sales criterion used in Jordan and Graves (1995) and Chou et al. (2010).

We seek to address the following research questions. (1) What happens to the effectiveness of chaining or sparse structure when response is not perfect? (2) Can we still analytically capture the value of chaining efficiency as in Chou et al. (2010)? (3) Given limited capital, how do we choose our investment between range flexibility and response flexibility?

The rest of the paper is organized as follows. In Section 2, we study the general effect of the response dimension on process flexibility. Section 3 presents our methods that analytically capture asymptotic chaining efficiency for various response levels. We then examine in Section 4 the investment trade-off between range and response. Finally, Section 5 concludes the paper.

2. Effect of response dimension

In this section, we examine the effect of the response dimension on general process flexible structures. To this end, we consider a firm with m products and n plants, such that $m \geqslant n$ which is typically the case. Product i has random demand D_i whereas plant j has fixed capacity S_j.[4] For this general setting, the process flexible structure is represented by $\mathcal{G}(m,n)$. To model production efficiency, we assume that each product has exactly one primary facility while each facility serves as primary facility for at least one product. We further denote by $\phi(i)$ the index of the primary facility designated to produce product i. Any link $(i,j) \in \mathcal{G}(m,n)$ such that $j \neq \phi(i)$ is considered secondary production. Each unit of product i sold earns the firm r dollars. Without loss of generality, we ignore the goodwill cost associated with unsatisfied demand.[5] If this unit of demand is produced by a primary plant $\phi(i)$, the production cost is c_p. On the other hand, this same product produced by a secondary plant $j \neq \phi(i)$ costs the firm at least as much at $c_s \geqslant c_p$. We call c_p and c_s the costs of primary and secondary production, respectively. To avoid triviality, we assume $c_s < r$. We can then use the cost parameter c_s to capture the system response level as summarized in the table below.

As in Chou et al. (2010), our goal is to measure the expected financial benefits of process flexible structures. To incorporate the response dimension, the previous expected sales criterion must be extended to the expected profit criterion. In lieu of the Maximum Flow problem, we solve the following profit maximization problem, where Π denotes the optimal profit.

Secondary cost	Range of c_s	System response
High	$c_s \geqslant \frac{1}{2}(r + c_p)$	Low
Low	$c_p < c_s < \frac{1}{2}(r + c_p)$	High
Same as primary	$c_s = c_p$	Perfect

$$\Pi(\mathcal{G}(m,n), \mathbf{D}, c_s) = \max (r - c_p) \sum_{i=1}^{m} x_{i,\phi(i)} + (r - c_s) \sum_{i=1}^{m} \sum_{j \neq \phi(i)} x_{ij} \tag{1}$$

$$\text{s.t.} \quad \sum_{j=1}^{n} x_{ij} \leqslant D_i \quad \sum_{i=1}^{m} x_{ij} \leqslant S_j$$

$$x_{ij} \geqslant 0 \quad \forall i = 1, \ldots, m, \forall \quad j = 1, \ldots, n$$

$$x_{ij} = 0 \quad \forall \ (i,j) \notin \mathcal{G}(m,n)$$

We define a measure called **F**lexibility **E**fficiency, which generalizes the expected benefits ratio in Table 2 from expected sales to expected profits. The measure is computed as follows.

$$\text{FE}(\mathcal{G}(m,n), c_s) = \frac{\text{E}[\Pi(\mathcal{G}(m,n), \mathbf{D}, c_s)] - \text{E}[\Pi(\mathcal{D}(m,n), \mathbf{D}, c_s)]}{\text{E}[\Pi(\mathcal{F}(m,n), \mathbf{D}, c_s)] - \text{E}[\Pi(\mathcal{D}(m,n), \mathbf{D}, c_s)]}$$

where $\mathcal{D}(m,n) = \{(i, \phi(i)) | i = 1, \ldots, m\}$ is the dedicated structure with only primary links, and $\mathcal{F}(m,n) = \mathcal{A}(m) \times \mathcal{B}(n)$ is the fully flexible structure.

Observe that the arguments of $\text{FE}(\mathcal{G}(m,n), c_s)$ adequately captures the dimensions of process flexibility, as $|\mathcal{G}(m,n)|$ and c_s, respectively represent the range and response levels. For $\mathcal{G}_1(m,n) \subset \mathcal{G}_2(m,n)$, it is easy to see that $\text{FE}(\mathcal{G}_1(m,n), c_s) \leqslant \text{FE}(\mathcal{G}_2(m,n), c_s)$ since $\mathcal{G}_2(m,n)$ has a larger feasible region. This means that upgrading system range improves system performance. The same can also be said about upgrading system response as shown in the following result.

[3] Production efficiency can also be modeled in terms of production time. In that case, we can approximate increased production time by increased production cost in the sense that in order to retain the original production speed, one has to spend more on other resources like labor.

[4] At this stage, we do not make any assumptions on the demand distribution nor on system symmetry.

[5] In the case where the firm incurs a goodwill cost of g for every unit of unsatisfied demand, we add the goodwill cost to the unit revenue and get the imputed revenue $\bar{r} = r + g$. Replacing r with $\bar{r}$, our analysis carries over.

Theorem 1. *For a fixed flexible structure $\mathcal{G}(m,n)$, such that $\mathcal{D}(m,n) \subseteq \mathcal{G}(m,n) \subseteq \mathcal{F}(m,n)$, its flexibility efficiency is non-increasing in c_s over the interval $[c_p,r)$.*

Proof. Please refer to Appendix A. □

It follows that for any flexible structure, its flexibility efficiency can only worsen or stay the same as system response worsens. In other words, when we take into account the possibility of low response flexibility, the value of any flexible structure may lessen.[6] This serves as a precaution not to oversell the benefits of any structure based on its range flexibility alone, and as a call to examine the response dimension when investing in flexibility.

Although the previous result may not have come unexpected, a more surprising result is that when system response deteriorates to a certain level (i.e. it enters the low response region), further deterioration will cause no more harm to the system than it does to the full flexibility system. The following theorem captures this insight.

Theorem 2. *For a fixed flexible structure $\mathcal{G}(m,n)$, such that $\mathcal{D}(m,n) \subseteq \mathcal{G}(m,n) \subseteq \mathcal{F}(m,n)$, $\mathrm{FE}(\mathcal{G}(m,n),c_s)$ is constant over the interval $[\frac{1}{2}(r+c_p),r)$.*

Proof. Note that the profit margin derived from secondary production is lower than or equal to that derived from primary production, i.e. $r-c_s \leqslant r-c_p$. This implies that one must perform secondary production only when either of two conditions holds. The first is when a particular facility has excess capacity after exhausting all possible primary production. When it has no excess capacity, secondary production can still occur if it can transfer some primary production to another facility (albeit secondary for that product) in order to free up capacity for secondary production. In other words, one trades a unit of primary production for two units of secondary production.

Since $c_s \in [\frac{1}{2}(r+c_p),r)$, it follows that $r-c_p \geqslant 2(r-c_s)$. This means that trading one primary unit for two secondary units leaves the system worse off. Hence, the optimal allocation can be obtained greedily by letting each facility produce as many units of its primary products as possible, and *only* thereafter use its extra capacity, if any, to produce the extra demand, if any, of secondary product. If X_p and X_s are the optimal primary and secondary production, then we have

$$X_p = \sum_{j=1}^{n} \min\left(\sum_{i:\phi(i)=j} D_i, S_j\right) \quad \text{and} \quad X_s = \min\left(\sum_{i=1}^{m} D_i, \sum_{j=1}^{n} S_j\right) - \sum_{j=1}^{n} \min\left(\sum_{i:\phi(i)=j} D_i, S_j\right)$$

Hence,

$$\begin{aligned}
\mathrm{FE}(\mathcal{G}(m,n),c_s) &= \frac{\mathrm{E}\left[(r-c_s)X_s - (r-c_p)\left(\sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right) - X_p\right)\right]}{\mathrm{E}\left[(r-c_s)\left(\min\left(\sum_{i=1}^{m} D_i, \sum_{j=1}^{n} S_j\right) - \sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right)\right)\right]} \\
&= \frac{\mathrm{E}[(r-c_s)X_s]}{\mathrm{E}\left[(r-c_s)\left(\min\left(\sum_{i=1}^{m} D_i, \sum_{j=1}^{n} S_j\right) - \sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right)\right)\right]} \\
&= \frac{\mathrm{E}[X_s]}{\mathrm{E}\left[\left(\min\left(\sum_{i=1}^{m} D_i, \sum_{j=1}^{n} S_j\right) - \sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right)\right)\right]}
\end{aligned}$$ □

Theorem 2 shows that once the response level hits the halfway mark between perfect response and worst-case response, the flexibility efficiency of the structure starts to plateau at a constant level. This is because at that point, any additional deterioration in response can cause only as much harm to the structure as it does to the fully flexible structure. We present two examples to illustrate this phenomenon. The first is a symmetric system with uncertain product demands and fixed facility capacities. The second is a real-life case study on a bread delivery system in Singapore. Unlike the first example, the "Food From the Heart" delivery system is asymmetric with uncertain bread supplies but fixed demands.

Example 1. Chaining Structure for a 3×3 system with uniform demand

Suppose all product demands are i.i.d. and uniformly distributed in $[0,2\mu]$. Let *CE* denote the flexibility efficiency of the chaining structure, i.e. $CE(3,c_s) = \mathrm{FE}(\mathcal{C}(3),c_s)$. It is not difficult, though it is tedious, to evaluate the $CE(3,c_s)$ in closed form for this special case.

$$CE(3,c_s) = \begin{cases} 1 - \frac{3}{11}\frac{c_s-c_p}{r-c_s} & \text{if} \quad c_p \leqslant c_s < \frac{1}{2}(r+c_p) \\ \frac{8}{11} & \text{if} \quad \frac{1}{2}(r+c_p) \leqslant c_s < r \end{cases}$$

Without loss of generality, we let $r = 1$ and $c_p = 0$, and we plot as follows.

As in Theorems 1 and 2, Fig. 2 shows that chaining efficiency deteriorates as system response worsens. However, the amount of deterioration does not decrease below the 72.7% mark. This means that upon entering the low response region, any further deterioration in response level will have no more effect on the chaining structure than it would have on full flexibility. Since we expect the 72.7% lower bound to decrease further as system size increases, we examine in Section 3 the asymptotic limit of this lower bound as n approaches infinity.

Example 2. Food from the Heart, Singapore

Food from the Heart (FFTH) is a charity organization based in Singapore, where the focus is on ensuring that end-of-day bread donations from bakeries go into the right hands, those of the people in the homes supported by the organization (Chou et al., 2009). The logistics involved is simple. Each bakery is served by one volunteer each night to bring the donated bread to his or her designated home. An FFTH

[6] Note that the result does not require any assumption about the demand distribution, facility capacity, or system symmetry.

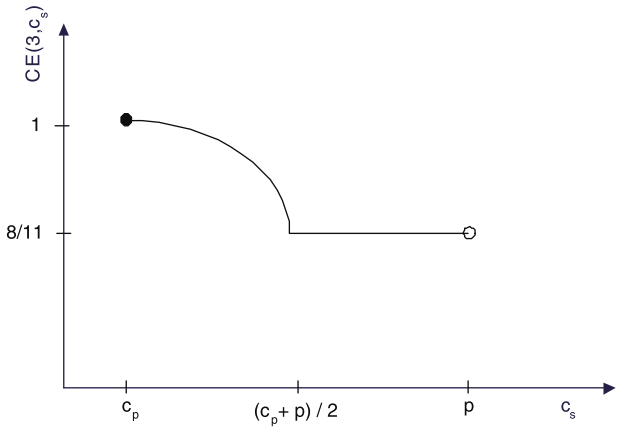

Fig. 2. Chaining efficiency as a function of response (Uniform 3×3).

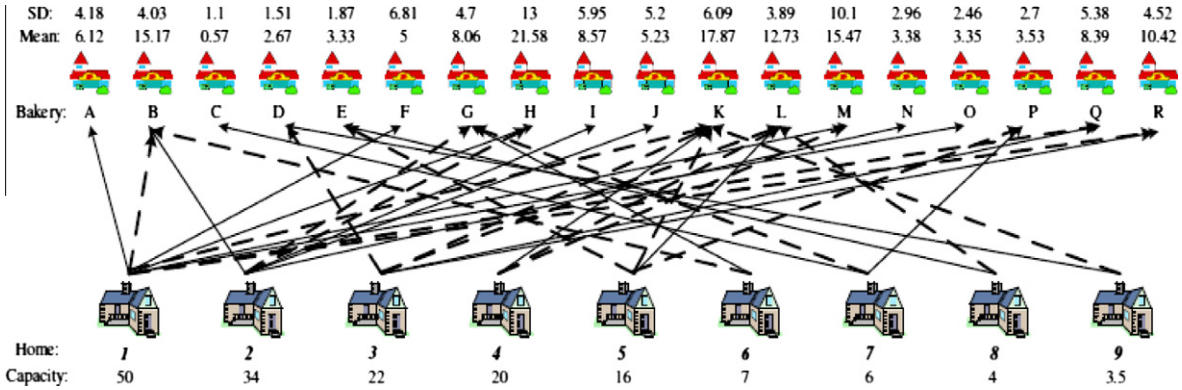

Fig. 3. Food from the heart – flexible routing system.

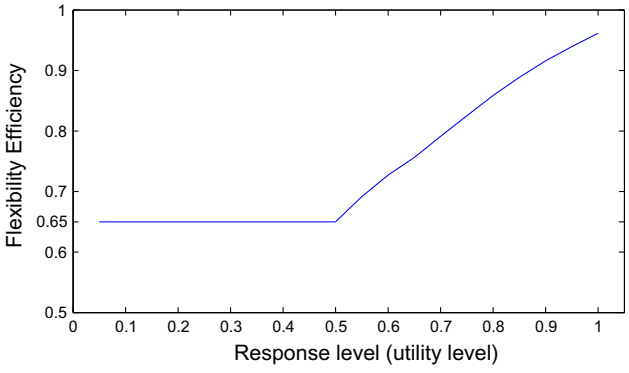

Fig. 4. Food from the heart – flexibility efficiency versus response level.

administrator selects routes (bakery-home assignments) and assigns a volunteer to each route. To reduce the burden on the volunteers, the administrator usually assigns a fixed route to each volunteer. Because the supply of bread from each bakery on each night is random, the rigidity of having only fixed primary routes inevitably leads to supply and demand mismatch. For an 18-bakery 9-home study of the FFTH system, Chou et al. (2009) examine how to optimally add a set of 18 secondary routes to the existing system of 18 primary routes.

Fig. 3 shows the flexible routing system generated by their algorithm. The solid lines pertain to the original primary routes, while the broken lines are the secondary routes. While the homes have fixed demands for bread, the bakeries' daily supplies are recorded for 66 days. The means and standard deviations, as well as the fixed demands are shown in Fig. 3. All units are in kilograms.

In this example, we investigate how the FFTH flexible routing system performs as system response deteriorates. While Chou et al. (2009) show that a system with uncertain supply and fixed demand is equivalent to one with uncertain demand and fixed supply, we also observe that the FFTH system has more uncertain nodes (18) than fixed nodes (9). Hence, the FFTH system fits into the general model in this section. Moreover, we let $r = 1$ and $c_p = 0$, and define $u = 1 - c_s$ as the utility from one kilogram of bread. Because $c_s \in [0,1)$, if follows that $u \in (0,1]$ can be a surrogate for the response level. We simulate the supplies at the bakeries over the 66-day empirical distribution. We generate

1000 demand scenarios and take the average maximum utility over all scenarios. Fig. 4 plots the flexibility efficiency of the FFTH flexible routing system as response level changes. As response deteriorates, so does flexibility efficiency. However, upon entering the low response region, the system will no longer achieve benefits any lower than 65% of the fully flexible system. This example illustrates the applicability of our results to asymmetric systems and uncertain supplies.

To sum up, our discussion yields the following fundamental insight on the performance of flexible system:

While a little range flexibility allows the system to reap most of the benefits that can be accrued from a fully flexible system in the perfect response scenario, there is a threshold below which any further loss in response flexibility has absolutely no impact on the performance of the flexible system.

3. Asymptotic analysis for range and response

In this section, we examine how the relationship between flexibility efficiency and response level changes as system size grows to infinity. Particularly, we extend Chou et al.'s (2010) results on asymptotic chaining efficiency to factor in less than perfect response. For ease of exposition, we consider a stylized model where all products have independent, identically distributed, and symmetric demands, with values in the range $[0, 2\mathrm{E}(D_i)]$. Examples of such demand distributions are uniform and (truncated) normal distributions. Note that our analysis can be extended easily to more general and asymmetric demand distribution. On the supply side, all facilities have capacities $S_i = \mathrm{E}(D_i) = \mu, \forall\ i$. This is the symmetric system, where demand and supply are identical and balanced.

Because of the assumption of symmetry, we can model demand in the following general form. Let $D_i = \mu + a_i Y_i$, where $0 \leqslant Y_i \leqslant \mu$ and

$$a_i = \begin{cases} 1, & \text{with probability } \frac{1}{2} \\ -1, & \text{with probability } \frac{1}{2} \end{cases}$$

Note that Y_i follows some distribution with support $[0, \mu]$ and represents the absolute deviation of demand D_i from the mean μ. Our interest is in characterizing the chaining efficiency defined as follows

$$CE(n, c_s) = \mathrm{FE}(\mathcal{C}(n), c_s) = \frac{\mathrm{E}[\Pi(\mathcal{C}(n), \mathbf{D}, c_s)] - \mathrm{E}[\Pi(\mathcal{D}(n), \mathbf{D}, c_s)]}{\mathrm{E}[\Pi(\mathcal{F}(n), \mathbf{D}, c_s)] - \mathrm{E}[\Pi(\mathcal{D}(n), \mathbf{D}, c_s)]}.$$

From (10)–(12), we can express the denominator of $CE(n,c_s)$ as follows:

$$\mathrm{E}[\Pi(\mathcal{F}(n), \mathbf{D}, c_s)] - \mathrm{E}[\Pi(\mathcal{D}(n), \mathbf{D}, c_s)] = (r - c_s)\mathrm{E}\left[\min\left(\sum_{i=1}^{n} D_i, n\mu\right) - \sum_{i=1}^{n} \min(D_i, \mu)\right] \tag{2}$$

The challenge in our analysis is to evaluate the term $\mathrm{E}[\Pi(\mathcal{C}(n), \mathbf{D}, c_s)]$. As shown in Theorem 2, the optimal production allocation varies dramatically depending on whether c_s falls in the low or high response region.

3.1. Low response region

In this region, $c_s \geqslant \frac{1}{2}(r + c_p)$, which implies $r - c_p \geqslant 2(r - c_s)$. That is, one unit of primary production is at least as profitable as two units of secondary production. Hence, optimal allocation must be greedy as follows. Let each facility produce as many units as possible of its primary product, and *only* thereafter use its extra capacity, if any, to produce the extra demand, if any, of its secondary product.

$$\begin{aligned} &x^*_{ii} = \min(D_i, \mu) \quad \forall\ i = 1, \ldots, n \\ &x^*_{i,i+1} = \min[(D_i - \mu)^+, (\mu - D_{i+1})^+] \quad \forall\ i = 1, \ldots, n-1, \\ &x^*_{n1} = \min[(D_n - \mu)^+, (\mu - D_1)^+] \end{aligned} \tag{3}$$

The following well-known facts on the normal distribution will be useful for our next result.

Lemma 1. *If X, X_1, $X_2 \sim N(0,\sigma)$, and X_1, X_2 are independent, then* $\mathrm{E}[X^+] = \frac{\sigma}{\sqrt{2\pi}}$ *and* $\mathrm{E}\left[\min\left(X_1^+, X_2^+\right)\right] = \frac{\sigma}{\sqrt{2\pi}}\left(1 - \frac{1}{\sqrt{2}}\right)$.

Theorem 3. *When $\frac{1}{2}(r + c_p) \leqslant c_s < r$ and for sufficiently large n, the chaining efficiency is decreasing in n and bounded below by the asymptotic chaining efficiency*

$$ACE(c_s) = \lim_{n\to\infty} CE(n, c_s) = \frac{1}{2}\frac{\mathrm{E}[\min(Y_1, Y_2)]}{\mathrm{E}[Y_1]} \leqslant 0.5.$$

Proof. We first write the expected optimal profit for the chaining structure.

$$\begin{aligned} \mathrm{E}[\Pi(\mathcal{C}(n), \mathbf{D})] &= (r - c_p)\mathrm{E}\left[\sum_{i=1}^{n} \min(D_i, \mu)\right] + (r - c_s)\mathrm{E}\left[\sum_{i=1}^{n} \min[(D_i - \mu)^+, (\mu - D_{i+1})^+]\right] \\ &= n(r - c_p)\mathrm{E}[\min(D_i, \mu)] + n(r - c_s)\mathrm{E}[\min[(D_1 - \mu)^+, (\mu - D_2)^+]] \\ &= n(r - c_p)\mathrm{E}[\min(D_i, \mu)] + \frac{1}{4}n(r - c_s)\mathrm{E}[\min(Y_1, Y_2)] \end{aligned} \tag{4}$$

The first equation is from (3), while the second equation comes from the identical distribution of the demands. The last equation is due to the definition of the absolute demand deviation Y_i. Since the first term in (4) is also the expected optimal profit for the dedicated structure, the numerator of $CE(n,c_s)$ becomes

$$\frac{1}{4}n(r-c_s)\mathrm{E}[\min(Y_1,Y_2)] \tag{5}$$

For the denominator, we let $T=\sum_i D_i$. Since n is sufficiently large, we invoke the Central Limit Theorem to get $T\sim N(n\mu,\sqrt{n}\sigma)$ and $X=T-n\mu\sim N(0,\sqrt{n}\sigma)$, where σ is the standard deviation of demand D_i. Then, we use (2), Lemma 1(a), and the definition of Y_i to obtain

$$\begin{aligned}(r-c_s)\mathrm{E}\left[\min(T,n\mu)-\sum_{i=1}^{n}\min(D_i,\mu)\right] &= (r-c_s)\mathrm{E}\left[\sum_i D_i - X^+ - \sum_i D_i + \sum_i (D_i-\mu)^+\right] = (r-c_s)\left[n\mathrm{E}(D_1-\mu)^+ - \frac{\sqrt{n}\sigma}{\sqrt{2\pi}}\right]\\ &= n(r-c_s)\left[\frac{1}{2}\mathrm{E}[Y_1]-\frac{1}{\sqrt{n}}\frac{\sigma}{\sqrt{2\pi}}\right]\end{aligned} \tag{6}$$

Combining (5) and (6), we obtain

$$CE(n,c_s)=\frac{\frac{1}{4}\mathrm{E}[\min(Y_1,Y_2)]}{\frac{1}{2}\mathrm{E}[Y_1]-\frac{1}{\sqrt{n}}\frac{\sigma}{\sqrt{2\pi}}}$$

which is decreasing in n. By taking limit, we arrive at the desired result. □

Our analysis above does not consider that the chaining structure usually exploits long chains – chains that link up as many supply and demand nodes as possible. In particular, we only consider that the secondary links in the chaining structure form a perfect matching. Such a matching can also be formed by a collection of short chains. As a result, a long chain performs identically to short chains under this scenario. This observation contrasts with the fundamental insight in the literature where system response is assumed perfect.

3.2. Perfect response: the expected sales criterion

When $c_s = c_p$, the expected profit criterion reduces to the expected sales criterion. While Chou et al. (2010) have developed a method that analytically obtains the asymptotic chaining efficiency (ACE), their method does not extend to the general high response region. In this section, we describe another method which can provide a lower bound for ACE in the perfect response case and can extend to other high response scenarios.

Since expected optimal profit becomes expected maximum flow, we revert to the maximum flow notation $Z^*(\mathcal{G}(n),\mathbf{D})$ for structure $\mathcal{G}(n)$ used in Section 1. If follows that

$$CE(n,c_p)=\frac{\text{Expected Chaining Gain}}{\text{Expected Full Flexibility Gain}}=\frac{\mathrm{E}[Z^*(\mathcal{C}(n),\mathbf{D})]-\mathrm{E}[Z^*(\mathcal{D}(n),\mathbf{D})]}{\mathrm{E}[Z^*(\mathcal{F}(n),\mathbf{D})]-\mathrm{E}[Z^*(\mathcal{D}(n),\mathbf{D})]}$$

Similar to (6), we can express the denominator as

$$\text{Expected Full Flexibility Gain}=n\left(\frac{1}{2}\mathrm{E}[Y_1]-\frac{1}{\sqrt{n}}\frac{\sigma}{\sqrt{2\pi}}\right) \tag{7}$$

For the numerator, consider any demand realization $\mathbf{D}$. Observe that each demand node i has either $D_i > \mu$ (positive node) or $D_i < \mu$ (negative node) with equal likelihood.[7] We define a *cluster* to be a run of consecutive positive nodes followed by a run of consecutive negative nodes. For example, suppose $n = 10$ and the demand outcome is {*N*,*P*,*P*,*P*,*N*,*N*,*N*,*P*,*N*,*N*} where *P* denotes a positive node while *N* denotes a negative node. The 2nd to 7th nodes form the first cluster {*P*,*P*,*P*,*N*,*N*,*N*} while the last 3 and the 1st form the next cluster {*P*,*N*,*N*,*N*}. This allows us to break the whole system into smaller pieces (clusters), and we can easily optimize the flow for each cluster. The aggregate solution from all clusters remains feasible for the max-flow problem of the whole system, and thus provides a lower bound for $Z^*(\mathcal{C}(n),\mathbf{D})$, that is,

$$\text{Expected Chaining Gain} \geqslant \mathrm{E}[\text{Sum of Cluster Chaining Gains}] = \mathrm{E}[\text{Number of Clusters}]\cdot\mathrm{E}[\text{Cluster Chaining Gain}] \tag{8}$$

The last equation holds for large n, and is the result of Wald's equation (Ross, 2003, p. 462) and the fact that all clusters are probabilistically identical and independent.

To obtain expected cluster chaining gain, we consider just one cluster. For this cluster, the lengths of the positive and negative runs as well as the deviations of realized demands from the mean are all random variables. Let M and N be the lengths of the positive and negative runs, respectively. Both M and N follow a geometric distribution with $p = 0.5$. From our earlier definition, we have Y_i for the demand deviations ($Y_i = D_i - \mu$ for positive nodes while $Y_i = \mu - D_i$ for negative nodes). To apply (8), we derive the following lemmas.

Lemma 2. *For any cluster with M positive nodes followed by N negative nodes, the maximum chaining flow is*

$$\text{Max Flow}=\min\left(\sum_{i=1}^{M+N}D_i,(M+1)\mu+\sum_{i=M+1}^{M+N}D_i,(M+N)\mu\right)$$

and the cluster chaining gain is

$$\text{Cluster Chaining Gain}=\min\left(\sum_{i=1}^{M}Y_i,\mu,\sum_{i=M+1}^{M+N}Y_i\right)$$

[7] We assume demand distribution has continuous support.

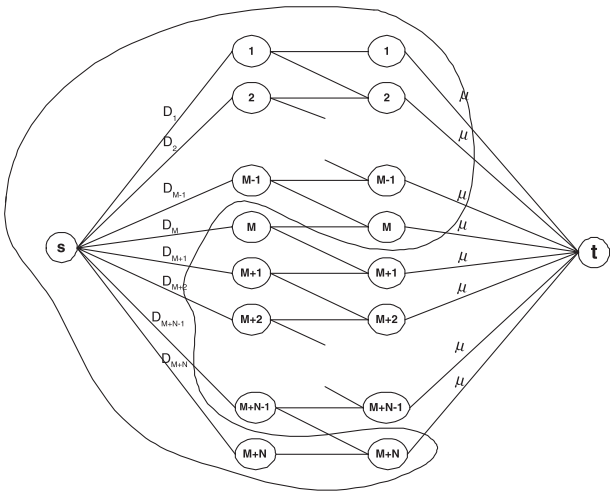

Fig. 5. Maximum flow problem: $\mathbb{C}_1 = \{s, 1, 2, \ldots, M-1, M+N\}$.

Proof. We use the equivalence between max-flow and min-cut to derive the above result. Consider the max-flow problem on the network shown in Fig. 5, from source s to sink t. All links are directed from left to right with capacities as indicated. Arcs from demand nodes to supply nodes have infinite capacities. Let $\mathbb{C}$ be a cut of the network and $V(\mathbb{C})$ be its cut value. Because of the infinite capacities of demand-to-supply arcs, every cut with finite cut value can be uniquely represented by $\{s\}$ union with a subset of $\mathbb{S} = \{1, 2, \ldots, M+N\}$. For example, $\mathbb{C}_1 = \{s, 1, 2, \ldots, M-1, M+N\}$ represents the cut in Fig. 5, with cut value $V(\mathbb{C}_1) = \sum_{i=M}^{M+N-1} D_i + (M+1)\mu$.

Let $\mathbb{S}^+ = \{1, 2, \ldots, M\}$ and $\mathbb{S}^- = \{M+1, M+2, \ldots, M+N\}$. Then every cut can be written as $\mathbb{C} = \{s\} \cup \mathbb{P} \cup \mathbb{N}$ where $\mathbb{P} \subseteq \mathbb{S}^+, \mathbb{N} \subseteq \mathbb{S}^-$. Recall that $D_i > \mu$ for $i \in \mathbb{S}^+$ and $D_i < \mu$ for $i \in \mathbb{S}^-$. Let $\mathbb{C}^* = \{s\} \cup \mathbb{P}^* \cup \mathbb{N}^*$ be a minimum cut of the above network flow problem. It is easy to see that the search for a minimum cut can be restricted to the three cuts $\{s\}$, $\{s\} \cup \mathbb{S}^+$, and $\{s\} \cup \mathbb{S}$, from which the first result follows.

The second result is an easy consequence of the first, since

$$\text{Cluster Chaining Gain} = \text{Max Chaining Flow} - \text{Max Dedicated Flow}$$
$$= \min\left(\sum_{i=1}^{M+N} D_i, (M+1)\mu + \sum_{i=M+1}^{M+N} D_i, (M+N)\mu\right) - \left(\sum_{i=1}^{M} \mu + \sum_{i=M+1}^{M+N} D_i\right) = \min\left(\sum_{i=1}^{M} Y_i, \mu, \sum_{i=M+1}^{M+N} Y_i\right) \qquad \square$$

Lemma 3. *For an $n \times n$ system,*

$$\frac{E[\text{Number of Clusters}]}{n} \to \frac{1}{4} \text{ as } n \to \infty$$

Proof. Without loss of generality, assume node 1 is negative. Then, the number of clusters from node 1 to node n can be viewed as a counting process, in fact, a renewal process whereby each occurrence of a cluster constitutes a renewal. By the Elementary Renewal Theorem (Ross, 2003, p. 409) and because E[cluster length] = E[$M + N$] = 4, the result follows. $\square$

Combining (7), (8), Lemmas 2 and 3, we obtain the following key result.

Theorem 4. *When system response is perfect ($c_s = c_p$), the asymptotic chaining efficiency is bounded below as follows:*

$$ACE(c_p) = \lim_{n\to\infty} CE(n, c_p) \geqslant \frac{1}{2} \frac{\mathrm{E}\left[\min\left(\sum_{i=1}^{M} Y_i, \sum_{i=1}^{N} \widetilde{Y}_i, \mu\right)\right]}{\mathrm{E}[Y_1]}$$

where M,N are geometric r.v. with p = 0.5, and Y_i, $\widetilde{Y}_i$ are i.i.d. random variables with support [0,μ].

3.3. High response region

In this region, $c_p < c_s < \frac{1}{2}(r + c_p)$, which implies $r - c_p < 2(r - c_s)$. That is, it is profitable to displace one unit of primary production in favor of two units of secondary production. Therefore, the greedy production allocation used in Section 3.1 no longer works. On the other hand, the maximum flow approach in Section 3.2 may include using the (extra) capacity of facility i to meet the extra demand for product $i + j$ for any i and j, through the intermediate facility $i + 1, i + 2, \ldots, i + j - 1$. We call such an allocation a j-order displacement. It is easy to see that a j-order displacement is justified only if j units of secondary production are as profitable as $j - 1$ units of primary production, a requirement not necessarily satisfied by $r - c_p < 2(r - c_s)$. Hence, our analysis requires dividing this high response region into countably infinite subcases, namely,

$$\frac{k}{k-1} \leqslant \frac{r - c_p}{r - c_s} < \frac{k-1}{k-2}, \quad \forall\, k = 3, 4, \ldots \tag{9}$$

For subcase k, a j-order displacement is profitable for $j < k$, but not for $j \geqslant k$. Therefore, if we use the maximum flow approach, we should distinguish flows that result in profitable displacement from flows that do not. In particular, the optimal allocation should not include displacements of order k or higher. Secondly, the flows must be assigned different weights, corresponding to different profit levels, depending on the amount of production displaced. Specifically, for a j-order displacement, the unit profit is $j(r - c_s) - (j-1)(r - c_p)$, provided $j < k$.

We use the same terminology in Section 3.2, and let $M + N$ denote the cluster length and write as CCG for Cluster Chaining Gain in short. Let $g(j)$ be the expected maximum flow net of maximum dedicated flow, for a cluster of length j. That is,

$$g(j) = E[CCG|M + N = j].$$

We obtain $g(j)$ in a manner similar to Section 3.2 and set $g(1) = 0$ for completeness. Let $\Delta g(j) = g(j+1) - g(j)$ represent the incremental gain in a cluster of length $j+1$ over a cluster of length j. The asymptotic chaining efficiency can then be bounded as follows.

Theorem 5. *For $k = 3,4,\ldots$, if $\frac{1}{k}[r + (k-1)c_p] \leqslant c_s < \frac{1}{k-1}[r + (k-2)c_p]$, then*

$$ACE(c_s) = \lim_{n\to\infty} CE(n, c_s) \geqslant \frac{\sum_{i=1}^{k-1}[i(r - c_s) - (i-1)(r - c_p)]\Delta g(i) \cdot \mathrm{P}\{M + N > i\}}{2(r - c_s)\mathrm{E}[Y_1]}$$

where $\Delta g(j) = g(j+1) - g(j)$, $g(j) = \mathrm{E}\left[\min\left(\sum_{i=1}^{M} Y_i, \sum_{i=1}^{N} \widetilde{Y}_i, \mu\right)|M + N = j\right]$, $g(1) = 0$, and M,N are geometric with $p = 0.5$, and Y_i, $\widetilde{Y}_i$ are i.i.d. random variables with support $[0,\mu]$.

We omit the technical details of the proof due to space constraint. Theorems 3–5 complete the characterization of asymptotic chaining efficiency over all relevant response levels. Next, we show some examples.

3.4. Computational examples

3.4.1. Uniform distribution

Suppose that demand $D_i \sim U[0,2\mu], \forall i$. It follows that $Y_i \sim U[0,\mu]$. It can be shown that

$$\mathrm{E}[Y_1] = \frac{1}{2}\mu, \quad \mathrm{E}[\min(Y_1, Y_2)] = \frac{1}{3}\mu, \quad \sigma = \frac{\mu}{\sqrt{3}}$$

Hence, for $\frac{1}{2}(r + c_p) \leqslant c_s < r$,

$$CE(n, c_s) = \frac{1}{3 - \frac{2\sqrt{6}}{\sqrt{n}\sqrt{\pi}}}, \quad ACE(c_s) = \frac{1}{3} \approx 33.33\%.$$

In this case, even with very poor response flexibility and system size n becomes extremely large, the chaining structure with only $2n$ links still manages to accrue 33.33% of the benefits of the fully flexible system with n^2 links. Moreover, if response flexibility improves, this worst-case performance will likewise improve.

3.4.2. Normal distribution

Suppose $D_i \sim N(\mu,\sigma), \forall i$. It follows that $X_i = D_i - \mu \sim N(0,\sigma)$ and $Y_i = |X_i|$. Assume further that $\mu \geqslant 3\sigma$ so that negative demand has negligible probability. It can derived that

$$\mathrm{E}[Y_1] = \frac{2\sigma}{\sqrt{2\pi}}, \quad \mathrm{E}[\min(Y_1, Y_2)] = \frac{4\sigma}{\sqrt{2\pi}}\left(1 - \frac{1}{\sqrt{2}}\right)$$

Hence, for $\frac{1}{2}(r + c_p) \leqslant c_s < r$

$$CE(n, c_s) = \frac{1 - \frac{1}{\sqrt{2}}}{1 - \frac{1}{\sqrt{n}}}, \quad ACE(c_s) = 1 - \frac{1}{\sqrt{2}} \approx 29.29\%$$

Note that this lower bound of 29.29% is not only independent of the actual magnitudes of μ and σ, but also independent of $CV \triangleq \sigma/\mu$. This contrasts with the perfect-response result in Chou et al. (2010), where the ACE for normal distribution varies with the CV. Hence, as long as $\mu \geqslant 3\sigma$ but regardless of the CV, the chaining structure with $2n$ links can still achieve 29.29% of the fully flexible system with n^2 links. As system response improves, this lower bound also improves.

3.4.3. Beta distribution

The methods can be extended to analyze the case when demand is not symmetric or not balance. We illustrate the approach using beta distribution as demand distribution. While the general beta distribution requires four parameters, we can focus (WLOG) on the standard

Table 3
Low response ACE for beta distribution.

Case	α	β	μ	σ	S/μ				
					1.2 (%)	1.1 (%)	1 (%)	0.9 (%)	0.8 (%)
1	0.5	0.5	0.50	0.3536	46.26	40.70	35.81	40.86	45.64
2	1	1	0.50	0.2887	45.98	39.67	33.90	40.98	46.98
3	2	2	0.50	0.2236	49.15	39.07	31.44	41.58	50.52
4	1	3	0.25	0.1936	39.38	34.55	29.88	33.85	38.46
5	4	1	0.80	0.1633	74.41	51.23	28.93	48.06	64.03
6	2	5	0.29	0.1597	45.04	35.44	29.38	34.68	44.21

beta distribution which has only two parameters; namely, α and β. This is because ACE is invariant over the scale of demand relative to supply. We consider 6 different sets of parameters and 5 different ratios of supply to mean demand. The results are summarized in Table 3. Observe that for Case 2, which is the uniform distribution, we reproduce the result in Section 3.4.1 for the balanced scenario (i.e. $S = \mu$).

By varying the values of c_s, our simulation result shows that all systems considered achieve at least about 30% of the benefits of full flexibility even in the worst case. This bound further improves as system response improves or as expected demand deviates further away from total supply. This is already evident in the low response case shown in Table 3. This implies that our earlier results obtained for the balanced case are conservative estimates for the non-balanced case.

4. Trade-offs and complements

4.1. Range versus response

Although we have seen that the chaining structure manages to accrue non-negligible benefits even in the worst case (e.g. 29.29% for normally distributed demands), one can certainly still improve his system performance by either upgrading response flexibility or range flexibility. With limited resources, it is of interest to know which upgrade provides greater improvement: a high response with limited range or a high range with low response. For example, chaining with low secondary cost or full flexibility with high secondary cost?

Let $\mathcal{S}_1(n)$ and $\mathcal{S}_2(n)$ be the high response (chaining) and high range (full flexibility) systems, respectively. i.e. $\mathcal{S}_1(n) = \mathcal{C}(n)$ and $\mathcal{S}_2(n) = \mathcal{F}(n)$. Denote their respective costs of secondary production by c_1 and c_2 such that $c_1 < c_2$. Our goal then is to compare the ratios of each system to the best possible system, which is full flexibility with secondary cost at c_p. That is,

$$\lim_{n\to\infty} \frac{\mathrm{E}[\Pi(\mathcal{S}_1(n), \mathbf{D}, c_1)]}{\mathrm{E}[\Pi(\mathcal{F}(n), \mathbf{D}, c_p)]} \quad \text{versus} \quad \lim_{n\to\infty} \frac{\mathrm{E}[\Pi(\mathcal{S}_2(n), \mathbf{D}, c_2)]}{\mathrm{E}[\Pi(\mathcal{F}(n), \mathbf{D}, c_p)]}$$

For the moment, suppose that $c_2 \geqslant \frac{1}{2}(r + c_p)$ and $c_1 = c_p$. It is easy to see that

$$\lim_{n\to\infty} \frac{\mathrm{E}[\Pi(\mathcal{S}_1(n), \mathbf{D}, c_1)]}{\mathrm{E}[\Pi(\mathcal{F}(n), \mathbf{D}, c_p)]} = \lim_{n\to\infty} \frac{\mathrm{E}[Z^*(\mathcal{C}(n), \mathbf{D})]}{\mathrm{E}[Z^*(\mathcal{F}(n), \mathbf{D})]} = \mathrm{ACE} + (1 - \mathrm{ACE})\left(\lim_{n\to\infty} \frac{\mathrm{E}[Z^*(\mathcal{D}(n), \mathbf{D})]}{\mathrm{E}[Z^*(\mathcal{F}(n), \mathbf{D})]}\right) = \mathrm{ACE} + (1 - \mathrm{ACE})\left(\frac{\mu - E[(\mu - D_i)^+]}{\mu}\right)$$
$$= 1 - (1 - \mathrm{ACE}) \cdot \frac{E[(\mu - D_i)^+]}{\mu}$$

where

$$ACE \triangleq \lim_{n\to\infty} CE(n, c_p) = \lim_{n\to\infty} \left(\frac{\mathrm{E}[Z^*(\mathcal{C}(n), \mathbf{D})] - \mathrm{E}[Z^*(\mathcal{D}(n), \mathbf{D})]}{\mathrm{E}[Z^*(\mathcal{F}(n), \mathbf{D})] - \mathrm{E}[Z^*(\mathcal{D}(n), \mathbf{D})]}\right).$$

Moreover, we can prove a bound on ACE following the generalized random walk approach used in Chou et al. (2010) – where the ACE is related to some performance indicators (stopping time and cycle overshoots) in a related generalized random walk. Due to space limitation, we refer the readers to Chou et al. (2010) for the technical details.

Lemma 4 (Chou et al. (2010)).

$$\mathrm{ACE} = 1 - \frac{E[\psi_0]}{2E[\tau_0]\mathrm{E}[(D_i - \mu)^+]} \geqslant \frac{1}{2}.$$

where ψ_0 and τ_0 are the cycle overshoot and cycle duration in the generalized random walk approach.

Proof. The result follows from Theorem 2 in Chou et al. (2010). The lower bound of 1/2 follows from the symmetry of demand distribution, and

$$\psi_0 \leqslant \sum_{i=1}^{\tau_0} (D_i - \mu)^- + (D_{\tau_0} - \mu)^+ = \sum_{i=1}^{\tau_0 - 1} (D_i - \mu)^- + (D_{\tau_0} - \mu)^+,$$

since $(D_i - \mu)^- = 0$ when $i = \tau_0$. □

We are now ready to present the following result:

Theorem 6. *If demands are i.i.d. and symmetric, then response is at least as good as range, that is,*

$$\lim_{n\to\infty} \frac{\mathrm{E}[\Pi(\mathcal{S}_1(n), \mathbf{D}, c_1)]}{\mathrm{E}[\Pi(\mathcal{F}(n), \mathbf{D}, c_p)]} \geqslant \lim_{n\to\infty} \frac{\mathrm{E}[\Pi(\mathcal{S}_2(n), \mathbf{D}, c_2)]}{\mathrm{E}[\Pi(\mathcal{F}(n), \mathbf{D}, c_p)]}$$

where response is chaining ($\mathcal{S}_1(n) = \mathcal{C}(n)$) with secondary cost at $c_1 = c_p$ while range is full flexibility ($\mathcal{S}_2(n) = \mathcal{F}(n)$) with secondary cost at $c_2 \geqslant \frac{1}{2}(r + c_p)$.

Proof. Please refer to Appendix B. □

Remark 1. If the response of the high range system improves, that is, $c_2 = \frac{1}{2}(r + c_p) - \epsilon$, and $ACE = \frac{1}{2}$ (e.g. 2-point distribution), then Theorem 6 no longer holds.

4.2. Upgrading response and reducing demand variability

The previous section shows the need to improve the response dimension of a system. This means reducing c_s to the level of c_p. A natural question to ask is how much benefit does this bring? Using Theorem 3 and results from Chou et al. (2010), we compare the asymptotic chaining efficiencies for high and low c_s for some (discrete and continuous) uniform and normal distributions. It is easy to see that for a discrete uniform distribution with 2Δ possible demand values, we have

$$\mathrm{E}[\min(Y_1, Y_2)] = \frac{(\Delta+1)(2\Delta+1)}{6\Delta^2}\cdot\mu$$

$$\mathrm{E}[Y_1] = \frac{\Delta+1}{2\Delta}\cdot\mu$$

$$\mathrm{ACE}(c_s) = \frac{\mathrm{E}[\min(Y_1, Y_2)]}{2\mathrm{E}[Y_1]} = \frac{2\Delta+1}{6\Delta} = \frac{1}{3}+\frac{1}{6\Delta} \quad \text{for high } c_s$$

$$\mathrm{ACE}(c_p) = \frac{7\Delta+2}{12\Delta+6} = \frac{7}{12}-\frac{1}{8\Delta+4} \quad \text{when} \quad c_s = c_p$$

$$\mathrm{CV} = \sqrt{\frac{2\Delta^2+3\Delta+1}{6\Delta^2}}$$

We tabulate the results for some values of Δ.

Δ	Distribution	CV	ACE (c_s) for high c_s	ACE (c_p)	Improvement
1	2-point	1.00	0.5000	0.5000	0.0000
2	4-point	0.79	0.4167	0.5333	0.1166
3	6-point	0.72	0.3889	0.5476	0.1587
4	8-point	0.68	0.3750	0.5556	0.1806
⋮	⋮	⋮	⋮	⋮	⋮
∞	continuous	0.58	0.3333	0.5833	0.2500

Similar results hold for normal distributions. Recall that ACE for low response is independent of CV.

CV	ACE (c_s) for high c_s	ACE (c_p)	Improvement
0.33	0.2929	0.7022	0.4093
0.31	0.2929	0.7145	0.4216
0.29	0.2929	0.7275	0.4346
0.27	0.2929	0.7413	0.4484
0.25	0.2929	0.7558	0.4629
0.23	0.2929	0.7708	0.4779
0.21	0.2929	0.7864	0.4935

These results suggest that upgrading system response brings more benefits as the demand coefficient of variation decreases. Although upgrading response is important, it becomes even more so if coupled with initiatives to reduce demand uncertainty. This contradicts the intuition that flexibility becomes less valuable under an environment with less uncertainty.

5. Conclusions

In this paper, we have introduced a new way of studying process flexibility by modeling its response dimension in addition to the traditional range dimension considered in the literature. We model system response in terms of uniformity of production cost. To this end, we distinguish between primary and secondary production, where low secondary production cost means high response, while high secondary cost means low response.

While we have shown that even a system with low range (chaining) and low response (high secondary cost) can already produce non-negligible returns relative to full flexibility, upgrading either system range or system response (or both) can clearly improve system performance. However, if one faces limited resources and has to choose investment between range flexibility and response flexibility, our results suggest that a system with high response but limited range performs at least as well as a system with high range but low response. This implies that one should focus first on upgrading system response and then use residual resources to widen system range.

The model can be enriched by considering endogenized pricing and/or capacity decisions. On the demand side, the distribution can also be extended to include correlated demands. Furthermore, an interesting direction to look at is the effect of competition (say, in an oligopoly) on the behavior of the flexibility-conscious firm. We leave these issues for future research.

Acknowledgments

This research was supported in part by A*STAR Grant 521160079 and NUS Academic Research Fund R-314-000-082-112.

Appendix A. Proof of Theorem 1

Regardless of the value of c_s, it is easy to derive the optimal production allocations for both the dedicated and the fully flexible structures. For the dedicated structure, each facility j can only produce its designated set of primary products $\{i|\phi(i)=j\}$. Because this set does not overlap with those of other facilities, the optimal allocation is for the facility to produce as many units of these primary products as possible, such that

$$\sum_{i:\phi(i)=j} x_{ij}^* = \min\left(\sum_{i:\phi(i)=j} D_i, S_j\right).$$

For each product that belongs to this set, we can allocate proportionately as follows

$$x_{ij}^* = \min\left(\sum_{i:\phi(i)=j} D_i, S_j\right)\cdot\frac{D_i}{\sum_{i:\phi(i)=j} D_i} \leqslant D_i \forall i : \phi(i)=j \tag{10}$$

We repeat this allocation procedure for all other facilities.

For the fully flexible structure, any facility can produce any product. Thus, it is optimal for each facility to produce as many units as possible of its primary products, and *only* thereafter, use its extra capacity, if any, to produce the extra demand, if any, of secondary products

$$x_{ij}^* = \min\left(\sum_{i:\phi(i)=j} D_i, S_j\right)\cdot\frac{D_i}{\sum_{i:\phi(i)=j} D_i} \leqslant D_i \forall i : \phi(i)=j, \quad \forall\, j=1,\ldots,n \tag{11}$$

$$\sum_{i=1}^{m}\sum_{j\neq\phi(i)} x_{ij}^* = \min\left(\sum_{i=1}^{m} D_i, \sum_{j=1}^{n} S_j\right) - \sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right) \tag{12}$$

To show our result, consider $c_s > c_s'$. For a fixed structure $\mathcal{G}(m,n)$ and a demand realization $\mathbf{D}$, we let $X_p = \sum_{i=1}^{m} x_{i,\phi(i)}$ and $X_s = \sum_{i=1}^{m}\sum_{j\neq\phi(i)} x_{ij}$ be the optimal primary and secondary production, respectively, when secondary production cost is c_s. Similarly, X_p' and X_S' are the optimal primary and secondary production when secondary production cost is c_s'. From model (1), Eqs. (10)–(12), and the definition of flexibility efficiency, we obtain the following:

$$\mathrm{FE}(\mathcal{G}(m,n),c_s) = \frac{\mathrm{E}\left[(r-c_s)X_s - (r-c_p)\left(\sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right) - X_p\right)\right]}{\mathrm{E}\left[(r-c_s)\left(\min\left(\sum_{i=1}^{m} D_i, \sum_{j=1}^{n} S_j\right) - \sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right)\right)\right]}$$

Hence,

$$\begin{aligned}\mathrm{FE}(\mathcal{G}(m,n),c_s) &= \frac{\mathrm{E}\left[X_s - \left(\frac{r-c_p}{r-c_s}\right)\left(\sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right) - X_p\right)\right]}{\mathrm{E}\left[\min\left(\sum_{i=1}^{m} D_i, \sum_{j=1}^{n} S_j\right) - \sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right)\right]} \leqslant \frac{\mathrm{E}\left[X_s - \left(\frac{r-c_p}{r-c_s'}\right)\left(\sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right) - X_p\right)\right]}{\mathrm{E}\left[\min\left(\sum_{i=1}^{m} D_i, \sum_{j=1}^{n} S_j\right) - \sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right)\right]} \\ &\leqslant \frac{\mathrm{E}\left[(r-c_s')X_s - (r-c_p)\left(\sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right) - X_p\right)\right]}{\mathrm{E}\left[(r-c_s')\min\left(\sum_{i=1}^{m} D_i, \sum_{j=1}^{n} S_j\right) - \sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right)\right]} = \mathrm{FE}(\mathcal{G}(m,n),c_s')\end{aligned}$$

The first inequality is because $c_s > c_s'$ and X_p is bounded above by $\sum_{j=1}^{n}\min\left(\sum_{i:\phi(i)=j} D_i, S_j\right)$. The second inequality results from the feasibility of (X_p, X_s) to model (1) when secondary cost is c_s'.

Appendix B. Proof of Theorem 6

Using Lemma 4, Eq. (10), and $c_2 \geqslant \frac{1}{2}(r+c_p)$,

$$\begin{aligned}\lim_{n\to\infty}\frac{\mathrm{E}[\Pi(\mathcal{S}_2(n),\mathbf{D},c_2)]}{\mathrm{E}[\Pi(\mathcal{F}(n),\mathbf{D},c_p)]} &= \lim_{n\to\infty}\frac{(r-c_p)\mathrm{E}\left[\sum_{i=1}^{n}\min(D_i,\mu)\right] + (r-c_2)\mathrm{E}\left[\min\left(\sum_{i=1}^{n} D_i, n\mu\right) - \sum_{i=1}^{n}\min(D_i,\mu)\right]}{(r-c_p)\mathrm{E}\left[\min\left(\sum_{i=1}^{n} D_i, n\mu\right)\right]} \\ &\leqslant \lim_{n\to\infty}\frac{(r-c_p)\mathrm{E}\left[\sum_{i=1}^{n}\min(D_i,\mu)\right] + \frac{1}{2}(r-c_p)\mathrm{E}\left[\min\left(\sum_{i=1}^{n} D_i, n\mu\right) - \sum_{i=1}^{n}\min(D_i,\mu)\right]}{(r-c_p)\mathrm{E}\left[\min\left(\sum_{i=1}^{n} D_i, n\mu\right)\right]} \\ &= \lim_{n\to\infty}\frac{\frac{1}{2}\mathrm{E}\left[\min\left(\sum_{i=1}^{n} D_i, n\mu\right) + \frac{1}{2}\mathrm{E}\left[\sum_{i=1}^{n}\min(D_i,\mu)\right]\right.}{\mathrm{E}\left[\min\left(\sum_{i=1}^{n} D_i, n\mu\right)\right]} = \frac{1}{2} + \frac{1}{2}\lim_{n\to\infty}\frac{\mathrm{E}\left[\sum_{i=1}^{n}\min(D_i,\mu)\right]}{\mathrm{E}\left[\min\left(\sum_{i=1}^{n} D_i, n\mu\right)\right]} = 1 - \frac{1}{2}\cdot\frac{E[(\mu-D_i)^+]}{\mu} \\ &\leqslant 1 - (1-\mathrm{ACE})\cdot\frac{E[(\mu-D_i)^+]}{\mu} = \lim_{n\to\infty}\frac{\mathrm{E}[\Pi(\mathcal{S}_1(n),\mathbf{D},c_1)]}{\mathrm{E}[\Pi(\mathcal{F}(n),\mathbf{D},c_p)]}\end{aligned}$$

References

Akşin, O., Karaesmen, O., 2007. Characterizing the performance of process flexibility structures. Operations Research Letters 35 (4), 477–484.

Bassamboo, A., Randhawa, R., Van Mieghem, J.A., 2009. A little flexibility is all you need: Optimality of tailored chaining and pairing. Working paper, Northwestern University. <http://www.kellogg.northwestern.edu/faculty/vanmieghem>.

Beach, R., Muhlemann, A.P., Price, D.H.R., Paterson, A., Sharp, J.A., 2000. A review of manufacturing flexibility. European Journal of Operational Research 122, 41–57.

Bish, E., Wang, Q., 2004. Optimal investment strategies for flexible resources, considering pricing and correlated demand. Operations Research 52 (6), 954–964.
Browne, J., Dubois, D., Rathmill, K., Sethi, S., Stecke, K., 1984. Classificiation of flexible manufacturing systems. FMS Magazine (April), 114–117.
Buzacott, J., 1982. The fundamental principles of flexibility in manufacturing systems. Proceedings of the First International Conference on Flexible Manufacturing Systems, 13–22.
Chandra, P., Tombak, M.M., 1992. Models for the evaluation of routing and machine flexibility. European Journal of Operational Research 60, 152–165.
Chou, M.C., Teo, C.P., Zheng, H., 2008. Process flexibility: Design, evaluation and applications. Flexible Services and Manufacturing Journal 20 (1–2), 59–94.
Chou, M.C., Teo, C.P., Zheng, H., 2009. Process flexibility Revisited: The Graph Expander and its Applications. Working Paper, National University of Singapore. <http://www.bschool.nus.edu.sg/staff/bizteocp/processrevision9.pdf>.
Chou, M.C., Chua, G., Teo, C.P., Zheng, H., 2010. Design for process flexibility: Efficiency of the long chain and sparse structure. Operations Research 58 (1), 43–58.
DesMarteau, K., 1999. [TC]2: Leading the way in changing times. Bobbin 41 (2), 48–54.
Eppen, G., 1979. Effects of centralization on expected costs in a multilocation newsboy problem. Management Science 25 (5), 498–501.
Fine, C., Freund, R., 1990. Optimal investment in product-flexible manufacturing capacity. Management Science 36 (4), 449–466.
Graves, S., Tomlin, B., 2003. Process flexibility in supply chains. Management Science 49 (7), 907–919.
Gupta, Y.P., Goyal, S., 1989. Flexibility of manufacturing systems: concepts and measurements. European Journal of Operational Research 43, 119–135.
Gurumurthi, S., Benjaafar, S., 2004. Modeling and analysis of flexible queueing systems. Naval Research Logistics 51, 755–782.
Hopp, W., Tekin, E., Van Oyen, M., 2004. Benefits of skill chaining in serial production lines with cross-trained workers. Management Science 50 (1), 83–98.
Iravani, S., Van Oyen, M., Sims, K., 2005. Structural flexibility: a new perspective on the design of manufacturing and service operations. Management Science 51 (2), 151–166.
Jaikumar, R., 1986. Postindustrial manufacturing. Harvard Business Review 64 (6), 69–76.
Jordan, W., Graves, S., 1995. Principles on the benefits of manufacturing process flexibility. Management Science 41 (4), 577–594.
Kusiak, A., 1986. Applications of operational research models and techniques in flexible manufacturing systems. European Journal of Operational Research 24, 336–345.
Mandelbaum, M., 1978. Flexibility in decision making: an exploration and unification. Ph.D. Thesis, Department of Industrial Engineering, University of Toronto, Canada.
McCutcheon, D., 2004. Flexible manufacturing: IBM's Bromont semiconductor packaging plant. Canadian Electronics 19 (7), 26.
Parker, R.P., Wirth, A., 1999. Manufacturing flexibility: measures and relationships. European Journal of Operational Research 118, 429–449.
Primrose, P.L., Leonard, R., 1984. Conditions under which flexible manufacturing systems is financially viable. Proceedings of the Third International Conference on Flexible Manufacturing Systems, 121–132.
Ross, S., 2003. Introduction to Probability Models. Academic Press.
Sethi, A., Sethi, S., 1990. Flexibility in manufacturing: a survey. International Journal of Flexible Manufacturing Systems 2, 289–328.
Slack, N., 1987. The flexibility of manufacturing systems. International Journal of Operations and Production Management 7 (4), 35–45.
Upton, D.M., 1994. The management of manufacturing flexibility. California Management Review 36 (2), 72–89.
Van Biesebroeck, J., 2007. Complementarities in automobile production. Journal of Applied Econometrics 22 (7), 1315–1345.
Van Mieghem, J., 1998. Investment strategies for flexible resources. Management Science 44 (8), 1071–1078.
Wahab, M.I.M., Wu, D., Lee, C.G., 2008. A generic approach to measuring the machine flexibility of manufacturing systems. European Journal of Operational Research 186, 137–149.
Wall, M., 2003. Manufacturing flexibility. Automotive Industries 183 (10), 44–45.
Wallace, R.B., Whitt, W., 2005. A staffing algorithm for call centers with skill-based routing. Manufacturing and Service Operations Management 7 (4), 276–294.