Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Guilt by Association: Strategic Failure Prevention and Recovery Capacity Investments

Sang-Hyun Kim, Brian Tomlin,

To cite this article:
Sang-Hyun Kim, Brian Tomlin, (2013) Guilt by Association: Strategic Failure Prevention and Recovery Capacity Investments. Management Science 59(7):1631-1649. http://dx.doi.org/10.1287/mnsc.1120.1658

MANAGEMENT SCIENCE
Vol. 59, No. 7, July 2013, pp. 1631–1649
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1120.1658

Guilt by Association: Strategic Failure Prevention and Recovery Capacity Investments

Sang-Hyun Kim
Yale School of Management, Yale University, New Haven, Connecticut 06511,
sang.kim@yale.edu

Brian Tomlin
Tuck School of Business at Dartmouth, Hanover, New Hampshire 03755,
brian.tomlin@tuck.dartmouth.edu

We examine technological systems that display the following characteristics: (i) unplanned outages incur significant costs, (ii) the system experiences an outage if one or more of its subsystems fails, (iii) subsystem failures may occur simultaneously, and (iv) subsystem recovery requires specific resources and capabilities that are provided by different firms. Firms may invest in two measures to enhance system availability—namely, failure prevention and recovery capacity. If recovery capacity investment is the only option, we find that the firms in a decentralized setting overinvest in capacity, resulting in higher system availability but at a higher cost. If both investments can be made, we find that the firms underinvest in failure prevention and overinvest in recovery capacity. That is, firms in a decentralized setting shift their focus from preventing failures to responding to failures. The net effect is lower system availability, reversing the conclusion above. We also find that, unexpectedly, firms are more willing to let their subsystems fail if a joint failure is more likely to occur.

Key words: supply chain disruption; business continuity; system reliability; game theory
History: Received November 2, 2011; accepted August 31, 2012, by Martin Lariviere, operations management.
Published online in *Articles in Advance* February 15, 2013.

1. Introduction

Maintaining a high level of system availability is of the utmost importance to many firms operating critical equipment, because prolonged outages can significantly impact their short-term and long-term profitability. In the energy industry, for instance, "unplanned downtime is exceedingly costly, whether from lost revenues, O&M [operations and maintenance] costs, contractual penalties, or regulatory fines" (GE Energy 2010, p. 2). British Energy lost 18% of its annual generating capacity as a result of unplanned outages in its power plants between 2000 and 2008 (British Energy 2008), with outages in 2003 alone causing British Energy to seek and receive a £75 million increase in its credit agreement with the UK government (Ford 2003). In the airline industry, it is estimated that an "aircraft-on-ground" (AOG) due to mechanical, electrical, or other types of failures costs an airline company as much as $150,000 per hour.[1] The consequences of a system outage go beyond financial impacts. The reputation of British Airports Authority, the operator of Heathrow Airport, was severely damaged in 2008 when a software glitch caused both baggage sorting machines in Terminal 4 to fail; as a result, passengers were allowed to travel with only hand luggage for two days (Chapman 2008).

Firms may enhance system availability in a number of ways. Whereas reducing the frequency of failures is an obvious route, availability can also be improved by shortening the time it takes to restore the system after a failure. How to prioritize between these availability enhancement options is an important decision to make, but it is further complicated by the facts that most systems are complex collections of subsystems (e.g., boilers and turbines in a power plant, engines and avionics in an aircraft, conveyor belts and software in a baggage handling system) and that these subsystems are often interconnected. System failure can result from a failure of even one subsystem, and different subsystems typically require distinct resources and capabilities for restoration. Moreover, multiple subsystems can fail simultaneously, either because some external shock damages multiple subsystems or because the failure of one subsystem causes the knock-on failure of another subsystem (see Scudder 1984, Palmer and Danaher 2004, Boyko and Popov 2010, and Milmo 2010 for examples of knock-on failures in power plants and aircraft systems). System availability therefore depends not only on the investments made in failure prevention and

[1] Boeing Commercial Airplanes Operations Center, http://www.boeing.com/commercial/global/opscenter.html.

recovery across multiple subsystems but also on the nature of failure interactions between subsystems.

System governance adds another layer of complexity. In some cases, a single firm is liable for any outage consequences and has maintenance/restoration responsibility for all subsystems, but in other cases, the overall system outage consequence flows to multiple firms who have responsibilities for different subsystems. Power plant builders such as Alstom, General Electric, and Siemens offer contractual service agreements (CSAs) to plant owners in which each provides unplanned maintenance coverage for all subsystems in a plant. Lacking certain technological capabilities, however, these companies have partnered with solar companies when building and maintaining solar thermal power plants. For example, Alstom and BrightSource Energy signed an agreement in 2010 to build and service integrated power plants (Alstom 2010). Similarly, different types of system governance exist for the maintenance of airport baggage handling systems. In some cases a single original equipment manufacturer (OEM) is responsible for the entire system, as was the case with Siemens' maintenance contract for Terminal 2 in Munich International Airport; in other cases joint CSAs involving multiple OEMs are also observed, as was the case with Siemens and Elsag Datamat's joint maintenance contract for Fiumicino Airport in Rome (*Airport Logistics* 2010).

The facts outlined above—that unplanned system outages incur significant costs, that simultaneous subsystem failures can occur, that subsystem recovery requires specific resources and capabilities, and that the recovery processes may be managed by different firms—present significant challenges for firms investing in failure prevention and recovery capacity. Not only must firms wrestle with tradeoffs between failure prevention and recovery capacity investments within a subsystem, they must also account for the possibility of joint subsystem failures and the strategic interactions between the firms responsible for different subsystems. From our conversations with practitioners in the aerospace and defense industry, we learned that they are indeed concerned about these issues as the entire industry moves toward performance-based service outsourcing models in which subsystem providers are compensated based on aircraft availability. Motivated by these examples, we seek to find answers to a set of questions that, to our knowledge, have not been addressed in prior research. Compared with the centralized setting in which a single firm manages all subsystems, do firms in a decentralized setting invest more or less in failure prevention and recovery capacity? Does the answer depend on how system outage cost is allocated among the firms? How does the probability of joint subsystem failures impact system efficiency?

We answer these questions by constructing and analyzing a stylized game-theoretic model that captures the system characteristics described above. One of the key features of our model is that the possibility of simultaneous subsystem failures creates an ambiguity in accounting for each firm's contribution to system performance. That is, when failure probabilities are *associated*, there is no clear-cut way to disentangle each firm's contribution to the system outage (hence, "guilt by association"), whose duration can be shortened by the inputs from all firms. As we demonstrate, the allocation rule employed to divide the system outage cost is critical in shaping the firms' incentives. In our model we lay out the set of conditions that allocation rules should possess in order to represent firm accountability, present their properties and interpretations, and derive general insights that naturally arise from them. Our main findings are as follows.

1. Firms in a decentralized system overinvest in recovery capacity compared with the centralized system if failure probabilities cannot be changed. Consequently, decentralization leads to *higher* system availability, but this benefit comes at a higher overall cost.

2. When firms may invest in failure prevention as well as recovery capacity, however, the above conclusion is reversed: decentralization leads to lower system availability. We find that, with an option to invest in both measures, decentralization shifts a firm's focus from preventing failures to responding to failures.

3. The possibility of joint subsystem failures influences centralized and decentralized systems differently. In particular, when a firm in a decentralized system faces a greater probability that its subsystem's failure cascades to another, it is more willing to let this happen, thus lowering system availability.

2. Related Literature

Our model builds on ideas found in a number of related but independently established literatures, including those of system reliability theory, public goods economics, and the operations management (OM) studies on disruption management.

System reliability theory (Barlow and Proschan 1975, Rausand and Høyland 2004) concerns itself with how a system's overall reliability is evaluated from its internal structure (e.g., components or subsystems in series and/or parallel configurations) given the random nature of failure processes, and what type of inspection and component replacement policies should be employed. A closely related area is probabilistic risk assessment (PRA), "a method widely used in engineering and in other fields to assess the probabilities and consequences of failures of a given

system" (Paté-Cornell 2007, p. 224). Sophisticated models and tools have been developed based on these streams of literature with a wide range of applications to quality control, construction, product development, and environmental protection, to name just a few. We borrow several concepts from system reliability theory—in particular, a model of dependent subsystem failures in a serial system—but add a new element of decentralized decision making; this aspect is largely ignored in this literature, which has traditionally focused on evaluating the data-driven statistical reliability measures from an engineering perspective.

There is a small but significant body of literature that connects system reliability theory and PRA with game theory, in the context of protecting a critical infrastructure from such risks as terrorist attacks. Bier (2005) provides an excellent survey of this area. A typical model found in this literature has a setup in which an attacker chooses a target within a system that has series or parallel structures to maximize the probability of disabling the system, whereas a defender invests in resources to protect vulnerable points in the system to minimize the chance of a successful attack. In contrast, our model does not feature a game between an attacker and a defender with opposing objectives; instead, we consider a game between parties who independently manage the subsystems and whose collective actions determine the performance and the cost at the system level. The article by Kunreuther and Heal (2003) is related to this paper in this regard, for they study a multiagent "interdependent security" game in which an externality is created by the probability of a spillover effect. However, many aspects of these two papers differ, including the problem focus and the modeling approach.

A stream of literature exists in economics that investigates how individual decisions on public good consumption are influenced by an aggregated output that exhibits either the weakest-link property or the best-shot property, in contrast to the additive property that is assumed in the majority of public good economics models. This literature is relevant to our model (which builds on the system reliability theory) because the weakest-link and best-shot properties are present in a serial system and a parallel system, respectively. In his seminal paper, Hirshleifer (1983) considers public goods with the production functions having each of the three properties above and finds that the degree of underprovision of goods (compared with the socially optimal level) is the smallest under the weakest-link property and the highest under the best-shot property. Cornes (1993) extends this analysis by incorporating heterogeneity in individual preferences. At the intersection of this literature and system reliability theory are Hausken (2002) and Varian (2004). Both consider strategic interactions among players whose utilities are linked in a serial or a parallel structure, and they analyze the equilibrium of the utility maximization games. The games presented in these papers are quite simple (deterministic utilities, binary decisions in Hausken 2002, etc.), and moreover, they focus exclusively on the decisions leading up to a system failure (or a nonfailure). Our work has a fundamentally different focus because we pay special attention to what happens afterwards, i.e., system recovery, and in addition, we uncover subtle and interesting insights based on a probabilistic analysis.

Various strategies for managing the risk of production system failures have been studied in the OM literature. Supplier diversification, under which multiple vendors are used to supply the same component, offers benefits that are analogous to those in a parallel system structure in that the system functionality (order delivery) breaks down if and only if all suppliers fail; see Tomlin (2006) and Dada et al. (2007). Emergency sourcing is an alternative option in which a backup supplier is used in the event of a failure to the primary supplier; see Chopra et al. (2007) and Tomlin (2009). Wang et al. (2010) consider process improvement investments to reduce supply risks. Although these papers share the same broad theme as in ours, they take a centralized-system perspective and do not consider strategic interactions among different firms. As surveyed in Aydin et al. (2012), a number of papers consider supply risk management in decentralized systems. Bakshi and Kleindorfer (2009), for example, consider a bargaining game between a retailer and a supplier who face an interdependent supply chain disruption risk. Most of these explore buyer–supplier, vertical coordination settings, but some (e.g., Babich et al. 2007, Yang et al. 2012) study horizontal competition among multiple suppliers. These papers differ from ours in that they develop their models in production distribution settings (as opposed to failure recovery settings), and they do not investigate incentive effects originating from dependent subsystem failures in a serially linked system, which is the main focus of this paper.

Finally, there are a number of related papers that do not fall exactly into the above categories. Kim et al. (2010) study a problem in which a supplier is incentivized to reserve a system recovery capacity through a performance-based contract that penalizes him for system downtime. However, their focus is on a contracting mechanism in a vertical customer–supplier relationship, instead of the horizontal competition that takes place in our model. Hu et al. (2013) consider a similar setting but study other contracting strategies. Kwon et al. (2010) feature a model of capacity ("work rate") competition in a project management

setting. Our model features a similar capacity competition, but it diverges significantly from Kwon et al. (2010) because of the differences in contexts and research questions.

3. Model

Building on the discussion in §1, we model a system with the following characteristics: (i) unplanned outages incur significant costs, (ii) the system experiences an outage if one or more of its subsystems fails, (iii) subsystem failures may occur simultaneously, and (iv) subsystem recovery requires specific resources and capabilities that are provided by different firms. To achieve model parsimony, we suppress ancillary features and make some simplifying assumptions, as outlined below.

3.1. System Description

The system consists of two distinct subsystems, $i \in \{1, 2\}$, in a series structure. When referring to subsystem i, we use j to denote the other subsystem. For simplicity, we treat each subsystem as the smallest unit and do not explicitly model the internal structure of each subsystem. We consider two types of system management with regard to failure prevention and recovery. In a *centralized system*, the two subsystems are managed by a single firm, e.g., Siemens' exclusive maintenance contract for the baggage handling system in Munich International Airport (see §1), whereas in a *decentralized system*, the recovery of subsystem i is managed by firm i, e.g., Siemens and Valoriza's joint maintenance contract for the Lebrija solar thermal power plant in Spain.[2] There is no third party in the decentralized system that can play a mediator role. In addition, we assume that the two firms do not have a leader–follower relationship. No side payments are made between the firms, but such a situation is considered in §7.3 as an extension. All parameters are common knowledge in the decentralized case.

The system has a finite life, scaled to 1 without loss of generality, reflecting the economic life of the system. For the purposes of this paper, the *economic life* refers to the length of time for which the firms are responsible for outages (e.g., the duration of the CSA).

3.2. System Failure

Because of the series structure, the system is operational (up) if and only if both subsystems are up; a failure of either subsystem causes the system to fail (down). We focus on failure events that are sudden, unpredictable, and complete (i.e., disable a subsystem instantaneously) rather than gradual and/or partial failure events. Subsystem failures occur as a result of some rare and exogenous shock, which arrives to the system with probability $\lambda > 0$. The shock may originate from a natural phenomenon (an earthquake, lightning, etc.), a human error, or other random fluctuations in the operating environment. The probability λ represents the aggregate chances of all such events. For simplicity, we assume that λ is sufficiently small so that the probability of more than one shock arriving during the system lifetime is negligible.[3]

Both subsystems can fail as a result of the external shock. We focus on a type of joint failure known in system reliability theory as a *cascading failure*, in which the failure of one subsystem may trigger a chain reaction that causes the other subsystem to fail. Cascading failures are quite common, and they occur when a malfunction of one subsystem creates stress for other subsystems through increased pressure, temperature, humidity, and so on (Rausand and Høyland 2004). Another type of joint failure often studied in the reliability literature is the *common-cause failure*, which describes situations where multiple subsystems fail directly due to a common source. To obtain sharper insights, we focus exclusively on the cascading failure scenario by assuming that the probability of a common-cause failure is negligible.

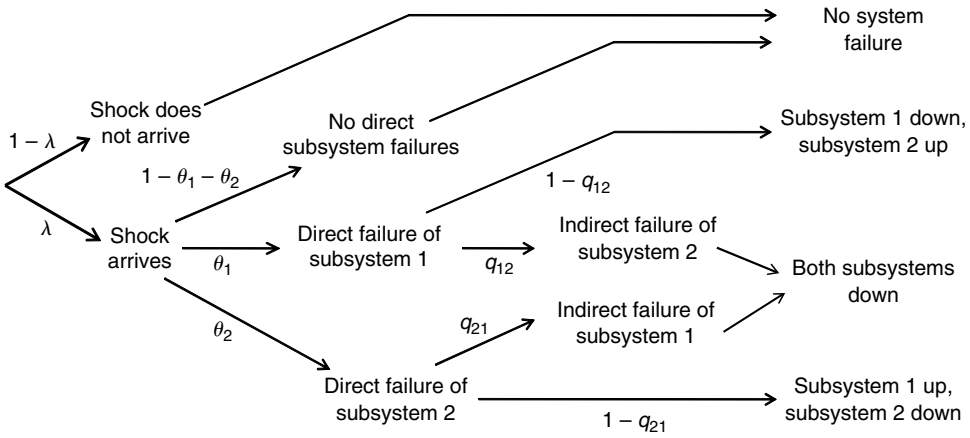
The external shock causes a *direct failure* of at most one subsystem, with $\theta_i > 0$ denoting the probability that subsystem i undergoes a direct failure. (Direct failures in the two subsystems are mutually exclusive as we rule out common-cause failures.) Therefore, the system is unharmed by the shock with probability $1 - \theta_1 - \theta_2$. We restrict the range of θ_1 and θ_2 such that $\theta_1 + \theta_2 < 1$. Subsequently, a direct failure of subsystem i leads to an instantaneous knock-on *indirect failure* of subsystem j with probability q_{ij}. As a result, the system experiences one of four possible scenarios during its lifetime (see Figure 1, which summarizes all cases): (i) no system failure, with probability $1 - (\theta_1 + \theta_2)\lambda$; (ii) system failure due to subsystem 1 down and subsystem 2 up, with probability $p_1 \equiv (1 - q_{12})\theta_1\lambda$; (iii) system failure due to subsystem 1 up and subsystem 2 down, with probability $p_2 \equiv (1 - q_{21})\theta_2\lambda$; and (iv) system failure due to both subsystems down, with probability $p_b \equiv (q_{12}\theta_1 + q_{21}\theta_2)\lambda$.

The probability θ_i reflects the *fragility* of subsystem i; the higher the value of θ_i, the smaller the chance that subsystem i can withstand the shock and

[2] "Siemens provided Lebrija's complete solar field, with all the solar receivers, collectors and mirrors, as well as the steam turbine. Valoriza provided the civil works, power block, and heat transfer fluid system . . . . The two companies are jointly responsible for operation and maintenance" (Siemens 2012).

[3] This assumption is not overly restrictive because, with some additional assumptions (e.g., Poisson shock arrivals), the analysis can be extended to a case with multiple shock arrivals without affecting the insights.

Figure 1 Event Diagram of Subsystem Failures



prevent the system from failing. The probability q_{ij} reflects subsystem j's *susceptibility* to a subsystem i failure; the higher the value of q_{ij}, the more subsystem j is prone to an indirect failure originating from subsystem i. We assume that the system is *cascade-prone*; that is, indirect failures occur with positive probabilities ($q_{12} > 0$ and/or $q_{21} > 0$). Furthermore, we assume that the susceptibilities q_{12} and q_{21} are exogenously given. This is a reasonable assumption especially in a decentralized system, because an indirect failure propagates through a common interface (e.g., electrical cable) whose physical properties are determined by system requirements for coordination functions and cannot be modified unilaterally by one firm. In contrast, we allow for the possibility that the fragilities θ_1 and θ_2 are endogenously determined by discretionary firm investments; this is considered in §5. We also study the impact of varying susceptibilities in §6.

3.3. System Recovery

Failures are not catastrophic in that a full system recovery can be made through repair of an affected subsystem. In the event that subsystem i fails, it takes a random amount of time Y_i to restore this subsystem from the disabled state to the operational state. As in numerous other papers (e.g., Tomlin 2006 and references therein), we assume that Y_i is exponentially distributed with parameter μ_i, so the expected value of Y_i, i.e., the mean-time-to-repair (MTTR), is given by $1/\mu_i$. Nonexponential distributions do not fundamentally change the insights, but they complicate the analysis significantly. We assume that Y_1 and Y_2 are independent. The system outage duration, i.e., the time from system failure until it is once again operational, is a random variable T that depends on the failure scenario and the subsystem recovery times Y_1 and Y_2. If only subsystem i fails, then $T = Y_i$, but $T = \max\{Y_1, Y_2\}$ if both subsystems fail. Throughout the paper we use the uppercase Y_1, Y_2, and T to denote random variables and lowercase y_1, y_2, and t to denote their realizations. We also use the notation τ for the expected system outage duration, which is weighted by the failure probabilities p_1, p_2, and p_b (see Equation (2) for a precise definition).

System availability is defined as the percentage uptime of the system. Because we normalize the length of system life to 1, system availability is simply $1-\tau$, and so availability increases as the expected outage duration τ decreases. Availability can be enhanced by reducing the probabilities of subsystem failures or by shortening subsystem recovery times (i.e., reducing the MTTRs). The latter can be achieved through investments in such resources as field service technicians and prepositioned spare parts. These resources cannot be shared between subsystems because the required expertise and parts are subsystem specific. For example, boilers and turbines in a power plant are repaired by different technicians because these subsystems require distinct training and skill sets. Hardware and software repairs in a baggage handling system are carried out by different technicians. Similarly, avionic and engine subsystems of an aircraft share very little common parts, the former being mostly electronic whereas the latter mainly mechanical, even though they are interconnected to enable flight controls.

The recovery capacity for subsystem i is measured by μ_i, the inverse of MTTR. We assume that subsystem i possesses a baseline recovery capacity of $\underline{\mu}_i > 1$; that is, there is sufficient existing capacity so that the baseline subsystem MTTR is less than the system life. The firms make discretionary decisions on how much additional recovery capacity to install, with subsystem i capacity costing κ_i per unit. Hence, firm i incurs the cost $\kappa_i(\mu_i - \underline{\mu}_i)$ for increasing its capacity from $\underline{\mu}_i$ to μ_i. The linear cost function is employed for analytical tractability, and a similar assumption has been made in other papers (e.g., Kim et al. 2010, Cachon and Feldman 2011). In §8.1 we

relax the linearity assumption and demonstrate that the insights are not altered. We assume that κ_i does not depend on the system governance (centralized or decentralized). Because the subsystems require distinct resources for restoring them, the capacities cannot be shared between them. Without loss of generality, we also assume that the feasible recovery capacity μ_i is bounded above by some arbitrarily large number $\bar{\mu}_i$. This ensures that in the decentralized system, each firm's strategy space $[\underline{\mu}_i, \bar{\mu}_i]$ is compact, which facilitates the game-theoretic analysis.

In addition to recovery capacity, firms may invest in failure prevention to reduce subsystem fragility θ_i and increase system availability. We postpone the description of this strategy until §5.

3.4. System Outage Penalty and Allocation

System outages have financial and reputational consequences that depend on the outage duration. For example, many service-level agreements found in the information technology, aerospace and defense, heavy equipment, and other industries stipulate a penalty based on the duration of system downtime or on system availability. Moreover, damage to a system operator's reputation accrues as system recovery is delayed. Motivated by these examples, we assume that the system incurs an outage penalty at a rate of π per unit time that the system is down. The time-independent constant penalty rate assumption is reasonable—especially if π represents the opportunity cost of lost revenue—and promotes analytical tractability. (For completeness, in §8.2 we study the impact of having time-dependent penalty rates.) An outage may also incur a fixed cost, but we omit it in our model to focus on the impact of time-based penalty. Then, as we ignore time discounting, a systemwide penalty equal to $\Pi(t) = \pi t$ is incurred in the event of an outage that lasts t time units. We assume that π is sufficiently larger than the capacity investment cost κ_i, which allows us to focus on the interior solutions in the subsequent analyses. We assume that if an outage occurs and recovery is not complete by time 1 (the economic life of the system), then the penalty still accrues until recovery is complete. For example, if an outage occurs in the last week of a maintenance contract, the party responsible for recovery would typically still bear contractual penalties until the system is brought back up.

For a given outage duration realization t, the systemwide penalty function $\Pi(t)$ is invariant to whether the system is centralized or decentralized. This reflects the fact that the economic consequence of an outage is measured at the system level regardless of the internal management structure. In a centralized system there is only one firm, so this firm internalizes the full systemwide penalty $\Pi(t)$. In a decentralized system, the penalty is assigned to the two firms according to a prescribed rule, which defines the penalty allocations $\Pi_1(y_1 \mid y_2)$ and $\Pi_2(y_2 \mid y_1)$. (Such a rule is necessary, for example, when the firms devise a joint CSA under which they promise to reimburse their customer together for the lost output $\Pi(t)$; we do not formally model how the rule is negotiated between the firms.) We adopt the convention of setting $y_j = 0$ when a system outage occurs due only to the failure of subsystem i.

In the event that an outage is caused by the failure of a single subsystem, there is no ambiguity in allocating the penalty: the firm responsible for restoring the failed subsystem should absorb the full penalty $\Pi(t)$. However, matters are not as clear-cut in the event that both subsystems suffer failures. What is each firm's contribution to the system outage when both are down? Should the penalty be equally divided in half? How should the penalty be allocated if one firm finishes recovery before the other? It is critical for the two firms to agree on a penalty allocation rule that resolves these issues.[4] For expositional clarity, we develop the model under the assumption that the chance of timely and undisputable root cause discovery is negligible and rule out ex post side compensations. We relax this assumption in §7.3 and examine the consequences.

Although the ambiguity created by simultaneous failures allows for many different ways to allocate the outage penalty, some allocation rules are not reasonable from a practical perspective. Hence, we focus on allocation rules that satisfy a minimal set of conditions that represent firm accountability.

Assumption 1. *For all* $y_i, y_j \geq 0$ *and* $i = 1, 2$, $i \neq j$, (i) $\Pi_i(y_i \mid y_j) + \Pi_j(y_j \mid y_i) = \Pi(\max\{y_i, y_j\})$; (ii) $\Pi_i(y_i = x \mid y_j = z) = \Pi_j(y_j = x \mid y_i = z)$; (iii) $\Pi_i(y_i \mid y_j) \leq \Pi(y_i)$; *and* (iv) $\Pi_i(y_i \mid y_j)$ *increases in* y_i.

Condition (i) ensures that the two firms together are fully responsible for the total systemwide outage penalty, which is proportional to the outage duration of the last-to-recover subsystem, $\max\{y_i, y_j\}$. Condition (ii) ensures impartiality; that is, the allocation depends strictly on subsystem recovery times and does not discriminate between firms based on their identity. This is reasonable under the earlier assumption that the root cause of a joint failure cannot be identified. Condition (iii) ensures that a firm is held accountable only for the duration of system outage

[4] This is especially true in many real-world situations where finding the exact chain of events that led to a cascading failure is extremely difficult. As Bier (1997, p. 7) notes, "The act of performing a root cause analysis does not ensure that the root causes are actually found." Even if the root cause can be identified, an allocation rule is still relevant because the outage penalty would have to be somehow shared as long as each contributes to the delay in system recovery.

to which it has contributed; the firm is not liable for the system-level penalty (if any) accrued after its subsystem is restored. Condition (iv) requires that a firm receives a higher penalty if its subsystem recovery is delayed, all else being equal. All stated conditions are reasonable. Note that the conditions apply to all failure scenarios—they are trivially satisfied in the case where a system outage is due to a failure of only one subsystem. Observe that condition (iii) implies that the penalty for the firm that recovers first does not exceed $\pi \min\{y_i, y_j\}$, the system penalty accrued while both subsystems are down. It then follows that we can express firm i's penalty $\Pi_i(y_i \mid y_j)$ as

$$\Pi_i(y_i \mid y_j) = \pi g_i(y_i \mid y_j) \min\{y_i, y_j\} + \pi(y_i - y_j)^+, \quad (1)$$

where $g_i(y_i \mid y_j) \le 1$ is the fraction of $\pi \min\{y_i, y_j\}$ assigned to firm i and where $(\cdot)^+ \equiv \max\{0, \cdot\}$. The fraction $g_i(y_i \mid y_j)$ may depend on the recovery times y_i and y_j. To facilitate analysis, we assume that $g_i(y_i \mid y_j)$ is monotonic and twice differentiable everywhere.

Finally, we impose the following condition on the allowed allocation rules to further operationalize firm accountability.

Assumption 2. *For* $y_i < y_j$, $\dfrac{\partial \Pi_i(y_i \mid y_j)/\partial y_j}{\partial \Pi_i(y_i \mid y_j)/\partial y_i} < \dfrac{y_i}{y_j}$.

With i denoting the faster firm, i.e., $y_i < y_j$, this condition ensures that a change in the penalty for the faster firm is more sensitive to that firm's own performance (recovery time) than it is to the slower firm's performance. Moreover, the relative sensitivity in general depends on performance outcomes (more pronounced if the performance gap is larger), as indicated by the ratio y_i/y_j in the condition. Assumption 2 precludes some undesirable characteristics of an allocation rule. For example, if an allocation rule violates Assumption 2, one may encounter an unrealistic situation where a firm is penalized (in expectation) for reducing the joint downtime $\min\{y_i, y_j\}$.

Assumptions 1 and 2 together accommodate a wide range of allocation rules, including two special ones that stand out because their simple and intuitive characteristics make them amenable to adoption in practice. These rules, analogous to two fair division principles discussed extensively in the social choice theory literature in the context of the "claims problem" (Young 1994), are as follows.

Allocation Rule 1 (A1; The "Contested Garment" Rule). *The penalty rate π is split equally for the duration in which both subsystems are down, but the firm that recovers later fully absorbs π for the duration in which only its subsystem is down. Thus,* $\Pi_i(y_i \mid y_j) = (\pi/2)\min\{y_i, y_j\} + \pi(y_i - y_j)^+$.

Allocation Rule 2 (A2; The "Proportional" Rule). *The total penalty $\pi \max\{y_i, y_j\}$ is allocated to the firms in proportion to each firm's respective subsystem recovery time. Thus,* $\Pi_i(y_i \mid y_j) = \pi(y_i/(y_i + y_j)) \max\{y_i, y_j\}$.

A1 is fair from the perspective of accounting for each firm's contribution to the system outage *at any given moment*; if the system is down due to failures of both subsystems, the systemwide penalty π for that moment is equally divided between the two firms. On the other hand, A2 is fair from the perspective of accounting for each firm's relative contribution toward the total system outage *at the completion of recoveries*, because the penalty share is proportional to the realized recovery time.[5] The logic behind A1 is analogous to that of the "Contested Garment" rule which dates back to Babylonian Talmud, whereas the proportionality principle of A2 is attributed to Aristotle (Young 1994).[6] At times in the paper we refer to these two rules to clearly convey the intuition for certain general results.

4. Investment in Recovery Capacity Only

In this section we study the optimal recovery capacity decisions assuming that the failure probabilities p_1, p_2, and p_b are exogenously given. This case represents the situation in which the failure prevention investment is infeasible, and therefore system availability can be enhanced only through recovery efforts, enabling us to isolate the outcome of firms' strategic decisions that are made without the first option. We examine the centralized system (§4.1) and the decentralized system (§4.2), and we then compare the outcomes (§4.3).

4.1. Centralized System

Given recovery capacities (μ_1, μ_2), the expected total cost for the firm in a centralized system is $\Psi_c(\mu_1, \mu_2) = \kappa_1(\mu_1 - \underline{\mu}_1) + \kappa_2(\mu_2 - \underline{\mu}_2) + E[\Pi(T)]$, where the first two terms reflect capacity investment costs and $E[\Pi(T)]$ is the expected outage penalty charge over the system life. Now, $E[\Pi(T)] = \pi\tau(\mu_1, \mu_2)$, where $\tau(\mu_1, \mu_2)$ is the expected outage duration. As discussed in §3, there are four outage scenarios: (i) no failure, in which case the outage duration $T = 0$; (ii) only subsystem 1 fails (with probability p_1), in which case $T = Y_1$; (iii) only subsystem 2 fails (with probability p_2), in which case $T = Y_2$; and (iv)

[5] A1 and A2 can be written in the form of (1) with $g_i(y_i \mid y_j) = 1/2$ and $g_i(y_i \mid y_j) = y_j/(y_i + y_j)$, respectively. The latter is true because $(y_i/(y_i + y_j)) \max\{y_i, y_j\} = (y_j/(y_i + y_j)) \min\{y_i, y_j\} + (y_i - y_j)^+$.

[6] Under the Contested Garment rule, if one person claims the entire garment and the other person claims half of it, they should receive 3/4 and 1/4 of the garment, respectively. This is because only half is in dispute: by claiming half of the garment, the second claimant concedes the other half to the first claimant. Because the disputed portion is divided equally, the second claimant receives $1/2 \times 1/2 = 1/4$ of the garment. The same logic appears in Cachon and Larivere (1999) under the name *linear allocation*.

both subsystems fail (with probability p_b), in which case $T = \max\{Y_1, Y_2\}$. Recall that Y_1 and Y_2 are independent and exponentially distributed with parameters μ_1 and μ_2, respectively. Therefore, $E[Y_1] = 1/\mu_1$, $E[Y_2] = 1/\mu_2$, and $E[\max\{Y_1, Y_2\}] = 1/\mu_1 + 1/\mu_2 - 1/(\mu_1 + \mu_2)$. It then follows that the expected outage duration is

$$\tau(\mu_1, \mu_2) = p_1 E[Y_1] + p_2 E[Y_2] + p_b E[\max\{Y_1, Y_2\}] = \frac{p_1 + p_b}{\mu_1} + \frac{p_2 + p_b}{\mu_2} - \frac{p_b}{\mu_1 + \mu_2}, \quad (2)$$

and therefore the expected total cost can be expressed as

$$\Psi_c(\mu_1, \mu_2) = \kappa_1(\mu_1 - \underline{\mu}_1) + \kappa_2(\mu_2 - \underline{\mu}_2) + \pi\left(\frac{p_1 + p_b}{\mu_1} + \frac{p_2 + p_b}{\mu_2} - \frac{p_b}{\mu_1 + \mu_2}\right). \quad (3)$$

The firm chooses μ_1 and μ_2 to minimize this function. Let (μ_1^C, μ_2^C) denote the optimal recovery capacities for the centralized system that solves this minimization problem.

Proposition 1. (a) $\Psi_c(\mu_1, \mu_2)$ *is convex*. (b) (μ_1^C, μ_2^C) *is given by the unique solution to the system of equations* $(p_i + p_b)/\mu_i^2 - p_b/(\mu_i + \mu_j)^2 = \kappa_i/\pi$, $i = 1, 2$, $i \neq j$.

All proofs are found in the appendix. Observe from (3) that the expected total system cost is inseparable in μ_1 and μ_2 because there is a positive probability of the subsystems simultaneously failing ($p_b > 0$). The optimal subsystem capacities are therefore linked, and they are in fact complementary; i.e., $\partial^2\Psi_c/\partial\mu_i\partial\mu_j < 0$. (Recall that $\Psi_c(\mu_1, \mu_2)$ is a cost, not a profit, so a negative cross partial indicates complementarity.) That is, an increase in subsystem j's capacity leads to an increase in subsystem i's capacity. To understand this, consider what happens when both subsystems fail. If the capacity level of subsystem j goes up, then this subsystem's contribution to the expected system outage duration diminishes, increasing the likelihood that the outage duration is driven by subsystem i. Therefore, the marginal benefit (i.e., reduction in outage duration) of increasing subsystem i's capacity is greater when subsystem j's capacity is higher.

4.2. Decentralized System

We now study the decentralized system in which different firms are responsible for subsystem capacity decisions. If subsystem j has a recovery capacity of μ_j, then firm i's expected total cost as a function of its recovery capacity μ_i is $\Psi_i(\mu_i \mid \mu_j) = \kappa_i(\mu_i - \underline{\mu}_i) + E[\Pi_i(Y_i \mid Y_j)]$, which can be written as

$$\Psi_i(\mu_i \mid \mu_j) = \kappa_i(\mu_i - \underline{\mu}_i) + p_i\pi E[Y_i] + p_b\pi E[g_i(Y_i \mid Y_j)\min\{Y_i, Y_j\}] + p_b\pi E[(Y_i - Y_j)^+] \quad (4)$$

using the expression in (1). Note that the expectations are functions of capacities μ_i and μ_j. Firms 1 and 2 competitively choose μ_1 and μ_2 to minimize their respective expected costs $\Psi_1(\mu_1 \mid \mu_2)$ and $\Psi_2(\mu_2 \mid \mu_1)$ in a simultaneous-move game. The next proposition establishes the existence and uniqueness of the Nash equilibrium of this capacity game.

Proposition 2. *A Nash equilibrium of the capacity game exists under an allocation rule that satisfies Assumption* 1. *Moreover, the equilibrium is unique if* $p_b \leq 2\sqrt{p_1 p_2}$.

Note that Assumption 2 is not needed for this result. The condition $p_b \leq 2\sqrt{p_1 p_2}$ (which implies $p_b \leq p_1 + p_2$ via geometric inequality) that establishes uniqueness of the equilibrium is quite reasonable, because it is satisfied in realistic scenarios where a joint subsystem failure is less likely to occur than a single subsystem failure.

The equilibrium capacities are found from the first-order conditions that minimize (4) for $i = 1, 2$. Although in general they are specified implicitly, simplified expressions are available for the two special allocation rules A1 and A2 introduced in §3.4. Let (μ_1^{A1}, μ_2^{A1}) and (μ_1^{A2}, μ_2^{A2}) be the equilibrium capacities under these rules.

Corollary 1. *The equilibrium capacities* (μ_1^{A1}, μ_2^{A1}) *and* (μ_1^{A2}, μ_2^{A2}) *are given by the solutions to the following systems of simultaneous equations*:

(a) *Under A*1,

$$\frac{p_i + p_b}{\mu_i^2} - \frac{p_b}{2(\mu_i + \mu_j)^2} = \frac{\kappa_i}{\pi}, \quad i = 1, 2;\ i \neq j.$$

(b) *Under A*2,

$$\frac{p_i + p_b}{\mu_i^2} - p_b\left(\frac{2\mu_j(2\mu_i + \mu_j)}{(\mu_i - \mu_j)^4}\ln\frac{(\mu_i + \mu_j)^2}{4\mu_i\mu_j} - \frac{2\mu_j(2\mu_i + \mu_j)}{(\mu_i^2 - \mu_j^2)^2}\right) = \frac{\kappa_i}{\pi} \quad \text{if } \mu_i \neq \mu_j \text{ and}$$

$$\left(p_i + \frac{13}{16}p_b\right)\frac{1}{\mu_i^2} = \frac{\kappa_i}{\pi} \quad \text{otherwise},$$

$i = 1, 2$; $i \neq j$.

If firms are symmetric, the Nash equilibrium of the capacity game under each allocation rule is unique and symmetric.

Notice that uniqueness under A1 and A2 is established without any restriction on the parameter values for the special case of symmetric firms, i.e., when all exogenous parameters are identical ($\theta_i = \theta_j$, $q_{ij} = q_{ji}$, and so on); the condition $p_b \leq 2\sqrt{p_1 p_2}$ in Proposition 2 is not required. In the discussions below, we focus on the symmetric case as it facilitates the analysis and helps generate sharp insights.

Before moving on to the comparison between the equilibrium capacities in a decentralized system and the optimal capacities in a centralized system, we examine the nature of capacity interactions under A1 and A2. Recall that the subsystem capacities were complementary in the centralized case. It is found that the same is true under A1, i.e., $\partial^2\Psi_i/\partial\mu_i\partial\mu_j < 0$, for the reason qualitatively similar to the one outlined in §4.1. Under A2, on the other hand, the capacity interaction is more nuanced. It can be shown that the capacities are complements if their values are close to each other, but they become substitutes if their values are sufficiently far apart. This happens because the degree of convexity embedded in the fraction $y_i/(y_i + y_j)$ creates highly nonlinear behavior at extreme ends. Therefore, whether the capacities are complements or substitutes in a decentralized system depends on how the systemwide penalty is allocated.

4.3. Comparing the Centralized and Decentralized Systems

We now compare the outcomes in the centralized and decentralized cases. We focus on three metrics of particular interest: recovery capacity, expected outage duration, and total system cost. Note that system availability increases as the expected outage duration decreases; see §3.3. To enable unambiguous comparisons and to generate sharp insights, in this subsection we focus on the symmetric system in which all subsystem-specific parameters are identical. (This allows us to normalize all confounding factors and isolate the effect of decentralization.) To this end, we drop the subscripts i and j from parameters κ, θ, and q for notational convenience. Because $\mu_1 = \mu_2$ in equilibrium for the symmetric system, we use μ^C and μ^A to denote the optimal/equilibrium subsystem recovery capacity of each firm in the centralized system and in the decentralized system under any allocation rule satisfying Assumptions 1 and 2. Similarly, μ^{A1} and μ^{A2} denote each firm's equilibrium capacity in the decentralized system under the specific allocation rules A1 and A2. The resulting expected outage duration τ and the total system cost Ψ_{tot} are superscripted in a similar manner, with the understanding that $\Psi_{\text{tot}} = \Psi_c$ in the centralized system and $\Psi_{\text{tot}} = \Psi_1 + \Psi_2$ in the decentralized system. The next proposition specifies the capacity decisions in equilibrium and compares the magnitudes of the resulting metrics.

Proposition 3. *In a symmetric system with* $p_b/2 \le p_1 = p_2$, $\mu^A > \mu^C$, $\tau^A < \tau^C$, *and* $\Psi^A_{\text{tot}} > \Psi^C_{\text{tot}}$. *Moreover, for any* p_b,

(a) $\mu^C = \sqrt{(1+\frac{1}{2}q)\rho}$, $\mu^{A1} = \sqrt{(1+\frac{3}{4}q)\rho}$, *and* $\mu^{A2} = \sqrt{(1+\frac{5}{8}q)\rho}$, *where* $\rho \equiv (\theta\lambda\pi)/\kappa$.

(b) $\mu^{A1} > \mu^{A2} > \mu^C$; $\tau^{A1} < \tau^{A2} < \tau^C$; $\Psi^{A1}_{\text{tot}} > \Psi^{A2}_{\text{tot}} > \Psi^C_{\text{tot}}$.

The expressions in part (a) of Proposition 3 show that the optimal/equilibrium subsystem recovery capacities increase in the shock arrival probability (λ), fragility (θ), and penalty rate (π), whereas they decrease in the capacity cost (κ). The capacities also increase in q; i.e., in equilibrium, a firm's capacity investment is larger when the subsystems are more susceptible to each other's direct failure. All of the results are in line with intuition.

The proposition ranks the three metrics in the centralized and decentralized settings evaluated at symmetric equilibria. It states that the firms in a decentralized system *overinvest* in recovery capacity compared with the centralized system. This is true under any allocation rule satisfying Assumptions 1 and 2,[7] and therefore it is true for the special cases A1 and A2. Capacity overinvestment in turn leads to lower expected outage duration and hence higher availability. However, this availability benefit comes at a higher overall system cost, creating inefficiency. The intuition behind this overinvestment result is not immediately obvious: given that each firm in a decentralized system internalizes only a portion of the system outage penalty through an allocation, why do the firms overinvest in recovery capacities rather than underinvest in them?

To understand this, it is instructive to consider a two-division equivalent of the centralized system. That is, suppose that the centralized system is composed of two subdivisions (division 1 and division 2), each managing its subsystem and receiving the same outage penalty allocation as the individual firms in decentralized systems do. Capacity decisions are made by a central manager to minimize the total expected system cost Ψ_{tot}, which is decomposed into subdivision costs Ψ_1 and Ψ_2. This decomposition allows for a fair comparison with a decentralized system, helping us isolate the source of inefficiency arising from decentralization.

Let us first consider A1, from which the key intuition is obtained. Under A1, division i in the centralized system and firm i in the decentralized system each expects to receive the penalty $(\pi/2)E[\min\{Y_i, Y_j\}] + \pi E[(Y_i - Y_j)^+]$ (see §3.4). Now suppose that subsystem i's capacity is increased by one unit while subsystem j's capacity is held constant. This leads to a reduction of both terms in the penalty function, thus benefitting both the centralized system and firm i in the decentralized system. Whereas this is the full impact on the decentralized firm, the

[7] This statement extends even to the cases where recovery times have general distributions that satisfy certain mild conditions. However, the exponential distribution assumption greatly simplifies the analysis, especially in proving the existence and uniqueness of the equilibrium (Proposition 2).

same is not true for the centralized firm because it also internalizes the penalty incurred by the other division (division j), equal to $(\pi/2)E[\min\{Y_i, Y_j\}] + \pi E[(Y_j - Y_i)^+]$. This penalty function increases with higher capacity of subsystem i (because the increase in the second term dominates the decrease in the first term), incurring a cost. Therefore the marginal benefit of increased capacity of subsystem i for division i is moderated by the concurrent marginal loss for division j, and this lowers the centralized firm's incentive to invest in capacity. In contrast, such a moderating effect is not present in the decentralized setting. Hence, overinvestment under A1 is driven by each individual firm's failure to account for the adverse effect of capacity increase—namely, that a shortened recovery time of its subsystem exposes the other subsystem to be the last one to be restored and causes the other firm to be penalized for it.

Proposition 3 proves that a similar effect exists under any allocation rule satisfying Assumptions 1 and 2. Therefore, overinvestment in recovery capacity in a decentralized system arises because individual firms fail to account for the negative externality that a capacity increase creates under these allocation rules. It is important to recognize that Assumptions 1 and 2 serve as sufficient conditions that lead to this outcome. Indeed, one may construct an allocation rule that results in capacity underinvestment if one of these assumptions is violated; see §7.1 for an example. However, given that the sufficient conditions reflect reasonable notions of firm accountability, our findings suggest that capacity overinvestment would be the norm rather than the exception.

Finally, observe from part (b) of Proposition 3 that the degree of capacity overinvestment is smaller under A2 than under A1. This is because a faster firm is penalized more heavily under A2: if firm i finishes recovery first, i.e., $y_i < y_j$, the fraction of the penalty $\pi \min\{y_i, y_j\} = \pi y_i$ it receives under A2 is $g_i(y_i \mid y_j) = y_j/(y_i + y_j) > 1/2$, whereas it is $g_i(y_i \mid y_j) = 1/2$ under A1. The higher penalty for being fast discourages quick recovery, thus dampening an incentive to overinvest.

5. Investments in Failure Prevention and Recovery Capacity

To this point, we have assumed that the firms could influence the subsystem recovery capacities μ_1 and μ_2 but not subsystem fragilities, i.e., the direct failure probabilities θ_1 and θ_2. Although this may be a good approximation of reality in some cases (especially when effective prevention technology does not exist or is economically infeasible to employ), in many instances firms are able to reduce fragility by deploying better failure-prevention capabilities in their subsystems. This may be achieved by, for example, material enhancement to withstand external shocks, redesign of internal subsystem architecture, and installation of shock-absorbing devices.

In this section, we extend our model to allow for failure prevention *and* recovery capacity investments. Anchoring on the findings of the previous section, we aim to understand how the former interacts with the latter. We describe the model in a decentralized setting, with the understanding that the centralized setting is analogous except that a single firm makes the decisions for both subsystems. Assume that firm i can exert an effort $a_i \in [0, \bar{\theta}_i)$ to reduce the failure probability θ_i from the default level $\bar{\theta}_i < 1$, which is thought of as the fragility inherent in the initial subsystem design. Assume $\bar{\theta}_1 + \bar{\theta}_2 < 1$. We choose a linear relationship between a_i and θ_i such that $\theta_i = \bar{\theta}_i - a_i$, but assume that the failure-prevention cost is convex increasing in the effort and approaches infinity as $a_i \to \bar{\theta}_i$, which implies that it becomes prohibitively expensive to make the subsystem completely resistant to the external shock. These features are captured in the cost function $\phi_i(1/(\bar{\theta}_i - a_i) - 1/\bar{\theta}_i)$, which we employ in the analysis. The constant $\phi_i > 0$ determines the marginal cost of investment, and its value is assumed to be sufficiently smaller than π, ensuring that the investment is justified and that the problem is well behaved. Before time 0, each firm chooses the failure-prevention effort a_i jointly with the recovery capacity μ_i. They are chosen to minimize the firm's expected cost, which now includes the failure-prevention cost. In the analysis we adopt the convention that the firms make sequential decisions on these variables: (a_1, a_2) followed by (μ_1, μ_2). Thus the game is played in two stages, and we seek a subgame-perfect equilibrium of this game.

To enable comparisons with the earlier results (Proposition 3), we consider a system consisting of identical subsystems and focus on the resulting symmetric equilibria. Unfortunately, the analysis of a decentralized system under the penalty allocation (1) with general penalty division function $g_i(y_i \mid y_j)$ is intractable even if symmetry is assumed. For this reason we focus on the special cases A1 and A2, which permit tractability as they feature simplified expressions for $g_i(y_i \mid y_j)$. It can be shown that, given parameter combinations that guarantee interior solutions, a symmetric equilibrium exists under A1. Existence of an equilibrium under A2 is still difficult to prove because of implicit characterization of equilibrium capacities (see Proposition 2(b)), but numerical examples show that a symmetric equilibrium exists. Based on these results, we characterize the symmetric equilibria in the next proposition, which parallels Proposition 3, in which we did not take into account the failure-prevention effort decisions. As before, we

drop the subscripts i and j and use θ^k and μ^k, $k \in \{C, A1, A2\}$, to denote the equilibrium fragilities and recovery capacities.

Proposition 4. *Assume a symmetric system and let* $\eta \equiv \phi^2/(\kappa\lambda\pi)$ *and* $\chi \equiv (\phi\lambda\pi)/\kappa^2$. *Then*

(a) $\theta^C = ((1+\frac{1}{2}q)^{-1}\eta)^{1/3}$, $\theta^{A1} = ((1-\frac{1}{4}q)^{-2}(1+\frac{1}{3}q)\eta)^{1/3}$, *and* $\theta^{A2} = ((1-\frac{1}{4}q)^{-2}(1+\frac{5}{8}q)\eta)^{1/3}$.

(b) $\mu^C = ((1+\frac{1}{2}q)\chi)^{1/3}$, $\mu^{A1} = ((1-\frac{1}{4}q)^{-1}(1+\frac{3}{4}q)^2\chi)^{1/3}$, *and* $\mu^{A2} = ((1-\frac{1}{4}q)^{-1}(1+\frac{5}{8}q)^2\chi)^{1/3}$.

(c) $\theta^{A1} > \theta^{A2} > \theta^C$; $\mu^{A1} > \mu^{A2} > \mu^C$; $\tau^C < \tau^{A1} < \tau^{A2}$; $\Psi^{A1}_{\text{tot}} > \Psi^{A2}_{\text{tot}} > \Psi^C_{\text{tot}}$.

We first examine how the equilibrium levels of the two decision variables, i.e., failure-prevention effort ($a = \bar{\theta} - \theta$, which linearly decreases in fragility θ) and recovery capacity (μ), vary with the exogenous parameters λ, ϕ, κ, and π. From parts (a) and (b) of Proposition 4 we find that, independent of system governance, (i) a higher probability of shock arrival (λ) leads to higher prevention effort and higher recovery capacity, (ii) a higher cost of failure prevention (ϕ) leads to lower prevention effort and higher recovery capacity, (iii) a higher cost of recovery capacity (κ) leads to higher prevention effort and lower recovery capacity, and (iv) a higher outage penalty rate (π) leads to higher prevention effort and higher recovery capacity. All results are in line with intuition. In addition, the results in (ii) and (iii) suggest that prevention effort and recovery capacity act as substitutes: the lower chance of a failure alleviates the need for recovery capacity. We discuss the impact of susceptibility q in §6.

Next, we compare the relative magnitudes of the two decision variables in equilibrium for the centralized system and the decentralized system under A1 and A2. Part (c) of Proposition 4 tells us that decentralization leads to higher overall cost, and this inefficiency is created as the firms in a decentralized system underinvest in failure prevention (leaving the subsystems more fragile) but overinvest in recovery capacity compared with the centralized system. In effect, decentralization leads the firms to shift their focus from *preventing failures* to *responding to failures*. The primary reason for underinvestment in failure prevention is that, regardless of how the postfailure outage penalty is divided between the firms (i.e., A1 or A2), the portion that each firm receives is always smaller than the total. Hence, the marginal benefit of reducing the chance of a failure is lower in a decentralized system. Note that the penalty splitting does not lead to overinvestment—unlike what happened with the recovery capacity decision—because failure prevention does not create a negative externality as the capacity increase does under A1 and A2. To the contrary, a lower chance of a direct subsystem failure is beneficial to the other firm exposed to the risk of indirect failure, and underinvestment arises because each firm in a decentralized system fails to internalize the full benefit.

As it turns out, this underinvestment problem is further exacerbated by the failure-prevention decision's interaction with the recovery capacity decision. This is because, as we observed above, these two decisions act as substitutes: a firm's tendency to overinvest in recovery capacity (as we found in the last section) leads to shorter system outage duration and hence reduces the outage penalty, which in turn lowers the firm's incentive to invest in failure prevention because the consequence of a failure is now smaller. Therefore, simultaneous decisions on failure prevention and recovery capacity reinforce each other to magnify the deviation due to decentralization. Interestingly, the flexibility of having two levers instead of one does not mitigate the inefficiency created by decentralization.

Another interesting question is how the combination of higher fragility and shorter recovery time together impact the expected system outage duration τ, which is the failure probability-weighted average of the conditional recovery times. Because the former increases τ whereas the latter decreases it, the answer to this question is not immediately clear. Part (c) of Proposition 4 establishes that the disadvantage of higher fragility dominates the advantage of shorter recovery time, resulting in a longer expected outage duration (hence lower system availability) in the decentralized system compared with the centralized system. Intuitively, this happens because, whereas a change in subsystem fragility directly impacts both subsystems (because it affects the chance of a joint failure), a direct impact of the change in a subsystem's recovery capacity is limited to that subsystem. Notice that this result reverses the analogous conclusion from Proposition 3(b), where it was assumed that the firms could invest only in recovery capacities; there, we found that decentralization leads to a shorter expected outage duration and thus higher system availability. Therefore, whether the system availability is higher or lower under decentralization depends critically on the firms' ability to influence the probability of a direct failure.

6. Impact of Susceptibility

Our analysis thus far has assumed that the susceptibilities q_{12} and q_{21}, i.e., the probabilities of indirect failures, are fixed. Susceptibility is perhaps the most important exogenous parameter in our model because all of the insights derived above rest on the assumption that a joint failure is possible. In fact, there is no distinction between centralized and decentralized systems in a cascade-free system where $q_{12} = q_{12} = 0$.

Table 1 **Directional Changes for the Decision Variables in Equilibrium as q Increases**

	Centralized	Decentralized (A1)	Decentralized (A2)
Recovery only (§4)	$\mu^C \uparrow$	$\mu^{A1} \uparrow$	$\mu^{A2} \uparrow$
Prevention and recovery (§5)	$\theta^C \downarrow$ $\mu^C \uparrow$	$\theta^{A1} \uparrow$ $\mu^{A1} \uparrow$	$\theta^{A2} \uparrow$ $\mu^{A2} \uparrow$

To better understand the extent to which the insights depend on susceptibility, in this section we study the impact of varying this parameter. We continue to focus on the symmetric equilibria in order to facilitate comparisons.

First, we return to the results presented in Propositions 3 and 4 and examine how the decision variables in equilibrium (recovery capacity μ and the subsystem fragility θ, when it is endogenously determined) vary with susceptibility q. For a decentralized system, we consider A1 and A2 only because analytical results are available for them. Note $q = q_{12} = q_{21}$ in these results, since it is assumed that the subsystems are identical. Thus, a higher q means both subsystems are more susceptible to each other's direct failure. The following observations are made: (i) the optimal/equilibrium recovery capacity increases in q in all cases, i.e., regardless of whether the system is centralized or decentralized and whether the subsystem fragilities are exogenously given or endogenously determined; and (ii) when the firm(s) can simultaneously invest in failure prevention and recovery capacity, the optimal/equilibrium subsystem fragility decreases in q in a centralized system but increases in q in a decentralized system. These observations are summarized in Table 1.

That higher susceptibility leads to a larger recovery capacity is intuitive. However, the result on fragility is less clear: why do firms in a decentralized system invest less in failure prevention (higher θ) when the chance of an indirect failure is greater (higher q), whereas the opposite is true in a centralized system?

To understand this, consider what happens if q_{ij}, the probability that a direct failure of subsystem i leads to an indirect failure of subsystem j, is increased while q_{ji} and θ are held constant at the symmetric equilibrium. It can be shown analytically that an incremental increase in q_{ij} leads to a larger expected cost for firm j but a *lower* cost for firm i in a decentralized system. That is, the firm exposed to an indirect failure is hurt by high susceptibility, but the firm that originates it actually benefits from high susceptibility. On the surface it seems that there is no reason that a firm should enjoy having the other firm undergo its own subsystem failure. However, this is precisely what happens. Recall that firm i expects to be charged with the outage penalty $\pi E[Y_i]$ if only its subsystem fails, whereas it expects $(\pi/2)E[\max\{Y_i, Y_j\}]$ if both subsystems fail—exactly half of the systemwide expected penalty—given that the subsystems are identical and the resulting equilibrium is symmetric. Observe that the expected penalty in the latter case is smaller than that of the former case because the subadditivity of the maximum function implies $E[\max\{Y_i, Y_j\}] < E[Y_i + Y_j] = 2E[Y_i]$. Because the firm splits the total outage penalty that is smaller than the sum of two individual penalties, a joint failure is then more attractive than a single failure. In essence, a firm prefers having the other firm participate in the failure event and the subsequent recovery effort because it presents an opportunity to diffuse the firm's responsibility for the system outage.

Now suppose that failure prevention is feasible, and consider a firm deciding how much to invest in it. If this firm faces a low probability that a failure of its subsystem leads to an indirect failure of the other subsystem, then the firm is likely to be the only one bearing the entire system outage penalty. On the other hand, if the firm faces a high probability of indirect failure, it is likely to share the outage penalty with the other firm. Between these two scenarios, the firm is more willing to let its subsystem fail in the latter case because penalty sharing results in a net cost saving. This explains why subsystem fragility increases with susceptibility in a decentralized system. In contrast, the opposite is true in a centralized system because the strategic incentive described above is absent.

Additionally, numerical examples confirm our earlier intuition that underinvestment in failure prevention and overinvestment in recovery capacity reinforce each other, magnifying the degree of deviation relative to when only recovery capacity investment is possible. They also reveal that degradation in system performance as a result of decentralization is quite sensitive to the risk of indirect failure; the system can lose up to a quarter of the efficient level of system availability. These results offer a new perspective on the advantage of modular system design. By designing a subsystem that does not spread failures to other subsystems, a firm can suppress the underinvestment and overinvestment incentives that we have identified, and as a result, a more efficient level of system availability can be attained.

7. Alternative Allocation Rules

The preceding results hold for allocation rules that satisfy Assumptions 1 and 2. These assumptions reflect the important notion of accountability, and as a result, the derived insights apply to realistic situations. There are, however, simple and noteworthy allocation rules that violate these assumptions. We now examine some such rules to shed further light on the impact of allocation.

7.1. Capacity Underinvestment Under Egalitarian Allocation Rule

One of the simplest allocation rules is the following.

Allocation Rule 3 (A3; The "Egalitarian" Rule). *In the case of a joint failure, the total penalty* $\pi \max\{y_i, y_j\}$ *is split equally between the two firms. Thus,* $\Pi_i(y_i \mid y_j) = (\pi/2)\max\{y_i, y_j\}$ *if* $y_i, y_j > 0$.

Under this rule, the total penalty $\pi \max\{y_i, y_j\}$ is divided in half regardless of the difference in the realized subsystem recovery times. This rule violates Assumption 2, and more importantly, Assumption 1(iii), i.e., the requirement that a firm is not held accountable beyond the duration of system outage to which it contributed. It is straightforward to show that firms in a decentralized system *underinvest* in recovery capacity under this rule, a reversal of the earlier results. (Underinvestment in failure prevention is maintained, however.) This happens because tying a faster firm's penalty to that of a slower firm dampens the former's incentive to recover its subsystem quickly. Although this example shows that capacity overinvestment is not a universal result, it highlights the fact that the reversal occurs in an unlikely scenario in which a firm agrees to pay a penalty for the outage duration for which it is not responsible.

7.2. Allocation Rules That Replicate Centralized Decisions

All of the allocation rules we have examined so far resulted in equilibria that differ from the optimal solution in the centralized system. This observation leads to a natural question: is there an allocation rule under which decentralization does not lead to a deviation? The answer is yes for the following two allocation rules, but they have limitations, as we explain below.

Allocation Rule 4 (A4; The "Blame the Winner" Rule). *The first-to-recover firm fully absorbs the penalty rate* π *for the duration in which both subsystems are down but nothing thereafter, whereas the last-to-recover firm fully absorbs* π *for the duration in which only its subsystem is down but nothing beforehand. In case of a tie, the penalty is split equally. Thus,* $\Pi_i(y_i \mid y_j) = \pi[\mathbf{1}(y_i < y_j) + (1/2)\mathbf{1}(y_i = y_j)]\min\{y_i, y_j\} + \pi(y_i - y_j)^+$.

Allocation Rule 5 (A5; The "Blame No One" Rule). *The first-to-recover firm is charged with no penalty, whereas the last-to-recover firm fully absorbs the penalty rate* π *for the duration in which only its subsystem is down but nothing beforehand. In case of a tie, the penalty is split equally. Thus,* $\Pi_i(y_i \mid y_j) = (\pi/2)\mathbf{1}(y_i = y_j)\min\{y_i, y_j\} + \pi(y_i - y_j)^+$.

Note that $\mathbf{1}(\cdot)$ is the indicator variable. Unlike A1, under which firms equally split the outage penalty $\pi \min\{y_i, y_j\}$ for which both are responsible, under A4 the same penalty is borne entirely by a faster firm. On the other hand, neither firm bears this penalty under A5; firm i is penalized only for the excess delay, $(y_i - y_j)^+$. Both allocation rules have issues because neither fully satisfies Assumption 1, the requirements for accountability. A4 violates Assumption 1(iv) because firm i's penalty decreases in y_i at $y_i = y_j$; i.e., the firm may actually be rewarded for being late. This rule is also not intuitive, because a slower firm is completely free of paying for its culpability toward the joint downtime $\min\{y_i, y_j\}$ despite its share of contribution. A5 is an incomplete allocation rule because the penalty $\pi \min\{y_i, y_j\}$ is left unassigned. As such, it violates Assumption 1(i) because $\Pi_1(y_1 \mid y_2) + \Pi_2(y_2 \mid y_1) < \Pi(\max\{y_1, y_2\})$; this means that the two firms are not held fully responsible for the total system outage under A5. Despite these shortcomings, they may lead to desirable outcomes from the system perspective.

Proposition 5. *Assume a symmetric system. When firms invest in recovery capacity only,* $\mu^{A4} = \mu^{A5} = \mu^C$. *When firms invest in failure prevention and recovery capacity,* $\theta^C < \theta^{A4} < \theta^{A5}$ *and* $\mu^C < \mu^{A4} < \mu^{A5}$.

As Proposition 5 reveals, the centralized capacity decision is replicated under A4 and A5 when subsystem fragilities are exogenously given and the decentralized firms make independent decisions on recovery capacity. That A4 lowers the incentive to overinvest in capacity is expected because it penalizes a faster firm disproportionately. However, it is surprising that overinvestment is completely removed. As it turns out, this is a by-product of assuming a symmetric system and exponentially distributed recovery times. If either of these assumptions is relaxed, the equilibrium deviates from the centralized solution, although the degree of deviation is in general smaller than that of other allocation rules studied in the previous sections. That the equilibrium under A5 coincides with the system optimum is also surprising, and even more so because this is true in any circumstance; the result does not depend on symmetry or the underlying distribution.[8] Intuitively, coordination occurs because the system-level performance (i.e., availability) is driven by the last-to-recover subsystem; it is the excess delay $(y_i - y_j)^+$ that ultimately determines how fast the system is restored.

Although potential elimination of efficiency loss under these allocation rules is intriguing, whether they can be implemented in practice is questionable because of the shortcomings identified above. Moreover, system optimality is no longer attained if the

[8] Observe that $(\partial/\partial\mu_i)E[\max\{Y_i, Y_j\}] = (\partial/\partial\mu_i)E[(Y_i - Y_j)^+]$, which follows from the relation $\max\{y_i, y_j\} = y_j + (y_i - y_j)^+$. Because the centralized firm chooses capacity μ_i to minimize the expected penalty $\pi E[\max\{Y_i, Y_j\}]$, whereas firm i in a decentralized system under A5 does so to minimize $\pi E[(Y_i - Y_j)^+]$, the optimality/equilibrium conditions in the two cases are identical.

firms invest in failure prevention as well as recovery capacity, according to Proposition 5. In this case we arrive at the same conclusion as before: there is an underinvestment in failure prevention and overinvestment in capacity. The net result is lower system availability. In summary, although there exist allocation rules that coordinate the system, they are limited in their scope and applicability.

7.3. Compensation After Root Cause Identification

Thus far we have focused on the allocation rules that reflect accountability and impartiality, under the premise that it is difficult or impractical to identify the exact chain of events that led to a cascading failure. In case it is identified, however, the firm whose subsystem originated the cascade may have a legal obligation to partially compensate the other firm for its role in allowing the shock transmission by not preventing the shock from entering into the system. We study such a scenario here, considering two simple partial compensation schemes. We focus on the recovery capacity investment case of §4.

First, suppose that compensation is based on the total penalty incurred by the receiving firm (i.e., $\Pi_j(y_j \mid y_i)$ as defined in (1), if it is subsystem j that undergoes indirect failure). We adopt the following sequence of events. Firms initially pay penalties according to an allocation rule satisfying Assumptions 1 and 2, and then, upon discovering the root cause of the joint failure (which might take a long time after the penalties are assessed), the firm whose subsystem is found to have originated the failure reimburses a fraction of the other firm's penalty that is subject to compensation. Let γ_{ij} be the fraction of firm j's total penalty $\Pi_j(y_j \mid y_i)$ reimbursed by firm i in the event that a joint failure is found to have originated from subsystem i, which happens with probability $q_{ij}\theta_i\lambda$. For simplicity, suppose that the firms are symmetric with $\gamma = \gamma_{ij} = \gamma_{ji}$. It can then be shown that the parameter γ plays the role of a weight between the no-compensation case of §4 and A3 (the "Egalitarian" rule) discussed in §7.1. (If $\gamma = 1$, a firm found to have originated the joint failure—an event that occurs with probability $p_b/2$ given the symmetry—is fully responsible for the systemwide penalty; the expected penalty allocation in this case is the same as that under A3.) Therefore, allowing for partial compensation results in a convex combination of these two cases. Given that decentralization results in overinvestment in recovery capacity in the former case whereas the opposite happens in the latter case, we then see that either over- or underinvestment in capacity may arise depending on the degree of partial compensation.

It is important to note that this compensation scheme violates (ii) and (iii) of Assumption 1, two of the requirements for firm accountability. Violation of (ii), i.e., nonimpartiality, is expected because ex post compensation means the originating firm has a heavier burden of penalty than the receiving firm does. However, violation of (iii) is more difficult to justify because it means that the compensating firm pays for the delay in system outage over which it has no control. Even if the firm has to bear some responsibility for the indirect failure cascaded to the other firm's subsystem, it is unreasonable to fault the former for not expediting recovery of the latter's subsystem. Therefore, a more reasonable compensation scheme would ensure that the maximum amount of compensation does not exceed $\Pi(y_i) = \pi y_i$ for $i = 1, 2$, thereby retaining Assumption 1(iii). A similar analysis as above shows that partial compensation under this scheme results in a convex combination of the no-compensation case of §4 and A4 (the "Blame the Winner" rule) of §7.2. This new scheme does not alter the main conclusion from the previous sections—that decentralization leads to overinvestment in recovery capacity.

8. Robustness of Results

8.1. General Investment Costs

The results presented in §§4–6 are derived based on specific functions representing the costs of investments. That is, the cost of installing recovery capacity μ_i was assumed to be linear as $\kappa_i(\mu_i - \underline{\mu}_i)$, and the cost of exerting effort a_i to prevent subsystem failure was assumed to be $\phi_i(1/(\bar{\theta}_i - a_i) - 1/\bar{\theta}_i)$. They were specifically chosen because they capture the expected behaviors of the cost functions while at the same time permit closed-form solutions. In this subsection we study the impact of generalizing these cost functions. To this end, assume that the general cost of capacity investment $K_i(\mu_i)$ is convex increasing with $K_i(\underline{\mu}_i) = 0$ and that the general cost of failure-prevention effort $\Phi_i(a_i)$ is also convex increasing with $\Phi_i(0) = 0$ and $\lim_{a_i \to \bar{\theta}_i} \Phi_i(a_i) = \infty$. Just as in §5, the effort variable a_i is defined in a linear scale with respect to fragility θ_i such that $\theta_i = \bar{\theta}_i - a_i$. Convexity is a natural assumption since a firm would choose among many technology options in the order of their cost effectiveness.

It is analytically proved that all of the results in §4 extend to their generalized versions with the capacity cost $K_i(\mu_i)$: existence and uniqueness of the capacity game are established as in Proposition 2, and the equilibrium comparisons remain the same as in Proposition 3. In particular, capacity overinvestment in a decentralized system under Assumptions 1 and 2 remains true. The only difference is that closed-form expressions for the equilibrium solutions are no longer available. Similarly, the qualitative conclusions

from §§5 and 6 are maintained in the generalized version as well. This is verified numerically under certain restrictions on the effort cost function $\Phi_i(a_i)$ which ensure that the problem is well behaved. Therefore, the main insights are robust to generalization of the cost functions.

8.2. Time-Dependent Outage Penalty Rate

We have assumed that a constant penalty π is incurred by the system at each instant it is down. Although this is a reasonable assumption in many cases—especially when the penalty represents the lost economic output normally generated at a constant rate—in some instances it may be more appropriate to assume that the penalty rate increases over time because of snowballing opportunity costs and other long-term effects. To study the impact of such effects, we replace the constant penalty rate π with $\pi t^{\beta-1}$, where $\beta \geq 1$. Thus, the cumulative penalty is convex increasing in time.

For the purpose of illustration, we focus on A1, which permits tractable analytical expressions even with the time-dependent penalty rate specified above. Under A1, the penalty assigned to firm i in a decentralized system is $\Pi_i(y_i \mid y_j) = (1/2) \cdot \int_0^{\min\{y_i, y_j\}} (\pi t^{\beta-1})\, dt + \int_{\min\{y_i, y_j\}}^{y_i} (\pi t^{\beta-1})\, dt = (\pi/\beta) y_i^{\beta} - (\pi/(2\beta))(\min\{y_i, y_j\})^{\beta}$. Combining this expression with exponentially distributed recovery times, we can solve for the symmetric equilibrium capacity under A1. Comparing it with the optimal capacity in the centralized system, we find $\mu^{A1} > \mu^{C}$. Hence, the conclusion from §4 that firms overinvest in recovery capacity in a decentralized system remains unchanged. In addition, a similar analysis shows that the firms under A1 in a decentralized system underinvest in failure prevention, consistent with the conclusion from §5. That decentralization may lead to higher or lower system availability depending on the scenarios continues to be true as well. These observations provide evidence that the main insights are robust to generalization of the outage penalty rate.

9. Conclusions

In this paper we study the dynamics that arise in an environment where multiple firms managing distinct subsystems of a serially linked system decide to invest in resources to enhance system availability. Availability can be improved by two measures: investment in failure-prevention technology and investment in capacity used to restore a failed subsystem. Given the costs and benefits of taking these measures, the firms decide how much they should invest in each. The system experiences a failure if one or more subsystems fail, and it incurs a systemwide penalty that accrues for the duration in which at least one subsystem is down. A complicating factor is that subsystem failures are not independent; that is, a possibility exists that a failure of one subsystem caused by an external shock cascades to an indirect failure of another subsystem. In a decentralized system, this dependency creates skewed incentives for the firms investing in availability-enhancing measures. We analyze the consequences of such incentives and compare them with those of the centralized setting, identifying the drivers of efficiency loss.

A simultaneous failure of multiple subsystems creates an ambiguity in assigning responsibility for system outage. Because it is not possible to disentangle each firm's contribution to the outage, there exists no clear-cut way to allocate the system outage penalty to the firms. We identify the minimal set of conditions for penalty allocation rules that represent firm accountability. We then consider two scenarios: (1) firms may invest only in recovery capacity, and (2) firms may invest in both failure prevention and recovery capacity. In the first case, we find that the firms in a decentralized setting tend to overinvest in capacity. This happens because, under the allocation rules satisfying the aforementioned conditions, a firm's capacity increase creates a negative externality to the other firm. Because the firms do not take this externality into account, they invest in capacity disproportionately more than a centralized firm would. This then results in higher system availability but at a higher overall cost. In the second case, the firms continue to overinvest in recovery capacity, but at the same time, they underinvest in failure prevention, because each firm internalizes only a portion of the total outage penalty. Therefore, firms shift their focus from preventing failures to responding to failures. The combined effect is that system availability is lower in a decentralized setting, reversing the result of the first case. We also explore the role played by failure susceptibility, i.e., the probability that a failure of one subsystem cascades to another. We find that, in decentralized settings, firms facing high susceptibility have smaller incentives to invest in failure prevention than firms facing low susceptibility do. In other words, a firm is more willing to let its subsystem fail when it is more likely that the failure will spread to the other firm's subsystem. Such a perverse incentive arises because a joint failure helps a firm diffuse the responsibility for system outage by involving the other firm.

The analysis of our model sheds light on the issues that have been largely ignored in the prior literature—namely, the incentive dynamics created by simultaneous subsystem failures that contribute to system outages. There are many practical examples that our model applies to, and we have identified some of the most important features and insights that may play

significant roles in managerial decisions. Our model is not without limitations, because certain complicating aspects are not captured as a result of the level of parsimony chosen so as to highlight the most salient features. For example, the model does not explicitly represent the process by which firms agree on the system outage penalty allocation rule. In reality, an agreement can be made through negotiations or by a third party's intervention (e.g., system owner). In addition, many complex systems consist of more than two serial- and/or parallel-linked subsystems managed by different firms, creating further ambiguity in outage penalty allocations and complicating the incentive structure. Although incorporating these real-world complexities is beyond the scope of this paper, our analysis paves the ground for future work to address these issues.

Acknowledgments

The authors thank the seminar participants at Washington University in St. Louis, the University of British Columbia, the University of Pennsylvania, the University of Michigan, and Yale University. The authors are also grateful for the suggestions made by Edward Kaplan, the anonymous reviewers, associate editor, and department editor.

Appendix. Proofs

Proof of Proposition 1. Differentiating $\Psi_c(\mu_i, \mu_j)$ yields $\partial\Psi_c/\partial\mu_i = \kappa_i - \pi((p_i + p_b)/\mu_i^2 - p_b/(\mu_i + \mu_j)^2)$, $\partial^2\Psi_c/\partial\mu_i^2 = \pi((2p_i)/\mu_i^3) + 2p_b\pi(1/\mu_i^3 - 1/(\mu_i+\mu_j)^3)$, and $\partial^2\Psi_c/\partial\mu_i\partial\mu_j = -\pi((2p_b)/(\mu_i+\mu_j)^3)$. Then the determinant of the Hessian $H = \{\partial^2\Psi_c/\partial\mu_i\partial\mu_j\}$ is, after collecting terms,

$$|H| = 4\pi^2\big[(p_i+p_b)(p_j+p_b)(3\mu_i^2\mu_j + 3\mu_i\mu_j^2) + p_j(p_i+p_b)\mu_j^3 + p_i(p_j+p_b)\mu_i^3)\big]\cdot\big[\mu_i^3\mu_j^3(\mu_i+\mu_j)^3\big]^{-1} > 0.$$

Since $\partial^2\Psi_c/\partial\mu_i^2 > 0$ for $i = 1, 2$ and $|H| > 0$, H is positive definite, and thus $\Psi_c(\mu_i, \mu_j)$ is convex. The optimality conditions that specify the interior solution are obtained by setting $\partial\Psi_c/\partial\mu_i = 0$ for $i = 1, 2$. □

Proof of Proposition 2. Note that $E[(Y_i - Y_j)^+] = (\mu_j/\mu_i)/(\mu_i + \mu_j)$ for exponentially distributed Y_i and Y_j. Then firm i's expected total cost (4) is equal to

$$\Psi_i(\mu_i \mid \mu_j) = \kappa_i(\mu_i - \underline{\mu}_i) + p_i\pi(1/\mu_i) + p_b\pi((\mu_j/\mu_i)/(\mu_i+\mu_j)) + p_b\pi\psi_i(\mu_i \mid \mu_j),$$

where

$$\psi_i(\mu_i \mid \mu_j) \equiv E[g(Y_i \mid Y_j)\min\{Y_i, Y_j\}] = \int_0^\infty\int_0^\infty g_i(y_i \mid y_j)\min\{y_i, y_j\}\mu_i e^{-\mu_i y_i}\mu_j e^{-\mu_j y_j}\,dy_i\,dy_j.$$

Differentiating $\psi_i(\mu_i \mid \mu_j)$ yields

$$\frac{\partial^2\psi_i}{\partial\mu_i^2} = \int_0^\infty\int_0^\infty g_i(y_i \mid y_j)\min\{y_i, y_j\}\cdot(-2y_i + \mu_i y_i^2)e^{-\mu_i y_i}\mu_j e^{-\mu_j y_j}\,dy_i\,dy_j$$

and

$$\frac{\partial^2\psi_i}{\partial\mu_i\partial\mu_j} = \int_0^\infty\int_0^\infty g_i(y_i \mid y_j)\min\{y_i, y_j\}\cdot(1-\mu_i y_i)e^{-\mu_i y_i}(1-\mu_j y_j)e^{-\mu_j y_j}\,dy_i\,dy_j.$$

Integrating $\partial^2\psi_i/\partial\mu_i^2$ by parts,

$$\begin{aligned}\frac{\partial^2\psi_i}{\partial\mu_i^2} &= \int_0^\infty\int_0^{y_j}\left(\frac{\partial g_i}{\partial y_i}y_i + g_i\right)y_i^2 e^{-\mu_i y_i}\mu_j e^{-\mu_j y_j}\,dy_i\,dy_j \\ &\quad + \int_0^\infty\int_{y_j}^\infty\left(\frac{\partial g_i}{\partial y_i}y_j\right)y_i^2 e^{-\mu_i y_i}\mu_j e^{-\mu_j y_j}\,dy_i\,dy_j \\ &> -\int_0^\infty\int_{y_j}^\infty y_i^2 e^{-\mu_i y_i}\mu_j e^{-\mu_j y_j}\,dy_i\,dy_j \\ &= -2\left(\frac{1}{\mu_i^3} - \frac{1}{(\mu_i+\mu_j)^3}\right),\end{aligned}$$

where the inequality follows from Assumption 1(iv), i.e., $\partial\Pi_i(y_i \mid y_j)/\partial y_i > 0$, which can be written as $(\partial g_i/\partial y_i)y_i + g_i > 0$ for $y_i \le y_j$ and $(\partial g_i/\partial y_i)y_j + 1 > 0$ for $y_i > y_j$. Since

$$\frac{\partial^2}{\partial\mu_i^2}\left(\frac{\mu_j/\mu_i}{\mu_i+\mu_j}\right) = 2\left(\frac{1}{\mu_i^3} - \frac{1}{(\mu_i+\mu_j)^3}\right) > -\frac{\partial^2\psi_i}{\partial\mu_i^2},$$

we have

$$\frac{\partial^2\Psi_i}{\partial\mu_i^2} = \frac{2p_i\pi}{\mu_i^3} + p_b\pi\frac{\partial^2}{\partial\mu_i^2}\left(\frac{\mu_j/\mu_i}{\mu_i+\mu_j}\right) + p_b\pi\frac{\partial^2\psi_i}{\partial\mu_i^2} > \frac{2p_i\pi}{\mu_i^3} > 0.$$

Thus we have established that the expected cost $\Psi_i(\mu_i \mid \mu_j)$ is convex in μ_i, or equivalently, firm i's payoff $-\Psi_i(\mu_i \mid \mu_j)$ is concave. Then a Nash equilibrium exists (Theorem 1 in Cachon and Netessine 2004). Next, we prove that if $p_b \le 2\sqrt{p_1 p_2}$, then the Hessian $H = \{\partial^2\Psi_i/\partial\mu_i\partial\mu_j\}$ is positive quasidefinite (i.e., $H + H^T$ is positive definite), which ensures that the equilibrium is unique (Theorem 6 in Cachon and Netessine 2004). With $\partial^2\Psi_i/\partial\mu_i^2 > 0$ for $i = 1, 2$, it remains to show

$$2\sqrt{\frac{\partial^2\Psi_1}{\partial\mu_1^2}\frac{\partial^2\Psi_2}{\partial\mu_2^2}} \ge \left|\frac{\partial^2\Psi_1}{\partial\mu_1\partial\mu_2} + \frac{\partial^2\Psi_2}{\partial\mu_2\partial\mu_1}\right|.$$

Using the condition $g_1(y_1 \mid y_2) + g_2(y_2 \mid y_1) = 1$ that follows from Assumption 1(i), we find

$$\frac{\partial^2\psi_1}{\partial\mu_1\partial\mu_2} + \frac{\partial^2\psi_2}{\partial\mu_2\partial\mu_1} = \int_0^\infty\int_0^\infty \min\{y_1, y_2\}(1-\mu_1 y_1)(1-\mu_2 y_2)\cdot e^{-\mu_1 y_1}e^{-\mu_2 y_2}\,dy_1\,dy_2 = \frac{2}{(\mu_1+\mu_2)^3}.$$

Note that

$$\frac{\partial^2}{\partial\mu_i\partial\mu_j}\left(\frac{\mu_j/\mu_i}{\mu_i+\mu_j}\right) = -\frac{2}{(\mu_i+\mu_j)^3}.$$

Since

$$\frac{\partial^2\Psi_i}{\partial\mu_i\partial\mu_j} = p_b\pi\frac{\partial^2}{\partial\mu_i\partial\mu_j}\left(\frac{\mu_j/\mu_i}{\mu_i+\mu_j}\right) + p_b\pi\frac{\partial^2\psi_i}{\partial\mu_i\partial\mu_j},$$

we have

$$\frac{\partial^2\Psi_1}{\partial\mu_1\partial\mu_2} + \frac{\partial^2\Psi_2}{\partial\mu_2\partial\mu_1} = -\frac{4p_b\pi}{(\mu_1+\mu_2)^3} + \frac{2p_b\pi}{(\mu_1+\mu_2)^3} = -\frac{2p_b\pi}{(\mu_1+\mu_2)^3}.$$

Combining this equality with $\partial^2\Psi_i/\partial\mu_i^2 > (2p_i\pi)/\mu_i^3$ found above,

$$2\cdot\sqrt{\frac{\partial^2\Psi_1}{\partial\mu_1^2}\frac{\partial^2\Psi_2}{\partial\mu_2^2}} > 4\pi\sqrt{\frac{p_1p_2}{\mu_1^3\mu_2^3}} > 4\pi\sqrt{\frac{p_1p_2}{(\mu_1+\mu_2)^3(\mu_1+\mu_2)^3}}$$
$$\geq \frac{2p_b\pi}{(\mu_1+\mu_2)^3} = \left|\frac{\partial^2\Psi_1}{\partial\mu_1\partial\mu_2}+\frac{\partial^2\Psi_2}{\partial\mu_2\partial\mu_1}\right|,$$

where we used the assumption $\sqrt{p_1p_2}\geq p_b/2$. Hence, the Hessian is quasidefinite, and as a result, the equilibrium is unique if $p_b\leq 2\sqrt{p_1p_2}$. □

Proof of Corollary 1. Omitted (available upon request). □

Proof of Proposition 3. A symmetric equilibrium exists because the subsystem parameters are identical (Cachon and Netessine 2004, Theorem 2), and it is unique under the condition $p_b/2 < p_1 = p_2$ by Proposition 2. We first show $\mu^A > \mu^C$ under an allocation rule that satisfies Assumptions 1 and 2. Rewriting (4) using $E[(Y_i - Y_j)^+] = (\mu_j/\mu_i)/(\mu_i+\mu_j)$ for exponential distributions,

$$\Psi_i(\mu_i\mid\mu_j) = \kappa_i(\mu_i-\underline{\mu}_i)+p_i\pi\frac{1}{\mu_i}+p_b\pi\frac{\mu_j/\mu_i}{\mu_i+\mu_j}$$
$$+p_b\pi\psi_i(\mu_i\mid\mu_j),$$

where $\psi_i(\mu_i\mid\mu_j)\equiv E[g(Y_i\mid Y_j)\min\{Y_i,Y_j\}]$. Differentiating $\Psi_i(\mu_i\mid\mu_j)$ with respect to μ_i yields

$$\frac{\partial\Psi_i}{\partial\mu_i} = \kappa_i - p_i\pi\frac{1}{\mu_i^2} - p_b\pi\left(\frac{1}{\mu_i^2}-\frac{1}{(\mu_i+\mu_j)^2}\right)+p_b\pi\frac{\partial\psi_i}{\partial\mu_i}.$$

Similarly, differentiating $\Psi_c(\mu_i,\mu_j)$ for the centralized system using (3) yields

$$\frac{\partial\Psi_c}{\partial\mu_i} = \kappa_i - \pi\left(\frac{p_i+p_b}{\mu_i^2}-\frac{p_b}{(\mu_i+\mu_j)^2}\right).$$

The symmetric equilibrium/optimal capacities μ^A and μ^C therefore satisfy the first-order conditions

$$\kappa_i - \frac{(4p_i+3p_b)\pi}{4(\mu^A)^2}+p_b\pi\left.\frac{\partial\psi_i}{\partial\mu_i}\right|_{\mu_i=\mu_j}=0 \quad\text{and}$$
$$\kappa_i-\frac{(4p_i+3p_b)\pi}{4(\mu^C)^2}=0,$$

which are obtained by setting $\mu_i=\mu_j$ in $\partial\Psi_i/\partial\mu_i=0$ and $\partial\Psi_c/\partial\mu_i=0$. It can be proved that $(\partial\psi_i/\partial\mu_i)|_{\mu_i=\mu_j}<0$ under Assumptions 1 and 2 (proof is available upon request). Hence,

$$\kappa_i-\frac{(4p_i+3p_b)\pi}{4(\mu^A)^2}>0=\kappa_i-\frac{(4p_i+3p_b)\pi}{4(\mu^C)^2},$$

and therefore $\mu^A>\mu^C$. The expected system downtime τ at a symmetric equilibrium capacity $\mu=\mu_1=\mu_2$ is $\tau=(4p_1+3p_b)/(2\mu)$ (from (2)), which is decreasing in μ. Since $\mu^A>\mu^C$, then, $\tau^A<\tau^C$. The expected system cost Ψ_{tot} at a symmetric equilibrium is $\Psi_{\text{tot}}=2\kappa(\mu-\underline{\mu})+\pi\tau$ (from (3)), which is minimized at μ^C. Hence, $\Psi_{\text{tot}}^A>\Psi_{\text{tot}}^C$. The expressions for μ^C, μ^{A1}, and μ^{A2} in part (a) are obtained by substituting the symmetric solution $\mu=\mu_1=\mu_2$ in the first-order conditions given in Proposition 1 and Corollary 1. The inequalities in part (b) are verified by substituting the expressions in (a) in τ and Ψ_{tot} for each case and comparing them. □

Proof of Proposition 4. Consider the failure-prevention effort decision problem described in §5 for a symmetric system and assume $\pi>(20\phi^2)/(\kappa\bar{\theta}^3\lambda)$. Then it can be proved that (a) in the centralized system, the firm's objective function is minimized at a unique symmetric interior point; and (b) in the decentralized system under A1, the effort game is supermodular, and a symmetric equilibrium exists at an interior point. (Proof is available upon request.) In the following, we characterize the equilibrium in each case. For given values of θ_i and θ_j, the firm in a symmetric centralized system chooses μ_i^* and μ_j^* that satisfies the optimality conditions in Proposition 1. Substituting these values in the expected cost function

$$\Psi_c=\phi\left(\frac{1}{\bar{\theta}-a_i}-\frac{1}{\bar{\theta}}\right)+\phi\left(\frac{1}{\bar{\theta}-a_j}-\frac{1}{\bar{\theta}}\right)+\kappa(\mu_i-\underline{\mu})$$
$$+\kappa(\mu_j-\underline{\mu})+\pi\lambda\left(\frac{\theta_i+q\theta_j}{\mu_i}+\frac{q\theta_i+\theta_j}{\mu_j}-\frac{(\theta_i+\theta_j)q}{\mu_i+\mu_j}\right)$$

and differentiating it with respect to $a_i=\bar{\theta}-\theta_i$ and $a_j=\bar{\theta}-\theta_j$ using the envelope theorem yields the first-order conditions

$$\frac{\phi}{\theta_i^2}-\pi\lambda\left(\frac{1}{\mu_i^*}+\frac{q}{\mu_j^*}-\frac{q}{\mu_i^*+\mu_j^*}\right)=0 \quad\text{and}$$
$$\frac{\phi}{\theta_j^2}-\pi\lambda\left(\frac{q}{\mu_i^*}+\frac{1}{\mu_j^*}-\frac{q}{\mu_i^*+\mu_j^*}\right)=0.$$

Evaluating these at the symmetric solution $\theta^*=\theta_i^*=\theta_j^*$, which leads to $\mu^*=\mu_i^*=\mu_j^*$, we get

$$\theta^*=\sqrt{\frac{\phi\mu^*}{\pi\lambda(1+q/2)}}.$$

The optimal values of θ and μ are obtained by substituting this expression in the optimality condition for μ^* (see Proposition 1) and evaluating it at the symmetric solution:

$$\theta^C=\left(\left(1+\frac{q}{2}\right)^{-1}\frac{\phi^2}{\kappa\lambda\pi}\right)^{1/3} \quad\text{and}\quad \mu^C=\left(\left(1+\frac{q}{2}\right)\frac{\phi\lambda\pi}{\kappa^2}\right)^{1/3}.$$

The expected system outage duration and the expected system cost at a symmetric equilibrium are $\tau=2(1+q/2)\cdot((\theta\lambda)/\mu)$ (derived from (2)) and $\Psi_{\text{tot}}=-2\phi-2\kappa\underline{\mu}+2\phi/\theta+2\kappa\mu+\pi\tau$, respectively. At μ^C and θ^C found above, these quantities are

$$\tau^C=2\left(\left(1+\frac{q}{2}\right)\frac{\phi\kappa\lambda}{\pi^2}\right)^{1/3} \quad\text{and}$$
$$\Psi_{\text{tot}}^C=-2\phi-2\kappa\underline{\mu}+6\left(\left(1+\frac{q}{2}\right)\phi\kappa\lambda\pi\right)^{1/3}.$$

Next, consider a symmetric decentralized system under A1. Firm i's expected cost in the first stage of the game is

$$\Psi_i=\phi\left(\frac{1}{\bar{\theta}-a_i}-\frac{1}{\bar{\theta}}\right)+\kappa(\mu_i-\underline{\mu})$$
$$+\pi\lambda\left(\frac{(1-q)\theta_i}{\mu_i^*}+(\theta_i+\theta_j)q\left(\frac{1}{2(\mu_i^*+\mu_j^*)}+\frac{\mu_j^*/\mu_i^*}{\mu_i^*+\mu_j^*}\right)\right),$$

where μ_i^* and μ_j^* are the second-stage equilibrium capacities as specified in Proposition 2 for fixed θ_i and θ_j. They satisfy the conditions

$$\frac{\kappa}{\pi\lambda} = \frac{\theta_i + q\theta_j}{(\mu_i^*)^2} - \frac{(\theta_i+\theta_j)q}{2(\mu_i^*+\mu_j^*)^2} \quad \text{and}$$

$$\frac{\kappa}{\pi\lambda} = \frac{q\theta_i + \theta_j}{(\mu_j^*)^2} - \frac{(\theta_i+\theta_j)q}{2(\mu_i^*+\mu_j^*)^2}.$$

Differentiating Ψ_i with respect to $a_i = \bar{\theta} - \theta_i$ using the envelope theorem yields the first-order condition

$$\frac{\phi}{\theta_i^2} - \pi\lambda\left((1-q)\frac{1}{\mu_i^*} + q\left(\frac{1}{2(\mu_i^*+\mu_j^*)} + \frac{\mu_j^*/\mu_i^*}{\mu_i^*+\mu_j^*}\right)\right) = 0.$$

A similar condition is derived for firm j, with the indexes i and j interchanged. Evaluating them at the symmetric solution $\theta^* = \theta_i^* = \theta_j^*$ with $\mu^* = \mu_i^* = \mu_j^*$, we get $\theta^* = \sqrt{(\phi\mu^*)/(\pi\lambda(1-q/4))}$. The rest of the analysis proceeds similarly and is omitted. □

Proof of Proposition 5. Under A4, $g_i(Y_i \mid Y_j) = \mathbf{1}(Y_i < Y_j) + (1/2)\mathbf{1}(Y_i = Y_j)$. Then $E[g_i(Y_i \mid Y_j)\min\{Y_i, Y_j\}] = E[\mathbf{1}(Y_i < Y_j)\min\{Y_i, Y_j\}] = \Pr(Y_i < Y_j)E[\min\{Y_i, Y_j\}] = \mu_i/(\mu_i+\mu_j)^2$, where we used the fact that $\mathbf{1}(Y_i < Y_j)$ and $\min\{Y_i, Y_j\}$ are independent random variables for exponentially distributed Y_i and Y_j (Kulkarni 1995, p. 193). Using (4) and $E[(Y_i - Y_j)^+] = (\mu_j/\mu_i)/(\mu_i+\mu_j)$, firm i's expected cost is then

$$\Psi_i(\mu_i \mid \mu_j) = \kappa_i(\mu_i - \underline{\mu}_i) + \pi\left(p_i\frac{1}{\mu_i} + p_b\left(\frac{\mu_j/\mu_i}{\mu_i+\mu_j} + \frac{\mu_i}{(\mu_i+\mu_j)^2}\right)\right)$$

under the capacity game of §4. Differentiating Ψ_i yields

$$\frac{\partial\Psi_i}{\partial\mu_i} = \kappa_i + \pi\left(-\frac{p_i+p_b}{\mu_i^2} + p_b\left(\frac{1}{(\mu_i+\mu_j)^2} - \frac{\mu_i-\mu_j}{(\mu_i+\mu_j)^3}\right)\right)$$

$$\text{and} \quad \frac{\partial^2\Psi_i}{\partial\mu_i^2} = \pi\left(\frac{2p_i}{\mu_i^3} + \frac{2p_b}{\mu_i^3}\left(1 - \frac{3\mu_j/\mu_i}{(1+\mu_j/\mu_i)^4}\right)\right).$$

Note that $\xi(\alpha) \equiv (3\alpha)/(1+\alpha)^4 < 1$ for $\alpha > 0$, where the inequality follows because $(3\alpha)/(1+\alpha)^4$ is bounded above by $(3/4)^4 < 1$ as it initially increases from 0 at $\alpha = 0$, peaks at $\alpha = 1/3$, then converges to 0 as $\alpha \to \infty$. As a result,

$$\frac{\partial^2\Psi_i}{\partial\mu_i^2} = \pi\left(\frac{2p_i}{\mu_i^3} + \frac{2p_b}{\mu_i^3}\left(1 - \xi\left(\frac{\mu_j}{\mu_i}\right)\right)\right) > 0;$$

i.e., $\Psi_i(\mu_i \mid \mu_j)$ is convex. Hence, an equilibrium of the capacity game exists, and it is characterized by the system of equations $\partial\Psi_i/\partial\mu_i = 0$, $i = 1, 2$, or $(p_i+p_b)/\mu_i^2 - p_b \cdot (1/(\mu_i+\mu_j)^2 - (\mu_i-\mu_j)/(\mu_i+\mu_j)^3) = \kappa_i/\pi$. At the symmetric equilibrium $\mu = \mu_1 = \mu_2$ with identical parameters, this reduces to

$$\left(p_1 + \frac{3}{4}p_b\right)\frac{1}{\mu^2} = \frac{\kappa}{\pi} \quad \text{or}$$

$$\mu^{A4} = \sqrt{\frac{\pi}{\kappa}\left(p_1 + \frac{3}{4}p_b\right)} = \sqrt{\left(1 + \frac{1}{2}q\right)\frac{\theta\lambda\pi}{\kappa}},$$

which is identical to μ^C (see Proposition 3). Then it immediately follows that $\tau^{A4} = \tau^C$ and $\Psi_{\text{tot}}^{A4} = \Psi_{\text{tot}}^C$. Now consider the prevention/capacity game of §5. Firm i's expected cost is

$$\Psi_i = \phi_i\left(\frac{1}{\bar{\theta}_i - a_i} - \frac{1}{\bar{\theta}_i}\right) + \kappa_i(\mu_i - \underline{\mu}_i) + \pi\left(p_i\frac{1}{\mu_i} + p_b\left(\frac{\mu_j/\mu_i}{\mu_i+\mu_j} + \frac{\mu_i}{(\mu_i+\mu_j)^2}\right)\right),$$

where $p_i = (1-q_{ij})\theta_i\lambda$ and $p_b = (q_{ij}\theta_i + q_{ji}\theta_j)\lambda$. Following the solution procedure outlined in the proof of Proposition 4, we find that at the symmetric equilibrium with identical parameters,

$$\theta^{A4} = \left(\frac{1+q/2}{(1-q/4)^2}\frac{\phi^2}{\kappa\lambda\pi}\right)^{1/3} \quad \text{and} \quad \mu^{A4} = \left(\frac{(1+q/2)^2}{1-q/4}\frac{\phi\lambda\pi}{\kappa^2}\right)^{1/3}.$$

Comparing these with the expressions for θ^C and μ^C appearing in Proposition 4, we have $\theta^{A4} > \theta^C$ and $\mu^{A4} > \mu^C$. Next, consider A5, under which $g_i(Y_i \mid Y_j) = (1/2)\mathbf{1}(Y_i = Y_j)$ and thus $E[g_i(Y_i \mid Y_j)\min\{Y_i, Y_j\}] = 0$. Then firm i's expected cost is $\Psi_i(\mu_i \mid \mu_j) = \kappa_i(\mu_i - \underline{\mu}_i) + \pi(p_iE[Y_i] + p_bE[(Y_i - Y_j)^+])$ under the capacity game of §4. Note that $\max\{Y_i, Y_j\} = Y_j + (Y_i - Y_j)^+$, and therefore $(\partial/\partial\mu_i)E[(Y_i - Y_j)^+] = (\partial/\partial\mu_i)E[\max\{Y_i, Y_j\}]$. This implies $\partial\Psi_i/\partial\mu_i = \partial\Psi_c/\partial\mu_i$ and $\partial^2\Psi_i/\partial\mu_i^2 = \partial^2\Psi_c/\partial\mu_i^2$, and as a result, the first- and second-order conditions of the centralized and decentralized cases coincide. Hence, $\mu^{A5} = \mu^C$, and as a result, $\tau^{A5} = \tau^C$. Now consider the prevention/capacity game of §5. Firm i's expected cost is

$$\Psi_i = \phi_i\left(\frac{1}{\bar{\theta}_i - a_i} - \frac{1}{\bar{\theta}_i}\right) + \kappa_i(\mu_i - \underline{\mu}_i) + \pi\left(p_i\frac{1}{\mu_i} + p_b\frac{\mu_j/\mu_i}{\mu_i+\mu_j}\right),$$

where $p_i = (1-q_{ij})\theta_i\lambda$ and $p_b = (q_{ij}\theta_i + q_{ji}\theta_j)\lambda$. Following the steps similar to above, we find that at the symmetric equilibrium,

$$\theta^{A5} = \left(\frac{1+q/2}{(1-q/2)^2}\frac{\phi^2}{\kappa\lambda\pi}\right)^{1/3} \quad \text{and} \quad \mu^{A5} = \left(\frac{(1+q/2)^2}{1-q/2}\frac{\phi\lambda\pi}{\kappa^2}\right)^{1/3}.$$

Comparing these with θ^{A4} and μ^{A4}, we find $\theta^{A5} > \theta^{A4}$ and $\mu^{A5} > \mu^{A4}$. □

References

Airport Logistics (2010) Rome: Siemens wins maintenance contract for Fiumicino Airport. *Airport Logistics* (4):3.

Alstom (2010) Alstom and BrightSource Energy step up solar partnership. Press release (October 20), Alstom, Levallois-Perret, France.

Aydin G, Babich V, Beil D, Yang Z (2012) Decentralized supply risk management. Kouvelis P, Dong L, Boyabatli O, Li R, eds. *Handbook of Integrated Risk Management in Global Supply Chains* (John Wiley & Sons, Hoboken, NJ), 389–424.

Babich V, Burnetas AN, Ritchken PH (2007) Competition and diversification effects in supply chains with supplier default risk? *Manufacturing Service Oper. Management* 9(2):123–146.

Bakshi N, Kleindorfer PR (2009) Co-opetition and investment for supply-chain resilience. *Production Oper. Management* 18(6):583–603.

Barlow RE, Proschan F (1975) *Statistical Theory of Reliability and Life Testing: Probability Models* (Holt, Rinehart, and Winston, New York).

Bier VM (1997) Illusions of safety. Paper presented at the Workshop on Organizational Analysis in High Hazard Production Systems: An Academy/Industry Dialogue, Dedham, MA. http://www.chpra.wisc.edu/pdfs/IllusionsofSafety.pdf.

Bier VM (2005) Game-theoretic and reliability methods in counterterrorism and security. Wilson A, Limnios N, Keller-McNulty S, Armijo Y, eds. *Modern Statistical and Mathematical Methods in Reliability* (World Scientific, Singapore), 17–28.

Boyko A, Popov S (2010) Investigating the Sayano-Shushenskaya Hydro Power Plant disaster. *Power Magazine* (December 1), http://www.powermag.com/issues/features/Investigating-the-Sayano-Shushenskaya-Hydro-Power-Plant-Disaster_3229.html.

British Energy (2008) Results for the three months ended 29 June 2008 of the financial year ending 31 March 2009. Presentation, August 13, British Energy, East Kilbride, UK. http://www.british-energy.com/documents/Results_Presentation_Q1_0809_FINAL_WEBSITE.pdf.

Cachon GP, Feldman P (2011) Pricing services subject to congestion: Charge per-use fees or sell subscriptions? *Manufacturing Service Oper. Management* 13(2):244–260.

Cachon GP, Larivere MA (1999) Capacity choice and allocation: Strategic behavior and supply chain performance. *Management Sci.* 45(8):1091–1108.

Cachon G, Netessine S (2004) Game theoretic applications in supply chain analysis. Simchi-Levi D, Wu SD, Shen Z-J, eds. *Handbook of Quantitative Supply Chain Analysis: Modeling in the eBusiness Era* (Kluwer Academic Publishers, Boston), 13–66.

Chapman S (2008) Software "glitch" causes baggage chaos at Heathrow. *Computer World UK* (February 8), http://www.computerworlduk.com/news/security/7592/software-glitch-causes-baggage-chaos-at-heathrow.

Chopra S, Reinhardt G, Mohan U (2007) The importance of decoupling recurrent and disruption risks in a supply chain. *Naval Res. Logist.* 54(5):544–555.

Cornes R (1993) Dyke maintenance and other stories: Some neglected types of public goods. *Quart. J. Econom.* 108(1): 259–271.

Dada M, Petruzzi NC, Schwarz LB (2007) A newsvendor's procurement problem when suppliers are unreliable. *Manufacturing Service Oper. Management* 9(1):9–32.

Ford J (2003) Adverse outages lead to loan increase. *Engineer* (November 28), http://www.theengineer.co.uk/news/adverse-outages-lead-to-loan-increase/269044.article.

GE Energy (2010) Power generation combined cycle: Bently Nevada asset condition monitoring. http://www.ge-mcs.com/download/appsolutions/GEA18239_Combined_Cycle_Broch_r14-LR.pdf.

Hausken K (2002) Probabilistic risk analysis and game theory. *Risk Anal.* 22(1):17–27.

Hirshleifer J (1983) From weakest-link to best-shot: The voluntary provision of public goods. *Public Choice* 41(3):371–386.

Hu X, Gurnani H, Wang L (2013) Managing risk of supply disruptions: Incentives for capacity restoration. *Production Oper. Management* 22(1):137–150.

Kim S-H, Cohen MA, Netessine S, Veeraraghavan S (2010) Contracting for infrequent restoration and recovery of mission-critical systems. *Management Sci.* 56(9):1551–1567.

Kulkarni VG (1995) *Modeling and Analysis of Stochastic Systems* (Chapman & Hall, Boca Raton, FL).

Kunreuther H, Heal G (2003) Interdependent security. *J. Risk Uncertainty* 26(2–3):231–249.

Kwon HD, Lippman SA, McCardle KF, Tang CS (2010) Project management contracts with delayed payments. *Manufacturing Service Oper. Management* 12(4):692–707.

Milmo C (2010) Superjumbo engine fire led to "cascade of failures." *Independent (UK)* (November 19), http://www.independent.co.uk/travel/news-and-advice/superjumbo-engine-fire-led-to-cascade-of-failures-2138113.html.

Palmer JA, Danaher DA (2004) A series of preventable events leads to a power plant explosion. *EC&M* (November 1), http://ecmweb.com/mag/electric_series_preventable_events/.

Paté-Cornell E (2007) Probabilistic risk analysis versus decision analysis: Similarities, differences and illustrations. Abdellaoui M, Luce RD, Machina MJ, Munier B, eds. *Uncertainty and Risk* (Theory and Decision Library C, Vol. 41) (Springer, Berlin), 223–242.

Rausand M, Høyland MA (2004) *System Reliability Theory: Models, Statistical Methods, and Applications*, 2nd ed. (John Wiley & Sons, Hoboken, NJ).

Scudder GD (1984) Priority scheduling and spares stocking policies for a repair shop: The multiple failure case. *Management Sci.* 30(6):739–749.

Siemens (2012) Siemens and Valoriza achieve crucial milestone: Lebrija solar thermal power plant feeds power to grid. Press release (January 18), Siemens, Erlangen, Germany. http://www.siemens.com/press/en/feature/2012/energy/2012-01-lebrija.php.

Tomlin B (2006) On the value of mitigation and contingency strategies for managing supply-chain disruption risks. *Management Sci.* 52(5):639–657.

Tomlin B (2009) Disruption-management strategies for short life-cycle products. *Naval Res. Logist.* 56(4):318–347.

Varian H (2004) System reliability and free riding. Camp LJ, Lewis S, eds. *Economics and Information Security* (Kluwer Academic Publishers, Boston), 1–16.

Wang Y, Gilland WG, Tomlin B (2010) Mitigiating supply risk: Dual sourcing or process improvement? *Manufacturing Service Oper. Management* 12(3):489–510.

Yang Z, Aydin G, Babich V, Beil DR (2012) Using a dual-sourcing option in the presence of asymmetric information about supplier reliability: Competition vs. diversification. *Manufacturing Service Oper. Management* 14(2):202–217.

Young HP (1994) *Equity: In Theory and Practice* (Princeton University Press, Princeton, NJ).