

Readiness at the Critical Point

Coupled degradation, submodularity, and the limits of greedy sustainment scheduling

Dayan Ibrar

RESEARCH AGENDA · PREPRINT, NOT PEER REVIEWED · NO FLEET EXPERIMENT WAS RUN

Research agenda with theoretical contributions. No fleet experiment was run for this work. Every result below is a proved statement about an explicitly stated model, a computation from stated parameters, or a protocol for an experiment that has not yet been executed. Section 7.1 contains claims about current hardware and solvers. Those are stated as structural obstacles rather than as sourced device specifications, and they are the only statements here about the state of the world outside the model. Every figure is labelled conceptual or illustrative accordingly.

What this is

A research agenda: one problem in fleet sustainment, worked through as theory, plus the experiment that would test it. Nothing here was measured on a real fleet, and the paper is explicit throughout about which statements are proved, which are computed from stated parameters, and which are proposals.

The problem

An organization keeps a fleet of aircraft flyable with a fixed number of depot slots and a parts inventory of finite depth. When a part is short, it is often taken off another aircraft. That aircraft is now grounded, and when its part is needed somewhere else the process repeats. A shortfall therefore causes further shortfall. Readiness does not slide down smoothly as capacity tightens. It holds, and then it falls over.

What we do about it

We model that feedback, show the collapse is a saddle-node bifurcation, and then ask what the bifurcation does to the scheduling problem sitting on top of it. The expected answer was that it destroys the diminishing-returns structure greedy algorithms depend on. It does not: the objective stays submodular right up to the edge of collapse, so the standard guarantees survive intact.

What breaks instead is conditioning. Near the tipping point, predicted readiness becomes extremely sensitive to the parameters it was given. The difficulty moves off the optimizer and onto estimation, which is why the learned components need calibration as a structural requirement rather than as good practice. Submodularity does fail, but for an unrelated reason: real mission requirements are conjunctive, of the form “at least m aircraft capable at once”, and that makes the first few repairs show increasing rather than diminishing returns.

The stack, and what each part carries

Complex systems. [CARRIES THE RESULT]

The scarcity feedback is a dynamical system, and its collapse is a saddle-node bifurcation. That analysis is where the model and the tipping point come from.

Sections 3.4, 4, 5.4

Approximation algorithms. [CARRIES THE RESULT]

Submodular function theory supplies the greedy guarantee the main theorem preserves, the submodularity ratio that bounds it when it degrades, and the matroid structure of the feasible region.

Sections 2, 5, 6

Graph machine learning. [WHERE IT LANDS]

The model's coupling parameters are edge weights on a fleet graph: shared part lots, shared environments, shared repair capacity. They are not observable, so a graph neural network has to estimate them from maintenance records.

Section 8.2

Language models and agents. [WHERE IT LANDS]

Most maintenance history is free text. A language model turns those narratives into the graph's edges, and agents assemble and explain candidate schedules. The optimizer, not the agent, decides.

Sections 8.1, 8.5

Quantum optimization. [OPEN QUESTION]

Stated as one falsifiable hypothesis about the inner assignment problem near degeneracy, with the experiment that would refute it. Classical solvers dominate at every fleet size of practical interest today, and we say so.

Section 7

The central theorem needs two of those five. The other three are where its consequences land or are left open, and Section 1.3 says so rather than claiming a five-way necessity the work does not demonstrate.

If you read three sections

Section 5.5 for the theorem, Section 5.6 for the one place submodularity genuinely fails and the exact threshold at which it does, and Section 8.4 for why calibration carries more weight here than a better combinatorial algorithm would.

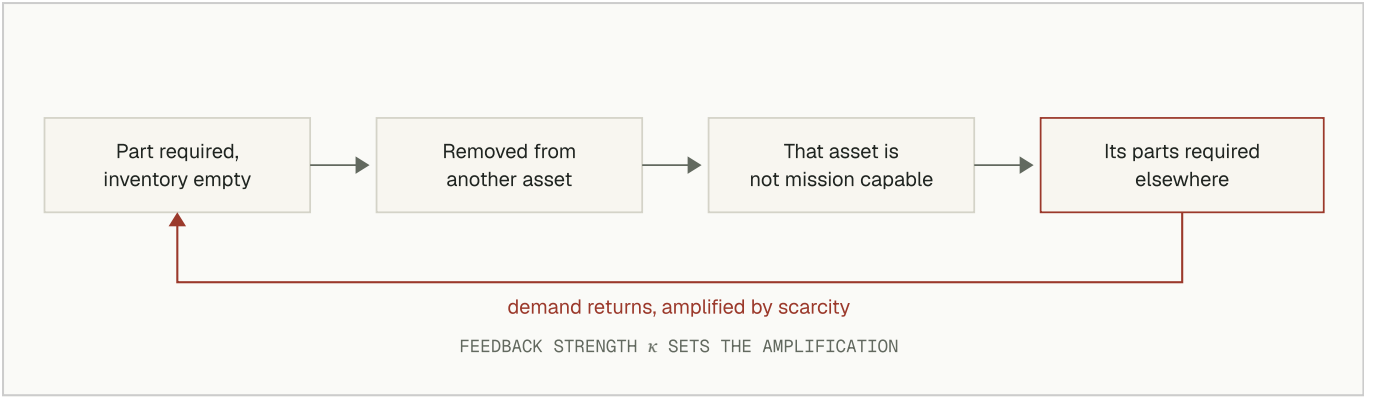


FIGURE 1 · CONCEPTUAL

The mechanism behind the collapse. A shortfall does not simply go unmet; it creates further demand, which is why the response to capacity is not proportional. This is the structure Section 3.4 parameterises, not an observation of any unit.

Abstract

Sustainment organizations that maintain fleets of complex repairable assets face a mechanism ordinary scheduling theory omits. It is cannibalization: deferred maintenance forces parts to be taken from other assets, which disables those assets, which creates further demand. The practice is documented across the services and inconsistently tracked [14]. What [14] establishes is that cannibalization is widespread; the behaviour of readiness under it is what this paper derives.

We show that a model of that mechanism does not predict a smooth decline in readiness as maintenance capacity tightens. It predicts readiness holding, and then collapsing.

This paper argues that the collapse is a saddle-node bifurcation, and traces what that implies for the optimization problem sitting on top of it. We formalize Readiness-Constrained Sustainment Scheduling (RCSS), a joint maintenance induction and logistics routing problem over a fleet whose failures are correlated through shared part lots, shared environments, and shared repair capacity.

Our first result is negative in a way that redirects the rest of the work. We show that expected readiness can be written as a scalar response function composed with a monotone submodular coverage function, and that under the feedback law we adopt this response is **concave** on the stable branch. Readiness is therefore monotone submodular, its submodularity ratio is one, and the standard guarantees apply. The cannibalization cascade, which is the feature practitioners find most alarming, does not damage the approximation guarantee. We expected the opposite and the algebra says otherwise.

That result is sharper than it first appears, and narrower. We give the exact condition on the feedback law under which it holds: writing b for the propensity as a function of the disabled fraction, the response is concave if and only if $x^2 b'(x)$ is nondecreasing. A linear propensity satisfies this; a saturating one need not, and we exhibit a strictly subcritical saturating law, with no bifurcation anywhere, under which readiness is strictly **super-**

modular. So the survival of the guarantee is a property of the feedback law, not of feedback, and the condition says which laws are safe.

What the bifurcation damages instead is conditioning. The derivative of the response diverges as $(\pi_{max} - \pi)^{-1/2}$ on approach to the tipping point, so the ratio of extreme slopes tends to zero even though concavity is preserved. The objective becomes stiff, not non-submodular: the marginal value of an induction varies by a large and growing factor across the operating range. Stiffness falls on estimation, not on optimization. We show that the sensitivity of predicted readiness to error in the coupling parameters diverges with the same exponent, which makes calibration of the graph learning and language model components a structural requirement rather than a refinement.

Submodularity does fail, but for an unrelated reason: mission requirements are conjunctive. When readiness is the probability that at least m assets are simultaneously capable, the response is S-shaped and early inductions show increasing returns. We give the exact condition under which this occurs, $m - 1 > (n - 1) q_{lo}$, and a general bound $\gamma \geq \inf h' / \sup h'$ covering both regimes.

Finally we ask which features of the problem leave many near-optimal schedules indistinguishable to single-element moves. We find two: proximity to the bifurcation, and the conjunctive objective. That is the only setting in which we think a quantum heuristic has a defensible reason to be tried. We state it as a falsifiable hypothesis with a test protocol, and we are explicit that classical solvers currently dominate at every fleet size of practical interest.

1. Introduction

1.1 The observation

Consider an organization responsible for keeping a fleet of aircraft mission capable. It has a fixed number of depot slots per period, a spare parts inventory with finite depth, and assets that degrade under use. The standard treatment is a scheduling problem: predict which assets will fail, induct them before they do, subject to capacity.

That treatment omits the dominant dynamic. When a part is needed and inventory is empty, it is frequently removed from another aircraft. The second aircraft is then not mission capable. If its part is later needed elsewhere, the process repeats. The practice is documented across the services, and the GAO has reported that it is both widespread and inconsistently tracked, with the Army at that time not recording it service-wide at all [14].

Structurally, **a shortfall creates demand**. In the model of Section 3.4, feedback of this kind produces two regimes: one in which capacity changes barely move readiness, one in which they move it enormously, and a sharp transition between them. Figure 1 draws the loop.

1.2 What this paper claims

The claim is not that cannibalization is harmful, which is obvious, or that readiness is nonlinear, which is known. The claim concerns where the difficulty lands.

A natural expectation, and the one this work began with, is that the cascade destroys the diminishing-returns structure that approximation algorithms rely on. Under the feedback law of Section 3.4 it does not: Section 5.5 proves readiness remains submodular on the stable branch, so the standard guarantees survive right up to the edge of collapse. The same section gives the exact condition on the feedback law for this to happen, and a counterexample where it fails.

The difficulty moves elsewhere. We establish four things:

1. **Readiness is submodular under the cascade** (Theorem 2, Section 5.5, proved). The feedback does not cost anything in approximation ratio.
2. **The objective becomes ill-conditioned near the bifurcation** (Section 5.4, derived). The slope ratio tends to zero while concavity survives.
3. **Sensitivity to parameter error diverges** with the same inverse square root exponent (Observation 1, Section 8.4, derived), so the data required to hold a fixed prediction error grows as the inverse of the margin. Section 8.4 is careful about what this does and does not imply for calibration.
4. **Conjunctive mission requirements do break submodularity** (Proposition 2, Section 5.6, proved with an exact threshold on the first induction), independently of the cascade.

We also state one refutable hypothesis about where non-classical optimization could matter (Section 7.2) and give the experiment that would settle it.

In practice, then, spend the next unit of effort on calibrated estimation and not on a better combinatorial algorithm. Optimization papers usually point the other way.

1.3 What this paper does not claim

It reports no measured results on any fleet. It demonstrates no quantum advantage and argues at length that none is available at

relevant scale. The experimental protocol in Section 9 is a design, not a report.

On the question of whether combining graph learning, quantum optimization, approximation theory, and complex systems is substantive or decorative, the honest answer is that the central theorem needs two of the four. Theorem 2 is proved from a bifurcation analysis of the scarcity feedback (complex systems) and from submodular function theory (approximation algorithms). Graph learning enters in Section 8 because the theorem's inputs have to be estimated from maintenance records, and Observation 1 shows that estimation is where the difficulty concentrates. Quantum optimization enters in Section 7 as a hypothesis about the inner subproblem and contributes nothing to the theorem. Two pillars carry the result; the other two are where its consequences land. We would rather state that than claim a four-way necessity the paper does not demonstrate.

2. Related work

Submodular maximization. Nemhauser, Wolsey and Fisher established that greedy maximization of a monotone submodular function under a **cardinality** constraint achieves $1 - (1 - 1/k)^k$, which tends to $1 - 1/e$ [1]. The companion paper gives $1/2$ for a single matroid and, more generally, $1/(p + 1)$ for an intersection of p matroids [2]. Călinescu, Chekuri, Pál and Vondrák later achieved $1 - 1/e$ for a **single** matroid using continuous greedy with pipage rounding [3]. These distinctions matter here, because the feasible region of RCSS is an intersection of two matroids and the three constants are different. Feige's $1 - 1/e$ inapproximability for maximum coverage, which holds under $P \neq NP$, makes the cardinality constant optimal [4].

Approximate submodularity. Das and Kempe introduced the submodularity ratio γ and proved a $1 - e^{-\gamma}$ guarantee for greedy on set functions that are not submodular, under a cardinality constraint [5]. We use it for the conjunctive objective of Section 5.6. It is not needed for the fractional objective, which Theorem 2 shows is submodular. Theorem 1 below bounds γ by a ratio of extreme response slopes; bounds of that shape, relating a guarantee to a ratio of curvature or derivative extremes, exist in the weak-submodularity literature, and we make no priority claim for the form of the argument.

Composition. That a nondecreasing concave function composed with a monotone submodular function is monotone submodular is a classical closure property, stated as Theorem 1 of Lin and Bilmes [18]. Lemma 1 is that result, restated on a bounded interval.

Cascading failure and critical phenomena. Motter and Lai analyzed load-redistribution cascades [6], Buldyrev and colleagues showed that interdependent networks fail more abruptly than isolated ones [7], and the percolation literature supplies the standard apparatus for locating critical points [8]. Branching-process facts used here are standard [15]. The closest prior art to our Section

3.4 is the power-systems line of Dobson, Carreras, Lynch and Newman, who model cascading blackouts as a branching process whose offspring mean rises with system loading, and who treat criticality and finite-size saturation carefully and against measured data [21]. Our feedback mechanism is the same shape as theirs, and we claim no novelty for it. Our contribution is not to this theory but to connecting it to an optimization guarantee.

Early warning signals. That the relaxation time and the variance of fluctuations diverge on approach to a fold is the basis of a large literature on anticipating critical transitions [20]. Section 8.4 uses it to argue that the margin to the tipping point is estimable from readiness data directly.

Inventory and sustainment. Sherbrooke's METRIC model is the foundation of multi-echelon repairable inventory [9] and, like most of that literature, treats demand as exogenous Poisson. The feature driving our analysis is that demand is endogenous.

Prognostics. Remaining useful life estimation on turbofan degradation data is a mature benchmark task following the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) datasets [10], whose generation methodology is documented separately [11].

Conformal prediction. Distribution-free coverage under exchangeability originates with Vovk, Gammerman and Shafer [12]; we follow the exposition in [13].

3. Problem formulation

3.1 Objects

Let A be a fleet of n repairable assets, D a set of depots, and $T = \{1, \dots, T\}$ a planning horizon.

An **induction** is a triple $a = (i, d, t)$ in $G = A \times D \times T$, meaning asset i enters depot d in period t . A **schedule** is a subset $S \subseteq G$.

Each asset is itself a graph of components, and is mission capable when every component in a designated critical set is serviceable. This conjunctive requirement returns in Section 5.6.

Table 1 lists the notation. The tables in this paper are inline exhibits, not numbered floats; each is read with the paragraph that introduces it. It is written in the same alphabet as the body: where the paper sets κ , the table says κ . Rows are ordered by where the symbol is first used.

SYMBOL	MEANING	FIRST USED
n	fleet size	3.1
S	a schedule, a set of inductions (i, d, t)	3.1
A, D, G	fleet, depot set, and the induction ground set $A \times D \times T$	3.1
T	planning horizon, and its cardinality	3.1
M_1, M_2	partition matroids for asset and depot capacity	3.2
k, k_i	a budget; and the induction cap for asset i	3.2
$c_{d,t}$	slots at depot d in period t	3.2
$\Lambda(S)$	the inductions in S a feasible routing can support	3.2
W, w_{ij}	coupling matrix over assets, and one of its entries	3.3
p_i	probability asset i fails within the horizon	3.3
β_0	baseline propensity for one failure to induce another	3.4
κ	feedback strength: how much scarcity amplifies that propensity	3.4
u	mission-capable fraction, $1 - X/n$	3.4
$\sigma(S), \sigma_a$	assets protected by S ; those protected by one induction a	3.5
$Q(S)$	covered failure mass, monotone submodular	3.5
$P(S)$	uncovered failure mass, $Q(G) - Q(S)$	3.5
X, X^*	expected disabled count, and its stable value	3.5
h	scalar response, readiness as a function of covered mass	5.1
q	covered failure mass as the argument of h	5.1

SYMBOL	MEANING	FIRST USED
ν	a generic monotone, normalized, submodular set function	5.2
ρ	slope ratio, $\inf h' / \sup h'$	5.3
$\gamma, \gamma_{U,k}$	submodularity ratio; and its value at budget k	5.3
π	uncovered failure mass per asset, $P(S)/n$	5.4
π_{max}	critical uncovered mass, the fold	5.4
Δ	discriminant $(1 - \beta_0)^2 - 4 \beta_0 \kappa \pi$; zero at the fold	5.4
x, x^*	disabled fraction X/n , and its stable value	5.4
x_c	disabled fraction at the fold	5.4
m	assets the mission requires simultaneously	5.6
R_m	conjunctive readiness, $Pr[N(S) \geq m]$	5.6
$N(S)$	number of capable assets under S	5.6
M_k	capable assets among the other $n - 1$, a Poisson-binomial count	5.6
q_{lo}, q_{hi}	per-asset capability when uncovered and covered	5.6

Two symbols are overloaded. We flag them here instead of re-naming, because context disambiguates each.

- T is both the horizon set and its cardinality.
- M is a matroid in 3.2 and a superset in Lemma 1, and M_k in 5.6 is a Poisson-binomial count, unrelated to either.

Three more collisions have been removed outright: the generic set function of Lemma 1 is ν , not m ; the independent-rounding indicators of Section 6 are Y_i , not X_i ; and the disabled count is X , since D is the depot set.

3.2 Constraints

Asset constraint. Each asset may be inducted at most k_i times over the horizon. Partitioning G by asset gives a partition matroid M_1 .

Capacity constraint. Depot d has $c_{d,t}$ slots in period t . Partitioning G by the pair (d, t) gives a partition matroid M_2 .

A schedule is admissible when it is independent in both, so the feasible region is the intersection $M_1 \cap M_2$. This is a bipartite degree-constrained structure, and it is important to be precise about what that costs. For a **linear** objective it is a transportation problem solvable in polynomial time. For a monotone submodular objective, greedy gives $1/3$ over an intersection of two matroids [2], not the $1/2$ of a single matroid and not the $1 - 1/e$ of continuous greedy on a single matroid [3]. When $T = 1$ and a single depot is available, M_2 collapses to a cardinality constraint and the $1 - 1/e$ results apply. We state which regime is in force wherever a constant appears.

Logistics feasibility. An induction (i, d, t) is *effective* only if the parts it requires are present at d by t . We write $\Lambda(S) \subseteq S$ for

the inductions supportable by some feasible routing. Maintenance and logistics are therefore not separable, which is why the problem is posed jointly.

3.3 Degradation with coupling

Let W be a symmetric nonnegative **coupling matrix** on A , where w_{ij} aggregates shared part lots, shared operating environment, and shared maintenance crews. Section 8 concerns how W is estimated; here it is given. Absent intervention, asset i suffers a primary failure within the horizon with probability p_i .

3.4 The scarcity feedback

Let $u \in [0, 1]$ be the mission-capable fraction. Secondary failures induced by cannibalization are modeled as a branching process whose mean offspring number depends on the state:

$$\beta(u) = \beta_0 \cdot (1 + \kappa \cdot (1 - u))$$

where β_0 , baseline propensity for one failure to induce another in a healthy fleet; κ , feedback strength: how much scarcity raises the chance of the next removal; $\beta(u)$, mean offspring number of the cascade at mission-capable fraction u ; u , fraction of the fleet that is mission capable.

β_0 is the baseline propensity for one failure to induce another in a healthy fleet, and $\kappa \geq 0$ is the **feedback strength**. Setting $\kappa = 0$ recovers exogenous demand.

For a Galton-Watson process with mean offspring $\beta < 1$, the expected total progeny including the root is $1 / (1 - \beta)$ and the process is almost surely finite. At $\beta = 1$, for a non-degenerate

offspring distribution, the process remains almost surely finite but its expectation diverges [15].

3.5 Coverage and the readiness objective

Inductions replace parts, and a replaced part protects every asset depending on that lot. Let $\sigma(S) \subseteq A$ be the set of assets protected by effective inductions in S , and define the **covered failure mass**

$$Q(S) = \sum_{i \in \sigma(S)} p_i$$

where $\sigma(S)$, the set of assets protected by the effective inductions in S ; $Q(S)$, covered failure mass; monotone, normalized and submodular; S , a schedule; in the submodularity-ratio definition, the disjoint test set; p_i , probability that asset i fails within the horizon.

Assumption 1 (per-induction effectiveness). Each induction α protects a fixed asset set σ_α , so that $\sigma(S) = \bigcup_{\alpha \in S} \sigma_\alpha$.

Assumption 1 is what makes Q a coverage function, and it is a genuine restriction rather than bookkeeping. Section 3.2 allows a

$$X = P(S) / (1 - \beta(1 - X/n)) \quad (1)$$

where X , expected number of disabled assets; $\beta(u)$, mean offspring number of the cascade at mission-capable fraction u ; $P(S)$, uncovered failure mass, $Q(G) - Q(S)$; n , fleet size; S , a schedule; in the submodularity-ratio definition, the disjoint test set.

Readiness is the expected mission-capable fraction, $R(S) = (n - X^*(P(S))) / n$, where X^* is the stable equilibrium of the associated dynamics defined in Section 5.4.

Definition 1 (RCSS). Given A , D , T , matroids M_1 , M_2 , routing network and part requirements, coupling W , failure probabilities p , and feedback parameters (β_0, κ) , find an admissible schedule S with a supporting routing maximizing $R(S)$.

$$R(S) = 1 - (Q(G) - Q(S)) / (n(1 - \beta_0)) = c + Q(S) / (n(1 - \beta_0)), \quad c = 1 - Q(G) / (n(1 - \beta_0))$$

where β_0 , baseline propensity for one failure to induce another in a healthy fleet; $Q(S)$, covered failure mass; monotone, normalized and submodular; $R(S)$, expected mission-capable fraction under schedule S ; n , fleet size; S , a schedule; in the submodularity-ratio definition, the disjoint test set.

which is affine and strictly increasing in $Q(S)$. An affine strictly increasing map preserves the argmax, so maximizing R exactly is maximizing the weighted coverage $Q(S)$ exactly, and NP-hardness transfers at once. With $T = 1$ and one depot the constraint is a cardinality bound on the number of inductions, and maximum coverage under a cardinality bound is NP-hard by a standard reduction from set cover.

The ratio does **not** transfer for free, because the offset c is non-negative and an additive offset makes a maximization problem easier to approximate. We need $c = 0$, that is $Q(G) = n(1 - \beta_0)$,

schedule to be only partly supportable by a routing, and when effectiveness depends on the whole schedule through $\Lambda(S)$, adding an induction can consume the capacity that made another one effective. In that case σ need not be monotone and Q need not be submodular. Everything from Lemma 1 onward assumes Q is monotone submodular, so everything from Lemma 1 onward assumes Assumption 1; where routing feasibility couples inductions, $Q(S)$ must be read as $Q(\Lambda(S))$ and we do not prove submodularity.

Under Assumption 1, distinct inductions can still protect overlapping asset sets, so Q is a weighted coverage function: monotone, normalized ($Q(\emptyset) = 0$), and **submodular**, not modular. This overlap is what makes the problem combinatorially hard, and Section 4 depends on it. Write $P(S) = Q(G) - Q(S)$ for the uncovered mass.

Expected disabled count X satisfies a self-consistency condition, since each unit of primary failure mass generates $1 / (1 - \beta(u))$ disablements in expectation while u itself depends on X . We write X rather than D , which already denotes the depot set:

4. Hardness

Proposition 1. RCSS is NP-hard even when $\kappa = 0$, $T = 1$, and a single depot is available. On the subfamily of instances with homogeneous $p_i = 1 - \beta_0$, no polynomial-time algorithm achieves a ratio of $1 - 1/e + \varepsilon$ for any fixed $\varepsilon > 0$ unless $P = NP$.

Proof. Set $\kappa = 0$, so $\beta = \beta_0$ is constant and (1) gives $X^* = P(S)/(1 - \beta_0)$. Then

and it is worth being careful about how to get it. Rescaling the weights of an arbitrary weighted instance does not work, because the p_i are probabilities: the required factor can push p_i above one, as it does for $n = 10$, $\beta_0 = 0.5$ and weights $(1, 0.01, \dots, 0.01)$, where the largest rescaled value is 4.587 . Instead reduce from **unweighted** maximum coverage, where every element has weight $p_i = 1 - \beta_0 \leq 1$ uniformly and $\sigma(G) = A$, so $Q(G) = n(1 - \beta_0)$ and $c = 0$ exactly. Feige's inapproximability is for unweighted maximum coverage in any case, so nothing is lost. Since $T = 1$ and one depot give $|G| = n$, the construction real-

izes coverage instances with at most n sets; replicating each element a fixed number of times lifts that restriction while preserving unit weights and the approximation ratio. Then $R(S) = Q(S)/(n(1 - \beta_0))$ is a positive scalar multiple of $Q(S)$, the reduction is ratio-preserving, and Feige's bound, that maximum coverage cannot be approximated within $1 - 1/e + \varepsilon$ for any fixed $\varepsilon > 0$ unless $P = NP$ [4], carries over. ■

The restriction is not a technicality to be waved through. Without it the offset is strictly positive and the empty schedule can already beat $1 - 1/e$: with $n = 10$, $p_i = 0.1$ and $\beta_0 = 0.5$ the offset is $c = 0.8$, so $R(\emptyset)$ is 0.8 against a maximum of 1.0, a ratio of 0.8 attained by doing nothing. Note also that $c = 0$ forces $x^*(\emptyset) = 1$, so the ratio-preserving subfamily sits on the boundary of the physical range of Section 5.4: the hardness is real, but it is hardness of the worst case, not of a fleet anyone operates. The same algebra runs the other way in Section 6, where a nonnegative offset in the margin objective only improves the achieved ratio.

The reduction is the reason Section 3.5 defines Q as a coverage function rather than a modular one. Had Q been modular, the $\kappa = 0$ instance would be a linear objective over a matroid intersection, hence solvable exactly in polynomial time once logistics feasibility is relaxed, and no hardness would follow from the objective itself. The $\Lambda(S)$ constraint of Section 3.2 is not matroidal and remains a separate source of difficulty. The overlap between the asset sets protected by different inductions is the entire source of difficulty in the benign regime.

Two consequences. First, the problem is hard even with no feedback. Second, under a cardinality constraint the $1 - 1/e$ of greedy [1] is optimal, so improvement must come from structure, not from a better algorithm for the same objective.

5. The structure of the readiness function

5.1 Readiness is a scalar response composed with a coverage function

R depends on S only through $Q(S)$. Define the **response function**

$$h(q) = (n - X^*(Q(G) - q)) / n, \quad \text{so that} \quad R(S) = h(Q(S))$$

where X^* , stable expected disabled count; $x^* = X^*/n$; q , covered failure mass, the argument of the response h ; $Q(S)$, covered failure mass; monotone, normalized and submodular; $R(S)$, expected mission-capable fraction under schedule S ; h , scalar response: readiness as a function of covered mass; n , fleet size; S , a schedule; in the submodularity-ratio definition, the disjoint test set.

All combinatorial difficulty sits in Q , and all nonlinearity sits in the scalar h . This separation is what makes the analysis tractable: a question about a set function over an exponentially large domain becomes a question about the curvature of a one-dimensional curve, and the curvature is something a bifurcation analysis can settle.

5.2 Concavity of the response transfers submodularity

Lemma 1 (classical; see Lin and Bilmes [18], Theorem 1). Let $f(S) = h(v(S))$ with v monotone, normalized and submodular, and let h be nondecreasing and concave on an interval containing the range of v . Then f is monotone submodular.

This is a standard closure property of submodular functions and we claim no credit for it. We give the proof because the usual statement takes h concave on all of the reals, while we need it only on an interval containing the range of v , and on a bounded interval the order of the two steps matters.

Proof. Monotonicity is immediate. For submodularity take $L \subseteq M$ and $a \notin M$. Monotonicity of v gives $v(L) \leq v(M)$, and submodularity of v gives $v(L \cup a) - v(L) \geq v(M \cup a) - v(M) \geq 0$. Write δ_L and δ_M for these two increments. Then

$$\begin{aligned} f(L \cup a) - f(L) &= h(v(L) + \delta_L) - h(v(L)) \\ f(M \cup a) - f(M) &= h(v(M) + \delta_M) - h(v(M)) \end{aligned}$$

where h , scalar response: readiness as a function of covered mass; $v(\cdot)$, a monotone, normalized, submodular set function; f , the composed objective $h \circ v$; L , the base set in the submodularity-ratio definition.

Since h is nondecreasing, its increment from a fixed left endpoint is nondecreasing in the length; since h is concave, its increment over a fixed length is nonincreasing in the left endpoint. Shorten the length first and move the endpoint second:

$$h(v(L) + \delta_L) - h(v(L)) \geq h(v(L) + \delta_M) - h(v(L)) \geq h(v(M) + \delta_M) - h(v(M))$$

where h , scalar response: readiness as a function of covered mass; $v(\cdot)$, a monotone, normalized, submodular set function; L , the base set in the submodularity-ratio definition.

The order matters. Taken the other way round the argument would evaluate h at $v(M) + \delta_L$, which can exceed every value v

attains and so can fall outside the interval the hypothesis covers. In the order above every point lies in $[v(L), v(M \cup a)]$, and

both endpoints are values of v . ■

Two cautions about the converse, which we do **not** claim. Concavity is sufficient, not necessary: on a finite range only increments of realizable lengths are tested, so a non-concave h can still produce a submodular f . Monotonicity of v is also required, and nonnegativity of $v(S)$ is not a substitute. Take v **modular and normalized** with $v(\emptyset) = 0$, $v(a) = 1$, $v(b) = -3$, $v(ab) = -2$, and $h(z) = \sqrt{(z + 4)}$, which is increasing and concave on

$[-4, \infty)$. Then $f(a) + f(b) = 3.2361 < 3.4142 = f(ab) + f(\emptyset)$, so submodularity fails on a function that satisfies every hypothesis except monotonicity. Monotonicity is doing the work. Normalization is not used in the proof at all; we keep it only so that f inherits $f(\emptyset) = h(0)$.

5.3 The submodularity ratio is bounded by the slope ratio

Recall the submodularity ratio of Das and Kempe [5]: for monotone f and a budget k ,

$$\gamma_{U,k} = \min \text{ over } (L, S), S \cap L = \emptyset, |S| \leq k, \text{ of } \sum_{a \in S} [f(L \cup \{a\}) - f(L)] / [f(L \cup S) - f(L)]$$

where $\gamma_{U,k}$, submodularity ratio at budget k , minimised over disjoint pairs with $|S| \leq k$; f , the composed objective $h \circ m$; L , the base set in the submodularity-ratio definition; S , a schedule; in the submodularity-ratio definition, the disjoint test set; k , a budget, or the count of assets already covered.

taken over pairs with $f(L \cup S) > f(L)$, so the ratio is defined; $\gamma \in [0, 1]$, and $\gamma = 1$ when f is submodular.

Theorem 1. Let $f(S) = h(v(S))$ with v monotone, normalized and submodular, and h differentiable and strictly increasing on

an interval containing the range of v . Let $\rho = \inf h' / \sup h'$ over that interval. Then $\gamma \geq \rho$.

Proof. Fix L and disjoint S . For each $a \in S$, the mean value theorem gives some ξ_a between $v(L)$ and $v(L \cup a)$ with

$$f(L \cup a) - f(L) = [v(L \cup a) - v(L)] \cdot h'(\xi_a) \geq [v(L \cup a) - v(L)] \cdot \inf h'$$

where h' , first derivative of the response with respect to covered mass; $v(\cdot)$, a monotone, normalized, submodular set function; f , the composed objective $h \circ m$; L , the base set in the submodularity-ratio definition; ξ , intermediate point supplied by the mean value theorem.

Submodularity of v gives $\sum_{a \in S} [v(L \cup a) - v(L)] \geq v(L \cup S) - v(L)$, so

$$\sum_{a \in S} [f(L \cup \{a\}) - f(L)] \geq [v(L \cup S) - v(L)] \cdot \inf h'$$

where h' , first derivative of the response with respect to covered mass; $v(\cdot)$, a monotone, normalized, submodular set function; f , the composed objective $h \circ m$; L , the base set in the submodularity-ratio definition; S , a schedule; in the submodularity-ratio definition, the disjoint test set.

Applying the mean value theorem once more, $f(L \cup S) - f(L) = [v(L \cup S) - v(L)] \cdot h'(\xi) \leq [v(L \cup S) - v(L)] \cdot \sup h'$. Dividing gives a ratio of at least ρ for every admissible pair. ■

Remark (discrete form). Section 5.6 applies Theorem 1 to an h defined on $\{0, 1, \dots, n\}$, where no derivative exists. The same argument runs with increments in place of derivatives: replace $\inf h'$ by the smallest unit increment $\min_j [h(j+1) - h(j)]$ and $\sup h'$ by the largest, and every step of the proof goes through unchanged, since $h(l + s) - h(l)$ is a sum of s unit increments. We write ρ for either form and say which is meant where it matters.

Corollary 1. Under a **cardinality** constraint, greedy maximization of f returns at least $(1 - e^{-\rho}) \cdot \text{OPT}$, combining Theorem 1 with [5]. We state this for cardinality because that is the setting in which the $1 - e^{-\rho}$ guarantee is proved; it does not transfer to a matroid intersection, and we make no claim there.

The reading of Theorem 1 is that the guarantee is controlled by how much the marginal value of maintenance varies across the operating range. Sections 5.5 and 5.6 show the bound can be arbitrarily loose, so it is informative only where h is not concave.

5.4 The response near the tipping point

Substituting $\beta(u) = \beta_0(1 + \kappa(1 - u))$ and $u = 1 - X/n$ into (1), writing $x = X/n$ and $\pi = P(S)/n$:

$$x \cdot (1 - \beta_0 - \beta_0 \kappa x) = \pi \quad (2)$$

where β_0 , baseline propensity for one failure to induce another in a healthy fleet; x , disabled fraction, X/n ; κ , feedback strength: how much scarcity raises the chance of the next removal; π , uncovered failure mass per asset, $P(S)/n$.

Write $g(x)$ for the left side, a downward parabola with maximum

$$\pi_{max} = (1 - \beta_0)^2 / (4 \beta_0 \kappa) \quad \text{attained at} \quad x_c = (1 - \beta_0) / (2 \beta_0 \kappa)$$

where π_{max} , critical uncovered mass, the fold beyond which no healthy equilibrium exists; β_0 , baseline propensity for one failure to induce another in a healthy fleet; x_c , disabled fraction at the fold, where the two equilibria collide; κ , feedback strength: how much scarcity raises the chance of the next removal.

Write $\Delta(\pi) = (1 - \beta_0)^2 - 4 \beta_0 \kappa \pi$ for the discriminant of (2), which is zero exactly at π_{max} .

(a) Existence threshold. For $\pi > \pi_{max}$ equation (2) has no real root: above a critical uncovered mass the healthy equilibrium

ceases to exist. The two equilibria collide and annihilate, which is a saddle-node bifurcation, and $x_c - x^*(\pi) = \sqrt{\pi_{max} - \pi} / \sqrt{\beta_0 \kappa}$ reproduces the generic normal-form scaling.

(b) Square-root divergence. The stable root is

$$x^*(\pi) = [(1 - \beta_0) - \sqrt{\Delta(\pi)}] / (2 \beta_0 \kappa), \quad dx^*/d\pi = \Delta(\pi)^{-1/2}$$

where β_0 , baseline propensity for one failure to induce another in a healthy fleet; x^* , stable disabled fraction, the smaller root of equation (2); Δ , discriminant $(1 - \beta_0)^2 - 4\beta_0\kappa\pi$; it reaches zero exactly at the fold; κ , feedback strength: how much scarcity raises the chance of the next removal; π , uncovered failure mass per asset, $P(S)/n$.

Since $\Delta(\pi) = 4 \beta_0 \kappa (\pi_{max} - \pi)$, the derivative diverges as $(\pi_{max} - \pi)^{-1/2}$.

(c) Slope ratio. With $h(q) = 1 - x^*(Q(G)/n - q/n)$ the chain rule gives $h'(q) = (1/n) \Delta^{-1/2}$, so over an operating range from π_{lo} to $\pi_{hi} < \pi_{max}$,

$$\rho = \inf h' / \sup h' = \sqrt{\Delta(\pi_{hi}) / \Delta(\pi_{lo})} = \Theta((\pi_{max} - \pi_{hi})^{1/2})$$

where π_{max} , critical uncovered mass, the fold beyond which no healthy equilibrium exists; π_{lo} , π_{hi} , the endpoints of the operating range in uncovered mass; Δ , discriminant $(1 - \beta_0)^2 - 4\beta_0\kappa\pi$; it reaches zero exactly at the fold; ρ , slope ratio $\inf h' / \sup h'$, which bounds γ from below and lies in $[0, 1]$; h' , first derivative of the response with respect to covered mass.

Dynamics and stability. Equation (1) is an equilibrium condition, and stability belongs to a flow, not to an algebraic root, so we state the flow: $\dot{x} = \pi - g(x)$, whose equilibria are exactly the roots of (2). Then $\partial \dot{x} / \partial x = -g'(x)$, and at the smaller root $g'(x^*) = \sqrt{\Delta(\pi)}$, giving $\partial \dot{x} / \partial x = -\sqrt{\Delta} < 0$. The smaller root is therefore the stable equilibrium for all $\pi < \pi_{max}$, and it is the branch referred to throughout.

We call the point $\pi = \pi_{max}$ the **fold**. One clarification is worth making, because Section 3.4 may suggest otherwise: at the fold the mean offspring number is $\beta = \beta_0(1 + \kappa x_c) = (1 + \beta_0)/2 < 1$. The collapse happens while the branching process is still strictly subcritical. It is a fold of the self-consistency equation (2), not branching criticality.

Numerics. The naive fixed-point iteration $x \leftarrow \pi / (1 - \beta_0 - \beta_0 \kappa x)$ has $|T'(x^*)| = \beta_0 \kappa (x^*)^2 / \pi$, which stays below one for

every $\pi < \pi_{max}$ and reaches one exactly at the fold. Contraction is therefore lost at the fold itself, not before it. But it degrades steeply on approach: the modulus is 0.52 at $0.9 \pi_{max}$ and 0.94 at $0.999 \pi_{max}$, and since it depends only on the ratio π/π_{max} those two values hold for every (β_0, κ) . Near the tipping point, solve the quadratic directly or integrate the flow; do not iterate the map.

Physical range. These quantities are meaningful only where the fold is reachable with $x^* \in [0, 1]$. Since $x_c > 1$ whenever $\kappa < (1 - \beta_0)/(2 \beta_0)$, readiness would go negative before the tipping point in that region. The bifurcation is physically relevant for $\kappa \geq (1 - \beta_0)/(2 \beta_0)$, and statements about approaching it should be read with that restriction. Figure 2 shows the response and its curvature together, with the reparameterization that connects them.

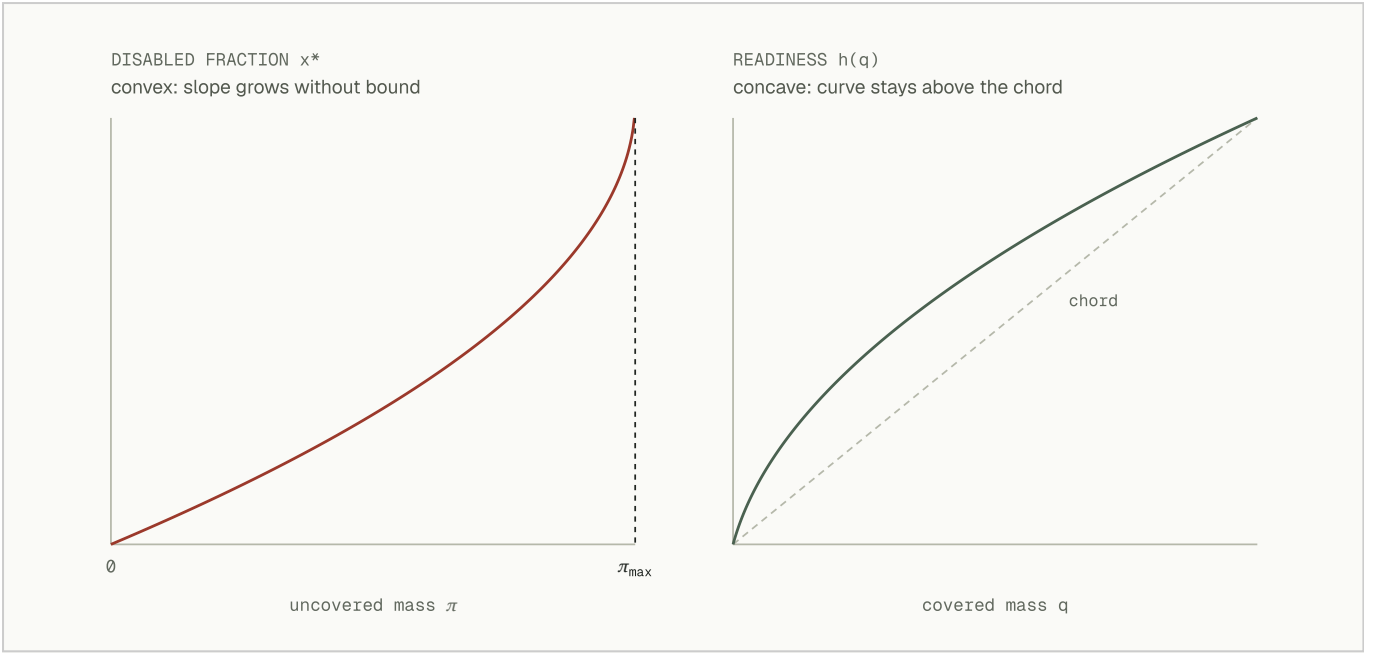


FIGURE 2 · ILLUSTRATIVE COMPUTATION

Why the cascade does not break submodularity. Computed from equation (2) with baseline propensity 0.45 and feedback strength 2.2. Disabled fraction is convex in uncovered mass and its slope diverges at the fold. Readiness is that quantity negated under a decreasing reparameterisation, which flips the curvature: it sits above the chord joining its endpoints, so it is concave, and Lemma 1 then gives submodularity. The divergence on the left is the stiffness of Section 5.5, not a loss of concavity on the right.

5.5 The cascade preserves submodularity

Theorem 2. Adopt Assumption 1, fix $\beta_0 \in (0,1)$, and let κ satisfy $Q(G)/n < \pi_{\max}$, so that every schedule including the empty one lies on the stable branch. Then in the model of Section 3.4 the response h is increasing and **concave** in q . Consequently R is monotone submodular and $\gamma = 1$.

Two remarks on the quantifiers, because an earlier draft of this paper got them wrong. First, $Q(G)/n < \pi_{\max}$ is in fact a constraint **on the feedback strength**, written out as $\kappa < n(1 - \beta_0)^2 / (4 \beta_0 Q(G))$, so the theorem cannot also hold for every $\kappa \geq$

0, and we do not claim it does; Section 10 gives the window in full. The hypothesis is not decorative: the worst case is the empty schedule, which leaves the whole failure mass uncovered at $\pi = Q(G)/n$, and if that exceeds $\pi_{\max}$ then h is undefined there, $R(\emptyset)$ does not exist, and the claim that R is monotone submodular is not a statement about a function. Second, the theorem is about the **particular feedback law** of Section 3.4. Proposition 3 below shows that is not a formality.

Proof. From 5.4(b), $h'(q) = (1/n) \Delta^{-1/2} > 0$. Differentiating, with $d\Delta/dq = 4 \beta_0 \kappa / n > 0$:

$$h''(q) = (1/n)(-1/2) \Delta^{-3/2} (4 \beta_0 \kappa / n) = -2 \beta_0 \kappa / (n^2 \Delta^{3/2}) \leq 0$$

where β_0 , baseline propensity for one failure to induce another in a healthy fleet; q , covered failure mass, the argument of the response h ; Δ , discriminant $(1 - \beta_0)^2 - 4\beta_0\kappa\pi$; it reaches zero exactly at the fold; κ , feedback strength: how much scarcity raises the chance of the next removal; h'' , second derivative of the response; non-positive, and strictly negative for $\kappa > 0$; n , fleet size.

strictly negative for $\kappa > 0$, and zero for $\kappa = 0$ where h is affine. So h is concave. Q is monotone, normalized and submodular by Section 3.5, so Lemma 1 gives submodularity of R , and a monotone submodular function has $\gamma = 1$. ■

The concavity is a property of the feedback law, not of feedback. Write $b(x) = \beta(1 - x)$ for the propensity as a function of the disabled fraction, so that (2) reads $g(x) = x(1 - b(x)) = \pi$. On the increasing branch $x^* = g^{-1}(\pi)$, and since h is $1 - x^*$ composed with a decreasing affine map, h is concave exactly when g is concave. Differentiating twice gives $g''(x) = -2 b'(x) - x b''(x)$, which proves the following.

Proposition 3. On the increasing branch, h is concave if and only if $2 b'(x) + x b''(x) \geq 0$, equivalently if and only if $x^2 b'(x)$ is nondecreasing.

The linear law of Section 3.4 has $b'(x) = \beta_0 \kappa$ constant and $b''(x) = 0$, so the criterion holds with room to spare, which is why Theorem 2 goes through. It is not a generic property of positive feedback. A **saturating** propensity, which is at least as plausible physically since the willingness to cannibalize has a ceiling, can fail it. Take

$$b(x) = 0.20 + 0.50 (1 - e^{-x/0.02})$$

where X , disabled fraction, X/n .

so that b stays within $[0.20, 0.70]$ and the feedback is **strictly subcritical everywhere**, $g' > 0$ on the whole range, and there is **no fold at all**. Then $2b' + x b'' = e^{-x/0.02}(50 - 1250x)$, which turns negative past $x = 0.04$. With $n = 10$ homogeneous assets at $p_i = 0.034$ and each induction protecting one fresh asset, the marginals of readiness are $0.011744, 0.011988, 0.012372$: they **increase**, and $R(a) + R(b) = 1.798165 < 1.798409 = R(ab) + R(\emptyset)$. Readiness is strictly supermodular. The same instance under the linear law has marginals

$0.008139, 0.007842, 0.007576$, diminishing as Theorem 2 requires.

We state this plainly because it bounds what the rest of the paper may claim. The cascade does not damage the approximation guarantee **under the feedback law of Section 3.4**, and Proposition 3 says exactly which laws share that property. It is not a general fact about cannibalization, and the counterexample above is not a boundary case: it has no bifurcation and never approaches criticality.

Consequences for the algorithm, stated by constraint type. The constants below are different and the literature is often quoted loosely, so we state which applies where.

CONSTRAINT	ALGORITHM	GUARANTEE	SOURCE
Cardinality bound k	greedy	$1 - (1 - 1/k)^k \rightarrow 1 - 1/e$	[1]
Cardinality bound k	any polynomial algorithm	$1 - 1/e$ is optimal	[4], Prop. 1
One matroid	greedy	$1/2$	[2]
One matroid	continuous greedy	$1 - 1/e$	[3]
Intersection of p matroids	greedy	$1/(p+1)$, so $1/3$ here	[2]
Intersection of p matroids	local search	$1/(p + \varepsilon)$, so $1/(2 + \varepsilon)$ here	[19]

The feasible region of RCSS is $M_1 \cap M_2$, so the applicable constant is $1/(2 + \varepsilon)$ by local search [19], which improves on the $1/3$ that greedy gives [2]; both degrade to the cardinality case only when $T = 1$ and a single depot reduce M_2 to a cardinality bound. We are not aware of a $1 - 1/e$ result for two matroids and do not claim one.

This runs against the intuition that motivated the investigation. The feedback loop producing the collapse is genuine positive feedback and it nonetheless leaves the set function submodular. The reason is visible in the proof: positive feedback makes the

disabled count convex in uncovered mass ($d^2x^*/d\pi^2 = 2\beta o \kappa \Delta^{-3/2} > 0$), and readiness is its negation under an affine decreasing reparameterization, which flips convexity into concavity.

What degrades instead is conditioning. The condition number $1/\rho$ diverges as $(\pi_{max} - \pi)^{-1/2}$. The objective is submodular but stiff: the marginal value of an induction varies by a large and growing factor across the operating range. Theorem 1 remains valid but is loose here, since the true γ is one. Figure 3 plots the slope ratio against the margin to the fold. Stiffness is not an optimization problem. Section 8.4 shows it is an estimation problem.

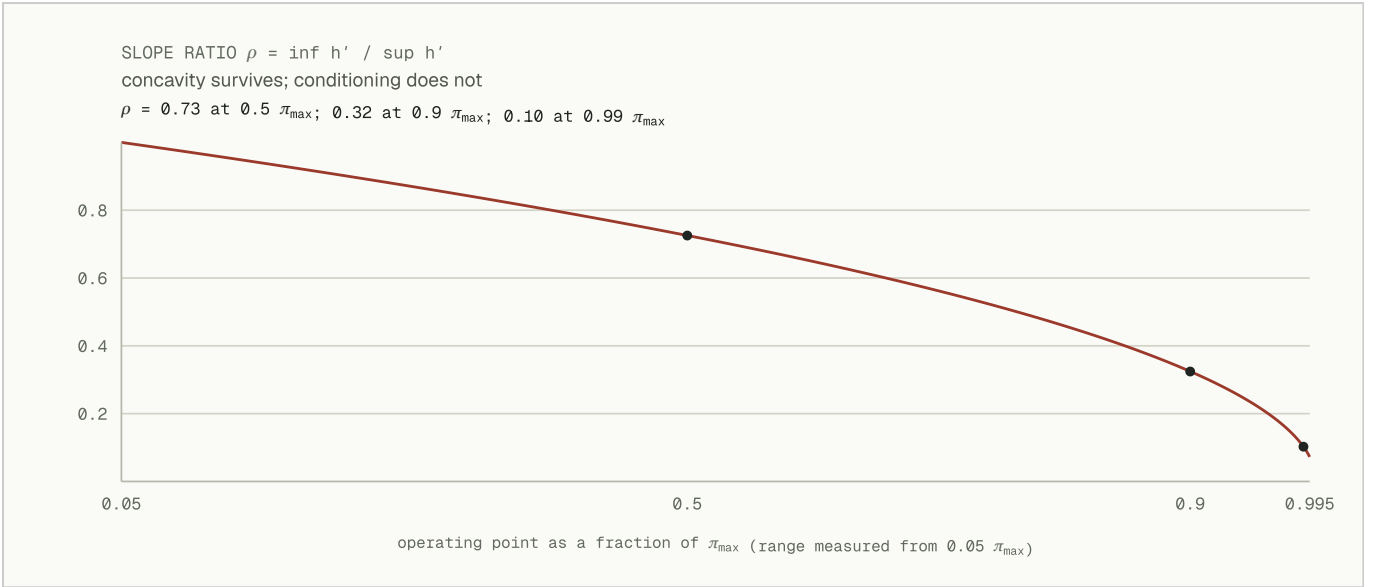


FIGURE 3 · ILLUSTRATIVE COMPUTATION

What the bifurcation actually costs. The ratio of extreme response slopes falls toward zero as the operating point approaches the fold, so the marginal value of an induction varies by a growing factor across the range. Theorem 2 still holds throughout: the objective stays submodular and becomes stiff. Stiffness is an estimation problem, which Figure 7 quantifies.

5.6 Where submodularity actually fails

Mission requirements are not fractional. An operator needs some number of assets simultaneously capable. Let $N(S)$ be the number of capable assets and

$$R_m(S) = \Pr[N(S) \geq m]$$

where R_m , conjunctive readiness: the probability that at least m assets are capable; $N(S)$, number of capable assets under schedule S ; m , the number of assets the mission requires simultaneously; S , a schedule; in the submodularity-ratio definition, the disjoint test set.

Assume homogeneous failure probabilities so that coverage reduces to a count, that each induction protects exactly one previously unprotected asset so the count rises by one at a time, and take per-asset capability q_{hi} when covered and q_{lo} when not, independent across assets. Then $N(S)$ is Poisson-binomial and R_m depends only on the covered count k , so the relevant h is defined on $\{0, \dots, n\}$ with unit increments, where concavity and the discrete increment condition coincide. The second assumption matters for the reading of the result: it removes the coverage overlap that Section 4 identifies as the source of difficulty in the benign regime, so the complementarity found below is separate from overlap, not a consequence of it.

Upgrading one asset from q_{lo} to q_{hi} gives the identity

$$h(k+1) - h(k) = (q_{hi} - q_{lo}) \cdot \Pr[M_k = m - 1]$$

where q_{lo} , per-asset capability probability when the asset is not covered; q_{hi} , per-asset capability probability when the asset is covered; M_k , number of capable assets among the other $n - 1$, a Poisson-binomial count; h , scalar response: readiness as a function of covered mass; m , the number of assets the mission requires simultaneously; k , a budget, or the count of assets already covered.

where M_k counts the other $n - 1$ assets. The marginal therefore increases at k exactly when $\Pr[M_{k+1} = m - 1] > \Pr[M_k = m - 1]$.

Proposition 2. With the assumptions above, and with $n \geq 2$ and $1 \leq m \leq n$, the marginal value of the **first** induction is increasing if and only if

$$m - 1 > (n - 1) \cdot q_{lo}$$

where q_{lo} , per-asset capability probability when the asset is not covered; n , fleet size; m , the number of assets the mission requires simultaneously.

When it is increasing, R_m is not submodular and $\gamma_{U,k} < 1$ for every budget $k \geq 2$: take $L = \emptyset$ and $S = \{a, b\}$, so the sum of singleton gains is $2(h(1) - h(0))$, which is below $h(2) - h(0)$ exactly when $h(2) - h(1) > h(1) - h(0)$.

Proof sketch. Evaluating the condition at $k = 0$ with $B \sim \text{Bin}(n - 2, q_{lo})$ reduces it to $\Pr[B = m - 2] > \Pr[B = m - 1]$, and the binomial mode inequality gives the stated form, with no slack. ■

The range $1 \leq m \leq n$ is required. For $m > n$ the requirement is unmeetable, R_m is identically zero and no marginal increases, while the stated inequality still holds.

What we have not proved. The converse direction we would like is global: that if the *first* marginal does not increase then *no* marginal does, so that the threshold characterizes submodularity of R_m outright and not merely the behaviour of the first induction. We believe this but do not prove it. An exhaustive search in exact rational arithmetic over $n = 2$ to 18 , with q_{lo} and q_{hi} on a grid of twentieths and every m , found no instance in which a later marginal increases while the first does not. The obvious route, showing that $Pr[B_k = m - 2] - Pr[B_k = m - 1]$ is monotone in k , does not work, since it fails for $m = 1$. We record the global statement as Open Problem 6 and use only the first-induction version above. A caution for anyone checking this numerically: a

floating-point sweep reports spurious violations at probabilities near machine epsilon, and exact arithmetic is needed.

The weaker hypothesis $n q_{lo} < m < n q_{hi}$ is **not** sufficient. For $n = 12$, $q_{lo} = 0.55$, $q_{hi} = 0.95$, $m = 7$, the expected capable counts are 6.60 and 11.40, so m lies strictly between them, yet the marginals decrease from the first step and R_m is exactly submodular with $\gamma = 1$. The gap between $m > n q_{lo}$ and $m - 1 > (n - 1) q_{lo}$ is precisely where such cases live.

Illustrative computation (not a measurement). For $n = 24$, $m = 16$, $q_{lo} = 0.55$, $q_{hi} = 0.95$, the condition holds since $(n-1) q_{lo} + 1 = 13.65 < 16$. Indexing the marginal by the covered count before the induction, as the identity above does, the gain from going from k to $k + 1$ rises for k from 0 to 6, peaks at $k = 6$, and first falls at $k = 7$. The Das-Kempe ratio depends on the budget k it is minimised over, so we give it by budget instead of quoting a single number. Table 3 lists it:

BUDGET K	Γ	GREEDY GUARANTEE $1 - e^{-\Gamma}$
5	0.757	0.531
6	0.721	0.514
7	0.694	0.500
10 and above	0.665	0.485

The ratio saturates at 0.665, attained at $|L| = 0$ and $|S| = 10$. The submodular case has $\gamma = 1$ and a greedy guarantee of $1 - 1/e = 0.632$; this instance has $\gamma = 0.665$ and a guarantee of 0.485. The comparison to make is 0.485 against 0.632, since 0.665 and 0.632 are quantities of different kinds. These numbers are computed from the stated parameters and the stated budget, and are not observed on a fleet. Figure 4 plots the marginals for this instance.

Theorem 1 is vacuous here. In its discrete form the slope ratio over the full range is $\rho \approx 4.8 \times 10^{-5}$, because the tail probabili-

ty flattens almost completely at large k , against the full-range $\gamma = 0.665$. Restricting to the convex window $k \leq 6$ gives $\rho = 0.566$, against the $\gamma = 0.721$ for the matching budget. Theorem 1 is a general sufficient bound and should not be expected to be tight; on this instance direct computation of γ is the right tool.

The complementarity has a plain reading. When the capable count is far below m , one more capable asset barely changes the probability of meeting the requirement. Several more change it considerably. That is increasing returns, and it is a property of the requirement, not of the degradation process.

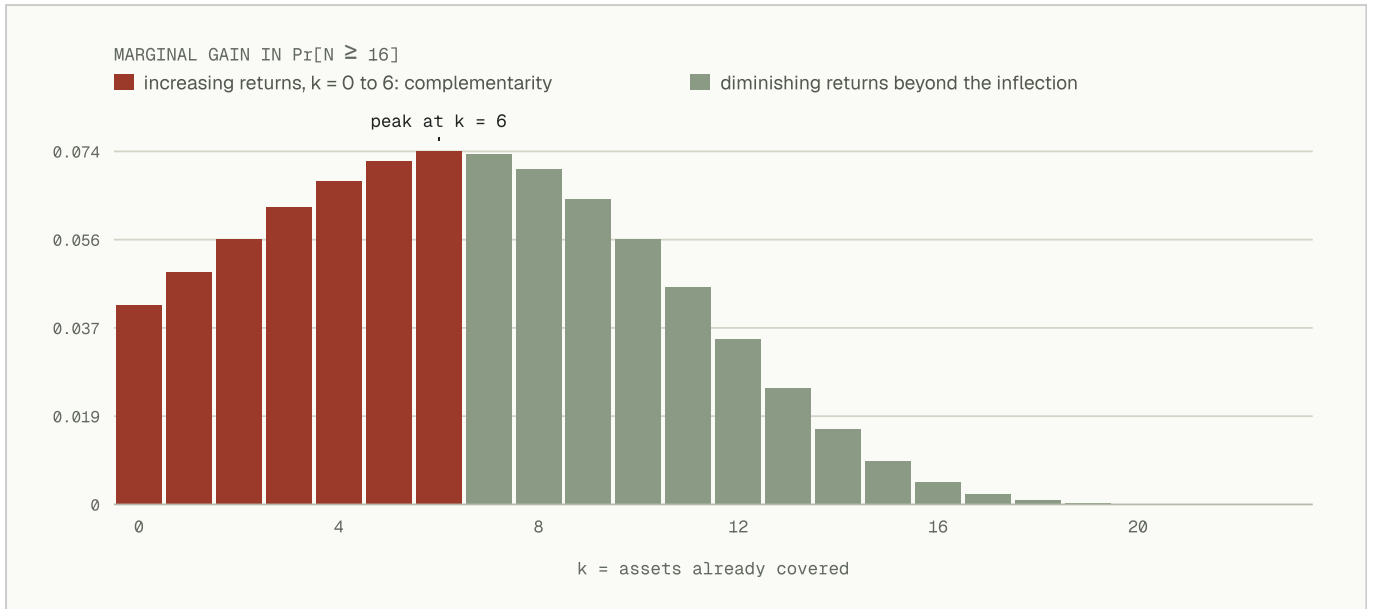


FIGURE 4 • ILLUSTRATIVE COMPUTATION

Where submodularity actually fails. Marginal gain in the probability that at least 16 of 24 assets are simultaneously capable, as each successive asset is upgraded from capability 0.55 to 0.95. The marginal rises through $k = 6$ before falling, so early inductions show increasing returns. That is complementarity, and it comes from the conjunctive form of the requirement rather than from the cascade. Computed from the stated parameters; not observed on a fleet.

5.7 Reading the result

Three statements, in order of how much they revise the starting intuition.

First, the cascade does not break the approximation guarantee. Under the feedback law of Section 3.4 readiness stays submodular on the stable branch at every admissible feedback strength, and the applicable constant depends only on which constraints bind. The alarming dynamic and the hard optimization problem are not the same thing. Proposition 3 bounds how far this gener-

alizes: it is the shape of the propensity curve, not the presence of feedback, that decides.

Second, the cascade makes the objective ill-conditioned, and ill-conditioning attacks the inputs. **The binding constraint near collapse is the quality of the estimated model, not the quality of the scheduler.**

Third, the genuine loss of submodularity comes from conjunctive mission requirements, is governed by an exact and checkable threshold, and would be present in a fleet with no coupling at all.

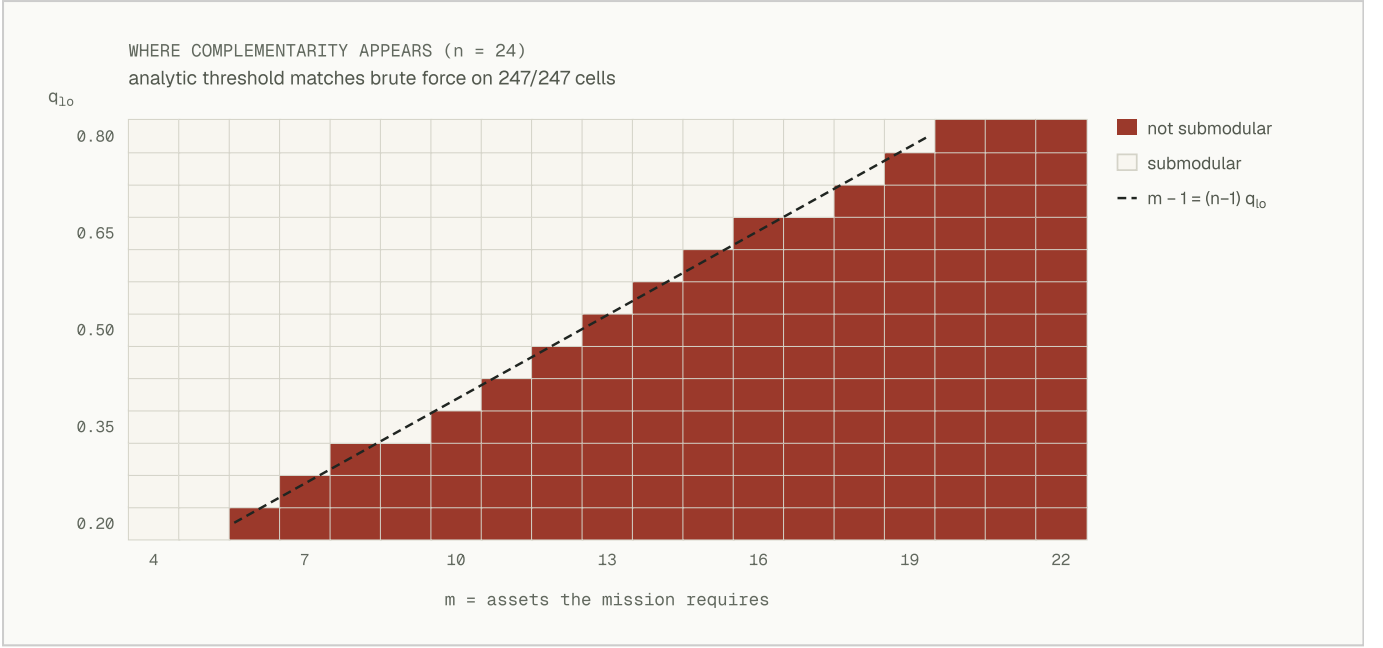


FIGURE 5 • ILLUSTRATIVE COMPUTATION

Proposition 2 as a map. Filled cells are parameter settings where some marginal increases, so the conjunctive objective is not submodular; in empty cells every marginal decreases. The check is exhaustive over all covered counts, not just the first induction. The dashed line marks equality in the analytic threshold $m - 1 > (n - 1) q_{10}$, which concerns the first induction alone, and it agrees with the exhaustive check on all 247 cells shown. The weaker condition that m lies between the expected capable counts is not sufficient, and the cells between the two conditions are exactly where that fails.

6. Algorithms

Theorem 2 settles the fractional case: the objective is submodular and the standard guarantees apply, with the constant determined by the binding constraints. The algorithmic question is therefore confined to the conjunctive objective of Section 5.6 and to the stiffness of Section 5.5.

Continuous greedy on the multilinear extension. Evaluating the multilinear extension requires the law of $Q(S)$ under independent rounding. We flag a limitation that is easy to state wrongly: when failure probabilities are heterogeneous, $Q(S) = \sum_i p_i Y_i$, with Y_i the indicator that asset i is protected, is a *weighted* sum whose law has up to 2^n atoms, and is **not** Poisson-binomial. Cheap evaluation via a one-dimensional integral is available only under homogeneous p_i , where Q reduces to a count, or through a discretization or central-limit approximation whose error would need to be controlled. Under the coverage structure of Section 3.5 the indicators are not independent in any case. We regard efficient evaluation of the multilinear extension for RCSS as open.

Target the bifurcation directly, and an argument we withdraw. If the practical obstruction is proximity to $\pi_{\max}$, an alternative objective is to maximize the margin $\pi_{\max} - \pi(S)$. Because $\pi(S) = (Q(G) - Q(S))/n$, the margin equals $\pi_{\max} - Q(G)/n + Q(S)/n$, an affine increasing function of $Q(S)$, so maximizing it is maximizing the coverage $Q(S)$, which is monotone submodular, so the same constants apply.

An earlier version of this paper claimed the margin objective is **better conditioned**, on the grounds that it removes the $\Delta^{-1/2}$ amplification from the optimization. That claim is wrong and we withdraw it. Since h is strictly increasing, $\arg\max h(Q(S)) = \arg\max Q(S)$ over any feasible family, and greedy selects the identical element at every step under either objective, because $h(Q(L) + \delta_a) - h(Q(L))$ is a strictly increasing function of δ_a with $Q(L)$ common to all candidates. The two objectives have the same exact optimum and the same greedy trajectory. Nothing about the conditioning of h reaches the combinatorial problem at all; the only thing it changes is floating-point arithmetic.

There is a real reason to prefer the margin, and it is a different one. The margin is well defined on the collapsed region where R is not, that is, past the fold where Theorem 2's hypothesis fails and h does not exist. That is precisely the stressed-fleet regime this paper is motivated by and currently excludes by assumption. An objective that still ranks schedules there is worth having for that reason alone.

The withdrawal sharpens the conditioning thesis rather than weakening it. Stiffness never was an optimization problem; it reaches the result only through the transfer of approximation ratios, which are not invariant under monotone transforms, and through estimation. Section 8.4 is where it lands.

Handle the conjunctive objective directly. For R_m , Proposition 2 identifies exactly when complementarity appears, and it appears at small k . A method that commits to blocks of inductions rather than single ones on the convex segment, then reverts to greedy once past the inflection, is the obvious response. We have not analyzed its guarantee.

6.1 Reference implementation

stead of taken on trust.

Every quantity in Sections 5 and 8 is a few lines of arithmetic. Listing 1 gives them in Python, so the claims can be checked in-

LISTING 1. PYTHON

```
1 import math
2
3 def pi_max(b0, kappa):
4     "Critical uncovered mass. The fold of (2). No fold exists when kappa = 0."
5     if kappa == 0:
6         return math.inf          # linear regime: never collapses
7     return (1 - b0) ** 2 / (4 * b0 * kappa)
8
9 def x_star(pi, b0, kappa):
10    "Stable disabled fraction: the smaller root of (2), or None past the fold."
11    if kappa == 0:
12        return pi / (1 - b0)      # the case Proposition 1 uses
13    disc = (1 - b0) ** 2 - 4 * b0 * kappa * pi
14    if disc < 0:
15        return None              # no healthy equilibrium
16    return ((1 - b0) - math.sqrt(disc)) / (2 * b0 * kappa)
17
18 def slope_ratio(pi_lo, pi_hi, b0, kappa):
19    "rho = inf h' / sup h' over [pi_lo, pi_hi]. Tends to 0 at the fold."
20    if pi_hi < pi_lo:
21        raise ValueError("pi_hi must be the larger endpoint")
22    if kappa == 0:
23        return 1.0               # h is affine, so the slopes agree
24    d = lambda pi: (1 - b0) ** 2 - 4 * b0 * kappa * pi
25    if d(pi_hi) <= 0:
26        raise ValueError("pi_hi is at or past the fold")
27    return math.sqrt(d(pi_hi) / d(pi_lo)) # lies in [0, 1] by construction
28
29 def sensitivity(pi, b0, kappa):
30    "|dR/dkappa| from Observation 1. Diverges as (pi_max - pi)^(-1/2)."
31    disc = (1 - b0) ** 2 - 4 * b0 * kappa * pi
32    if disc <= 0:
33        return math.inf          # at or past the fold
34    x = x_star(pi, b0, kappa)
35    return b0 * x * x / math.sqrt(disc)
36
37 def tail_at_least(ps, m):
38    "Pr[N >= m] for independent Bernoulli ps: exact Poisson-binomial."
39    if m <= 0:
40        return 1.0              # guard: dp[-1:] would be wrong
41    dp = [1.0] + [0.0] * len(ps)
42    for p in ps:
43        for k in range(len(dp) - 1, 0, -1):
44            dp[k] = dp[k] * (1 - p) + dp[k - 1] * p
45        dp[0] *= 1 - p
46    return sum(dp[m:])
47
48 def complementary(n, m, q_lo):
49    "Proposition 2: does the first induction show increasing returns?"
50    if not (n >= 2 and 1 <= m <= n):
51        raise ValueError("Proposition 2 assumes n >= 2 and 1 <= m <= n")
52    return (m - 1) > (n - 1) * q_lo
```

The last function, *complementary*, is the whole of Proposition 2. Checking it against a brute-force evaluation of *tail_at_least* over a parameter grid is the quickest way to confirm the threshold, and is what Figure 5 reports.

7. Quantum optimization: what would have to be true

7.1 What the hardware can and cannot do

The inner problem of assigning inducted assets to depot slots admits a standard quadratic unconstrained binary optimization (QUBO) encoding, and therefore an Ising form suitable for QAOA [16] or annealing hardware.

That does not mean quantum methods should be used. For $n = 200$ assets, $|D| = 5$ depots and $T = 12$ periods, the natural binary encoding has $200 \times 5 \times 12 = 12,000$ variables before routing variables are added. Two obstacles apply. Annealing hardware with enough qubits exists, but its connectivity is sparse. Mapping a densely coupled problem onto that sparse graph, a step called minor-embedding, spends several physical qubits per logical variable, and that cost grows quadratically in the logical problem size. The usable logical size is therefore far below the nominal qubit count. Gate-based devices have passed a thousand qubits, so raw count is no longer the binding limit there either; coherence time and two-qubit gate error are, and they cap the circuit depth that QAOA at this problem size would require. We state these as the structural obstacles rather than quoting a device specification, because specifications change and the obstacles have not.

We also note that a careful benchmarking study of quantum annealing on randomly generated instances found no evidence of speedup [17]. That study concerns one device and one instance class, and it explicitly leaves open the possibility of speedup on structured instances, which is the only possibility the next section entertains.

7.2 Why degeneracy is the thing to look for

The motivation must come from somewhere the paper has actually established, so we are explicit that it cannot come from the fractional model. Theorem 2 proves h is concave on the stable branch, strictly so for $\kappa > 0$, with no convex segment anywhere, and stiffness separates optima *better*, not worse, since a small change in Q produces a large change in R . Nothing about the fractional objective suggests a rugged landscape.

The conjunctive objective of Section 5.6 is different. On its convex segment the response is nearly flat in readiness terms, and many distinct schedules reach similar covered counts. A large number of configurations therefore sit inside a narrow band of objective value, separated only by the block structure Section 6 describes. Degeneracy of that kind, where near-optimal configurations are numerous and single-element moves do not distinguish them, is the structural feature that would motivate interest in quantum heuristics.

A caveat that may dissolve the motivation entirely. Section 5.6 assumes homogeneous failure probabilities and that each induction protects exactly one previously unprotected asset. Under precisely those assumptions $R_m(S)$ depends only on $|S|$, so every feasible schedule of a given size has the *identical* value.

That is not a narrow band, it is total degeneracy, and it makes the problem trivially easy rather than hard: maximizing a function of cardinality over an intersection of two matroids is a maximum-cardinality common independent set problem, solvable exactly in polynomial time by matroid intersection. The degeneracy that motivates Hypothesis 1 is therefore an artifact of the simplification introduced to obtain Proposition 2, not a property of RCSS. Restoring heterogeneous p_i and overlapping protection sets gives a genuinely hard problem, but then Proposition 2 no longer locates the complementarity and we no longer know where to look. Open Problem 5 asks whether the degeneracy survives that restoration. Until it is answered, Hypothesis 1 rests on nothing, and we would not spend hardware time on it.

Hypothesis 1. For the conjunctive objective R_m in the parameter regime identified by Proposition 2, the classical time-to-target for the inner assignment problem grows faster with instance size than it does outside that regime, and a quantum heuristic exhibits a smaller growth exponent.

7.3 The test that would refute it

Sweep instances across the Proposition 2 threshold, holding everything else fixed, so the only varying quantity is whether the convex segment exists. For each instance measure classical time-to-target, the wall-clock time to first reach a fixed target objective value, with at least two solvers of different type, for example a constraint-programming solver such as CP-SAT and a tuned simulated-annealing baseline, and record the scaling exponent as a function of instance size. Then run the same instances on available quantum hardware and on a classical simulator of the same algorithm, comparing exponents rather than wall-clock times, following the methodology in [17].

The hypothesis is refuted if the classical exponent does not increase across the threshold, which would mean the conjectured hard regime is not hard. It is also refuted if the exponent increases but the quantum exponent increases at least as fast. We consider the second outcome more likely than a positive result, and the first outcome the more interesting one, because it would simplify the algorithmic picture of Section 6. Reporting either negative outcome is the point of specifying the test.

8. The learned layer

Sections 3 through 7 treat W , p , and (β_0, κ) as given. They are not, and the estimation problem does not sit beside the theory. It is inside it. Figure 6 shows where each learned component sits relative to the solver.

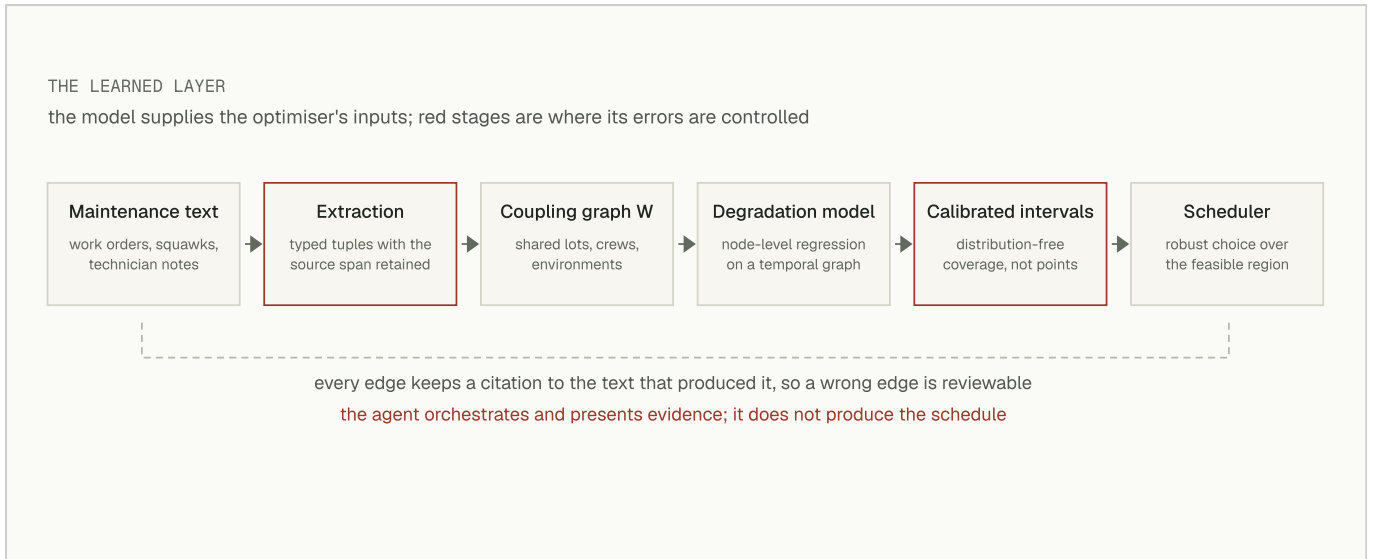


FIGURE 6 • CONCEPTUAL

Where the machine learning sits. The model does not choose the schedule; it supplies the coupling and degradation estimates the scheduler consumes, which is why Observation 1 makes calibration structural rather than optional. The two highlighted stages are where model error is controlled: extraction keeps a source span for every edge, and calibration replaces point predictions with intervals a robust formulation can use.

8.1 Language models build the graph

Most of what you need to populate W is written in prose, not recorded in telemetry, and the reason is specific: narratives carry coupling and sensors do not. A sensor channel reports that a temperature rose on one aircraft. A maintenance note reports that the same bracket cracks on every aircraft that operated from one location. Work orders, discrepancy reports (squawks), supply requisitions and technician free-text notes record what failed, what was done, and often what the technician believed caused it, at a specificity structured fields never reach.

The defensible use of a language model here is **extraction, not decision**. It parses narratives into typed tuples (component, failure mode, suspected cause, part lot, corrective action), resolves entities against the maintenance catalogue, and emits candidate edges for W . Every edge keeps the source span that produced it.

This matters for a specific reason. W appears in no equation in this paper: the analysis runs on the protection sets σ_a , the failure probabilities p , and the feedback parameters $(\beta o, \kappa)$. W is the

record from which those are estimated, and it reaches the result only through them. That makes the extraction step the first place an error can enter, and also the only place an error can be audited, because an extracted edge can be checked against the span that produced it. A wrong edge is then a reviewable claim, not an unexplained number. Two requirements follow and should be treated as hard: every extracted edge carries a citation to its source span, and extraction confidence is retained rather than thresholded away, for the reason given in 8.4.

8.2 Graph neural networks estimate degradation

Given the heterogeneous graph of assets, components, part lots and depots, remaining useful life (RUL) estimation is node-level regression over a temporal graph. The motivation is direct: components sharing a lot or an environment have correlated degradation, which is what a graph model represents and a per-asset model cannot.

The benchmark starting point is the C-MAPSS turbofan data [10], whose generation methodology is documented separately [11].

PROPERTY	C-MAPSS	WHAT THIS PAPER NEEDS
Source	simulated, NASA C-MAPSS	real maintenance records
Failure data	run-to-failure trajectories	run-to-failure, with lot identifiers
Units	independent by construction	correlated through shared lots
Covariates	sensor channels per unit	sensors plus narrative text
Validates	per-unit remaining life	the coupling record W behind σ and $(\beta o, \kappa)$

Those last two rows are why C-MAPSS is not enough. C-MAPSS can validate the per-unit prediction component and nothing more, because units that never interact cannot exhibit the

coupling this paper is about. We state that rather than let a reader infer a stronger claim from a familiar benchmark.

8.3 Policy confounding

Maintenance records are not an observational sample from the degradation process. They are filtered by the policy in force. Assets that were inducted did not fail, so the records systemati-

cally lack the counterfactual, and a model trained naively learns the policy as much as the physics. Its errors are then correlated with the decisions the optimizer is about to make. The record of which assets were inducted is a treatment assignment, and a model that consumes it as a feature has already gone wrong.

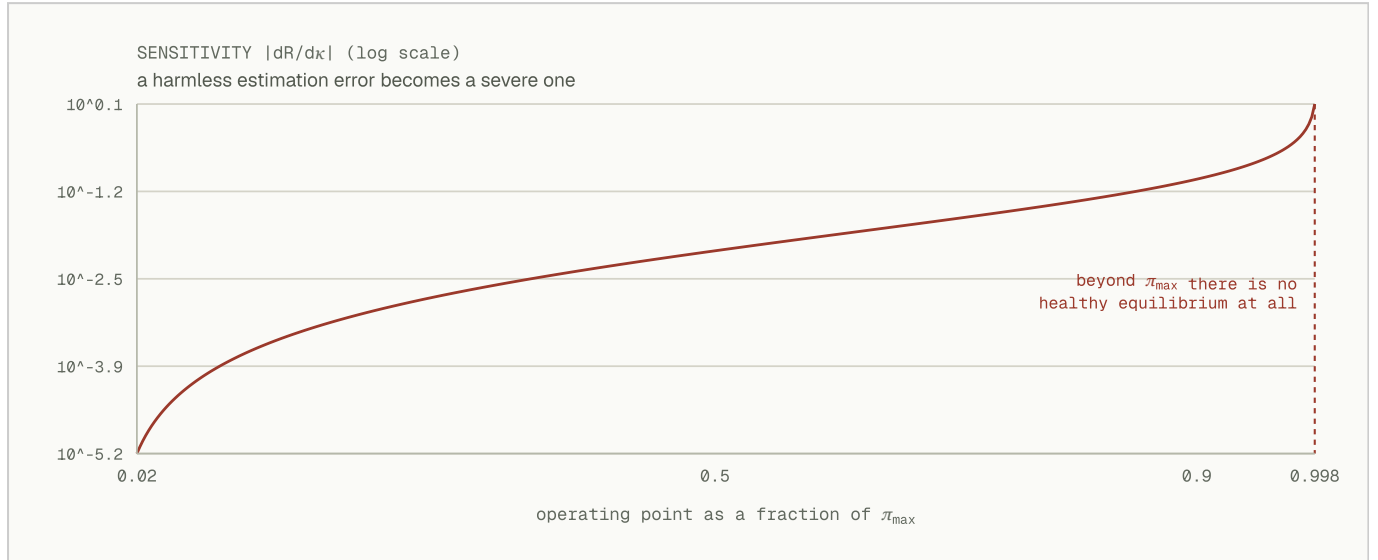


FIGURE 7 · ILLUSTRATIVE COMPUTATION

Observation 1, drawn. Sensitivity of predicted readiness to an error in the feedback strength, on a logarithmic scale, computed from the stated parameters. A relative error that is harmless for a healthy fleet moves the prediction by orders of magnitude near the fold, and beyond the fold at the right edge the healthy equilibrium does not exist at all, so the failure there is loss of solution rather than a bounded error. Two cautions on reading this curve. Most of the rise across the plotted window comes from the growth of the equilibrium disabled fraction itself, not from the inverse-square-root divergence, which dominates only in the last few percent before the fold; the exponent is a statement about the approach to the fold and is not measurable off this axis. And the right-hand marker sits at 0.998 of the critical mass, not at it.

8.4 Calibration is a structural requirement

Here the learned layer rejoins the theory. The quantity of interest is how much an error in κ moves predicted readiness. Differentiating (2) implicitly with respect to κ at fixed π :

$$dx^*/d\kappa = \beta_0 (x^*)^2 / \sqrt{\Delta}$$

where β_0 , baseline propensity for one failure to induce another in a healthy fleet; x^* , stable disabled fraction, the smaller root of equation (2); Δ , discriminant $(1 - \beta_0)^2 - 4\beta_0\kappa\pi$; it reaches zero exactly at the fold; κ , feedback strength: how much scarcity raises the chance of the next removal.

and since $\Delta = 4\beta_0\kappa(\pi_{\max} - \pi)$, the sensitivity $|dR/d\kappa| = \beta_0 (x^*)^2 / \sqrt{\Delta}$ grows as $(\pi_{\max} - \pi)^{-1/2}$. The same differentiation in β_0 gives $dx^*/d\beta_0 = x^*(1 + \kappa x^*) / \sqrt{\Delta}$, with the same exponent, so the statement below covers both parameters.

Observation 1. The sensitivity of predicted readiness to error in the feedback parameters (β_0 , κ) grows without bound as the margin to the bifurcation closes, with the same $(\pi_{\max} - \pi)^{-1/2}$ exponent that governs the response slope. At any fixed operating point short of the fold the sensitivity is finite and computable; the statement is about the rate at which it grows as the margin shrinks.

A second and sharper failure mode deserves separate mention. $\pi_{\max}$ itself depends on κ , so an error in $\hat{\kappa}$ does not merely scale the predicted readiness: it can place the true operating point beyond the fold, where the healthy equilibrium does not exist at all. Near collapse, a bad estimate does not give you a wrong answer. It gives you no answer. Figure 7 plots this sensitivity on a logarithmic scale, including the loss-of-solution regime past the fold.

This is where the corrected picture from Section 5 does its work. Theorem 2 removes optimization quality as the binding constraint in the fractional case. What remains is estimation quality, and Observation 1 says it degrades in the least convenient way: an error that is harmless for a healthy fleet is severe for a stressed one.

What Observation 1 does and does not imply. It is tempting to conclude that diverging sensitivity makes calibration a structural requirement, and an earlier version of this paper argued exactly that. The argument does not hold, and the reason is worth stating because it leads somewhere better. Since R is monotone in κ , a valid $1 - \alpha$ interval $[\kappa_{lo}, \kappa_{hi}]$ maps to $[R(\kappa_{hi}), R(\kappa_{lo})]$ with **exactly** the same coverage, however steep the map. Steepness destroys informativeness, not calibration. Worse for the original argument, the amplification can cancel outright: if κ is inferred from readiness observations with fixed noise σ , then the Fisher information is $(\partial R / \partial \kappa)^2 / \sigma^2$, so $\text{Var}(\hat{\kappa}) = \sigma^2 / (\partial R / \partial \kappa)^2$, and

the delta method returns $\text{Var}(\hat{R}) = \sigma^2$, independent of the margin. We have verified this numerically across operating points from $0.5 \pi_{\max}$ to $0.999 \pi_{\max}$, where $|dR/d\kappa|$ grows by a factor of 243 while the induced prediction standard deviation stays flat.

The correct statement is about **precision and sample size**, and it is stronger than the one we withdrew. To hold a target error δR fixed as the fleet approaches the fold, the parameter error must shrink as $\delta\kappa \propto (\pi_{\max} - \pi)^{1/2}$, so the sample size required grows as $(\pi_{\max} - \pi)^{-1}$. Observation 1 bites in practice when κ is estimated from a source whose informativeness does **not** scale with $\partial R/\partial\kappa$, for example cannibalization counts rather than readiness itself, or when a model fitted on a healthy fleet is extrapolated to a stressed one. Both are the realistic cases, and both should be stated as the condition rather than assumed.

The margin is estimable from readiness data alone. A claim in an earlier version, that the amplification is a property of the equations which the data cannot reveal, is false in our own model and we withdraw it. The relaxation rate on the stable branch is $\lambda = \text{sqrt}(\Delta)$, so the correlation time of readiness fluctuations grows as $(\pi_{\max} - \pi)^{-1/2}$ and the stationary variance with it. This is critical slowing down, the standard early-warning apparatus for an approaching fold [20]. The variance and lag-one autocorrelation of a readiness time series are therefore margin estimators carrying the same exponent, and they require no estimate of κ , no coupling matrix, and no language model. That is a cheaper and more direct route to the quantity that matters than estimating the parameters and inferring the margin from them, and it is testable on readiness data that operators already hold. We regard establishing its exponent in a discrete-event simulation as the single most informative experiment in Section 9.

Two consequences for practice. The margin to $\pi_{\max}$ is itself the quantity most worth estimating accurately, more than any individual asset's remaining life, and standard prognostics evaluation does not measure it. And the learned layer should emit calibrated intervals, not point predictions, so that a robust formulation can select against the worst case inside the interval. One caution about the tool: conformal prediction supplies distribution-free coverage for **future observations** under exchangeability [12], [13], whereas selecting against the worst case in $(\beta o, \kappa)$ requires a confidence region for **parameters**, which conformal methods do not provide. Likelihood, bootstrap or Bayesian inference on $(\beta o, \kappa)$ is the right instrument there, and conformal prediction remains appropriate for the per-asset failure probabilities p . Exchangeability fails under the confounding of 8.3, so the coverage guarantee weakens there too, and that weakening should be stated and not assumed away.

8.5 Agents orchestrate; they do not optimize

The appropriate role for an agentic system here is narrow, and worth stating because the tempting role is wrong.

Appropriate:

- Monitor the estimated margin to $\pi_{\max}$, and raise an alert when it narrows.
- Assemble the evidence behind a proposed plan, including the source spans that justify the coupling edges it relies on.
- Call the solver and present the tradeoffs between candidate schedules.
- Identify which assumption a recommendation is most sensitive to.
- Route the decision to a human with the authority to make it.

Not appropriate: producing the schedule. A language model asked to generate a maintenance plan will produce a fluent and plausible one. It will not respect depot capacity, it will not respect routing feasibility, and it has no access to the equilibrium of (1). The output will be well formed and wrong, which in a safety-relevant setting is worse than no output because it is harder to reject.

The boundary is that the agent owns evidence and orchestration and the solver owns the schedule. It should be enforced in the system design, not left to prompting.

9. Proposed experimental protocol

This protocol has not been executed. It is specified so that it can be.

Phase 1: response curve recovery (simulation). Build a discrete-event sustainment simulator implementing Section 3.4 with explicit inventory and cannibalization. Sweep κ within the physical range of 5.4 and measure the empirical response and the location of $\pi_{\max}$. *Success criterion:* the measured $dx^*/d\pi$ exhibits the $(\pi_{\max} - \pi)^{-1/2}$ divergence of 5.4(b). A different exponent would indicate the bifurcation is not of saddle-node type and that Section 5.4 needs revision.

Phase 2: guarantee behavior (simulation). Compute the empirical submodularity ratio by sampling the defining pairs, and compare greedy against exact optima on instances small enough to solve exactly. *Success criterion:* under fractional readiness, greedy stays within the constant appropriate to the binding constraint at every κ , confirming that stiffness does not manifest as an optimality gap. Under R_m , the measured γ falls below one exactly when Proposition 2's threshold is crossed. Theorem 2 is a proved statement about the model, so a contrary simulation result would indicate that the simulator does not implement the model of Section 3.4, or a coding error, rather than falsifying the theorem.

Phase 3: sensitivity (simulation). Perturb κ by a fixed relative amount and measure the change in the chosen schedule's value as a function of margin. *Success criterion:* the $(\pi_{\max} - \pi)^{-1/2}$ growth of Observation 1, and the appearance of the loss-of-solution regime. This phase tests the paper's main practical claim and should run before any learning work.

Phase 4: per-unit RUL (real data). Train and evaluate graph and non-graph RUL models on C-MAPSS [10], [11] with the standard split. *This phase validates only the per-unit prediction component*, for the reason given in 8.2.

Phase 5: coupling extraction (requires data we do not have). Validating W requires maintenance narratives with known shared-lot structure. This is the binding constraint on the program. Public corpora of this kind are scarce; the realistic paths are a partnership with an operator under a data agreement, or a synthetic narrative corpus with known ground-truth coupling, which tests the extraction pipeline but not its behavior on real prose.

Phase 6: the quantum sweep. As specified in 7.3, across the Proposition 2 threshold.

Phase 5 is stated as the limiting step deliberately. A reader should know which part of the program is blocked and why.

10. Limitations

The feedback model is a choice, and the headline result depends on it. $\beta(u) = \beta_0(1 + \kappa(1 - u))$ is the simplest parameterization producing the behavior. The square-root scaling of 5.4(b) is generic to saddle-node bifurcations and robust to the form, and π_{max} and the constants are not. More importantly, neither is the **concavity**: Proposition 3 gives the exact criterion, and Section 5.5 exhibits a strictly subcritical saturating propensity, with no fold anywhere, under which readiness is supermodular. An earlier version of this paper presented the submodularity as a fact about cannibalization. It is a fact about a linear propensity curve.

The cascade requires a mission-capable donor, and the model does not enforce it. In (1) a cannibalization event always dis-

ables a new asset. In a finite fleet a removal needs a capable donor and a target that is not already down, which multiplies both the primary and the induced terms by a factor of the capable fraction. Writing that correction as $(\pi + \beta_0(1 + \kappa x)x)(1 - x) = x$, a fold exists only when $\beta_0 + 2\beta_0\kappa x$ exceeds $1/(1 - x)^2$ somewhere in $(0, 1)$. At $\beta_0 = 0.45$ and $\kappa = 2.2$, the parameters of every computed figure in this paper, it does not: the corrected response is monotone with no fold at all, and the critical feedback strength rises to about 5.87, roughly ten times the lower end of the window above. The existence of the tipping point is therefore an assumption of the stated model, not a consequence of cannibalization, and we do not claim otherwise. The scaling results are unaffected where a fold does exist. Settling which form is right is an empirical question and the first thing Section 9 should answer.

The healthy branch is metastable in a finite fleet. The deterministic flow treats the stable root as an attractor, but demographic noise at finite n allows escape over the unstable root, so the operational tipping point arrives before π_{max} and the deterministic fold location is an optimistic bound. The width of the smeared region shrinks only as a fractional power of n , which at realistic fleet sizes places the operating points discussed in 5.4 inside it. A simulation at a few hundred assets should be expected to show a rounded transition, and that would confirm the finite-size physics rather than contradict the model.

Theorem 2 holds on a bounded window of feedback strengths. Two conditions bracket κ from opposite sides. The fold is reachable with $x^* \in [0, 1]$ only for $\kappa \geq (1 - \beta_0)/(2\beta_0)$; below that the bifurcation belongs to the equations and not to any fleet. And Theorem 2 needs $\kappa < n(1 - \beta_0)^2 / (4\beta_0 Q(G))$, or the empty schedule already sits past the fold. Together:

$$(1 - \beta_0) / (2\beta_0) \leq \kappa < n(1 - \beta_0)^2 / (4\beta_0 Q(G))$$

where β_0 , baseline propensity for one failure to induce another in a healthy fleet; κ , feedback strength: how much scarcity raises the chance of the next removal; $Q(S)$, covered failure mass; monotone, normalized and submodular; n , fleet size.

The window is non-empty exactly when $Q(G) < n(1 - \beta_0)/2$, that is when the average primary failure probability is below half the baseline survival margin. A fleet failing faster than that is past the fold with nothing scheduled. The window also narrows as κ grows, so the theorem's reach shrinks exactly as the mechanism it describes gets stronger: at $\beta_0 = 0.45$ and $\kappa = 2.2$ it requires a mean primary failure probability below 0.076, which is a lightly loaded fleet. We give no worked parameter set that satisfies every constraint in this paper simultaneously, and one should be supplied.

Branching approximates a finite system. Galton-Watson assumes unbounded population and independent offspring. A fleet is finite and cannibalization targets are not independent draws. The approximation is weakest exactly at the critical point.

Proposition 2 assumes homogeneity and independence. It requires homogeneous failure probabilities so coverage reduces to a count, and conditional independence of asset capability. Coupling violates the second. We expect the S-shape to survive because it comes from the threshold and not from independence, and we have not proved it under dependence.

Theorem 1 can be arbitrarily loose. Section 5.6 exhibits an instance where $\rho \approx 4.8 \times 10^{-5}$ against a true $\gamma = 0.665$. The bound is a general sufficient condition, not a useful estimate.

Lemma 1 is one-directional. Concavity is sufficient for submodularity, not necessary. We do not characterize the non-concave responses that still yield submodular objectives.

Efficient evaluation of the multilinear extension is unresolved. See Section 6.

No measured fleet results. Every number here is a model parameter or a derived quantity.

11. Open problems

1. **How should a fixed budget be split between estimation and optimization?** Theorem 2 says a better scheduler buys nothing in the fractional case, and Observation 1 says better coupling estimates buy more the closer the fleet is to $\pi_{\max}$. An allocation rule as a function of margin would be directly useful and follows from neither literature alone. This question follows from the corrected result, not from the one we originally anticipated.
2. **A guarantee for block-greedy on the conjunctive objective.** Proposition 2 localizes the complementarity to small k . Whether a block-then-greedy method beats $1 - e^{-\gamma}$ there is open.
3. **Proposition 2 under dependence.** Extend the threshold to coupled capability, where the Poisson-binomial structure is

lost.

4. **Multilinear evaluation for weighted coverage.** Either an efficient scheme for heterogeneous p_i or a hardness result.
5. **Does the degeneracy of 7.2 appear?** Answerable by simulation alone, before any quantum hardware is involved, and it should be settled first.
6. **Does the first-induction threshold characterize submodularity globally?** Proposition 2 proves the condition for the first induction. Whether the first marginal failing to increase implies that no marginal increases is verified exhaustively for small fleets in exact arithmetic but not proved, and the obvious monotonicity route fails.
7. **Which feedback laws preserve the guarantee, among those a fleet actually exhibits?** Proposition 3 gives the mathematical criterion, that $x^2 b'(x)$ be nondecreasing. Which propensity curves satisfy it empirically is a question about cannibalization behaviour, not about mathematics, and it decides whether Theorem 2 describes anything real.

References

- [1] G. L. Nemhauser, L. A. Wolsey, M. L. Fisher. An analysis of approximations for maximizing submodular set functions I. *Mathematical Programming*, 14(1):265–294, 1978. doi.org/10.1007/BF01588971
- [2] M. L. Fisher, G. L. Nemhauser, L. A. Wolsey. An analysis of approximations for maximizing submodular set functions II. *Mathematical Programming Study*, 8:73–87, 1978. doi.org/10.1007/BFb0121195
- [3] G. Călinescu, C. Chekuri, M. Pál, J. Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 40(6):1740–1766, 2011. doi.org/10.1137/080733991
- [4] U. Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998. doi.org/10.1145/285055.285059
- [5] A. Das, D. Kempe. Approximate submodularity and its applications: subset selection, sparse approximation and dictionary selection. *Journal of Machine Learning Research*, 19(3):1–34, 2018. Earlier version: Submodular meets spectral: greedy algorithms for subset selection, sparse approximation and dictionary selection. *ICML*, 2011; arXiv:1102.3975. jmlr.org/papers/v19/16-534.html
- [6] A. E. Motter, Y.-C. Lai. Cascade-based attacks on complex networks. *Physical Review E*, 66(6):065102(R), 2002. doi.org/10.1103/PhysRevE.66.065102
- [7] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, S. Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1028, 2010. doi.org/10.1038/nature08932
- [8] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, D. J. Watts. Network robustness and fragility: percolation on random graphs. *Physical Review Letters*, 85(25):5468–5471, 2000. doi.org/10.1103/PhysRevLett.85.5468
- [9] C. C. Sherbrooke. METRIC: a multi-echelon technique for recoverable item control. *Operations Research*, 16(1):122–141, 1968. doi.org/10.1287/opre.16.1.122
- [10] A. Saxena, K. Goebel. Turbofan engine degradation simulation data set. *NASA Prognostics Data Repository*, NASA Ames Research Center, 2008. www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoc/pcoc-data-set-repository/
- [11] A. Saxena, K. Goebel, D. Simon, N. Eklund. Damage propagation modeling for aircraft engine run-to-failure simulation. *International Conference on Prognostics and Health Management (PHM)*, 1–9, 2008. doi.org/10.1109/PHM.2008.4711414
- [12] V. Vovk, A. Gammerman, G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi.org/10.1007/b106715
- [13] A. N. Angelopoulos, S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511, 2021. arxiv.org/abs/2107.07511
- [14] U.S. General Accounting Office (now the Government Accountability Office). *Military aircraft: services need strategies to reduce cannibalizations*. GAO-02-86, 2001. www.gao.gov/products/gao-02-86
- [15] K. B. Athreya, P. E. Ney. *Branching Processes*. Springer, 1972. doi.org/10.1007/978-3-642-65371-1
- [16] E. Farhi, J. Goldstone, S. Gutmann. A quantum approximate optimization algorithm. arXiv:1411.4028, 2014. arxiv.org/abs/1411.4028
- [17] T. F. Rønnow, Z. Wang, J. Job, S. Boixo, S. V. Isakov, D. Wecker, J. M. Martinis, D. A. Lidar, M. Troyer. Defining and detecting quantum speedup. *Science*, 345(6195):420–424, 2014. doi.org/10.1126/science.1252319
- [18] H. Lin, J. Bilmes. A class of submodular functions for document summarization. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 510–520, 2011. aclanthology.org/P11-1052/
- [19] J. Lee, M. Sviridenko, J. Vondrák. Submodular maximization over multiple matroids via generalized exchange properties. *Mathematics of Operations Research*, 35(4):795–806, 2010. doi.org/10.1287/moor.1100.0463
- [20] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. van Nes, M. Rietkerk, G. Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009. doi.org/10.1038/nature08227

- [21] I. Dobson, B. A. Carreras, V. E. Lynch, D. E. Newman. Complex systems analysis of series of blackouts: cascading failure, critical points, and self-organization. *Chaos*, 17(2):026103, 2007.
doi.org/10.1063/1.2737822