Cybersecurity at Scale

Why Risk, Coordination, and Leadership Stop Growing Linearly



A practitioner's field paper for studying growth, complexity, operating pressure, and cyber risk.

Carlos Morales

Version 4.0 | November 2026

Abstract

Cybersecurity programs are often planned as though growth were linear. A larger organization is expected to need proportionally more controls, analysts, governance, and technology. That assumption becomes unreliable when growth adds relationships as well as components. Every identity can create permissions. Every application can create integrations and data flows. Every supplier can create technical and operational dependencies. Every additional team can increase the coordination needed to make a decision. The environment becomes more connected, and the work or exposure created by those connections may grow faster than the number of things being counted.

This paper adapts the language of scaling laws to cybersecurity, cyber risk management, leadership, and organizational behavior. It introduces a practical model that separates four layers: scale inputs, relationship structure, operating demand and sustainable capacity, and risk consequences. The separation keeps practitioners from treating entitlement growth, alert volume, control workload, and business loss as though they were the same variable. They are related, and each link still requires evidence.

Version 4.0 brings together the cited analytical framework and an expanded identity example covering legacy systems, non-human identities, and vendor AI connections. It clarifies sustainable capacity, separates conditional scaling illustrations from forecasts, and makes the synthetic case's reproducibility limits explicit. The case illustrates analysis, model comparison, intervention planning, and verification design; it does not report validated organizational findings.

The paper remains deliberately reserved. It does not claim a universal cybersecurity exponent or turn a fitted curve into a risk decision. Its purpose is to help practitioners see where growth changes the system, where an operating model will begin to strain, and which redesign can change that relationship.

Practitioner Summary

The central problem is not size alone. It is the growth of relationships, dependencies, and possible failure paths compared with the organization's ability to understand and govern them. Scale tells us how much of something exists. Structure tells us what those things can do together. Cyber risk emerges from both, along with threats, controls, business consequence, response, and recovery.

The basic scaling relationship is compact:

$$Y(N) = Y_0 \left(\frac{N}{N_0}\right)^{\beta}$$

Here, N is a defined measure of scale, such as identities, applications, assets, suppliers, external connections, agents, or incident workstreams. Y is the outcome being studied. The exponent β describes how responsive that outcome is to changes in scale across the range where the relationship fits.

Scaling regime	What happens to total Y	What happens to Y per unit of N
$\beta < 1$	Y grows more slowly than N	The amount per unit falls
$\beta = 1$	Y grows in proportion to N	The amount per unit remains stable
$\beta > 1$	Y grows faster than N	The amount per unit rises

Cybersecurity can contain all three regimes at the same time. Common platforms, standardized services, and automation can create sublinear economies. Basic inventory obligations may remain close to linear. Permissions, integrations, coordination demands, shared dependencies, and candidate paths may grow superlinearly. None of those exponents should be borrowed from biology or cities and presented as a cybersecurity fact. They must be estimated for a defined variable, population, period, and operating model.

The operating question is whether security demand grows faster than sustainable capacity:

$$D(N) = D_0 \left(\frac{N}{N_0}\right)^{\beta_D}$$

$$K(N) = K_0 \left(\frac{N}{N_0}\right)^{\beta_K}$$

$$\rho(N) = \frac{D(N)}{K(N)} = \rho_0 \left(\frac{N}{N_0}\right)^{\beta_D - \beta_K}$$

D is incoming work in quality-adjusted units per period. K is the maximum sustainable service capacity under the current operating model, expressed in the same units and period. It is not the number of cases that happened to be completed. The ratio ρ is operating pressure. If β_D is larger than β_K, pressure rises as the environment grows. Added staff, external services, or tools may improve the baseline. Lasting relief usually requires structural changes such as simplification, modularity,

reusable controls, better information, automation, reduced relationship density, and distributed authority.

The framework also sharpens risk analysis. Workload, exposure, and risk influence one another, although they are not substitutes. A superlinear increase in entitlement edges does not prove an equal increase in expected loss. A sublinear increase in incident count does not prove that risk is improving. Architecture, reachability, privilege, control strength, threat conditions, business dependence, and recovery still determine what the structure means.

For a practitioner, the test is simple: the analysis must improve a decision. It should identify which variable is driving demand, where relationship density is increasing, where a process changes regime, whether a shared capability creates a genuine economy, where concentration has become fragile, and which intervention should change the result. A clean power law can be useful. A breakpoint, a residual pattern, or evidence that a linear model performs better can be equally useful.

Table of Contents

ABSTRACT 2

PRACTITIONER SUMMARY 3

INTRODUCTION 9

WHY SCALE DESERVES ATTENTION 9
WHAT SCALE MEANS IN CYBERSECURITY 10
COMPONENTS, RELATIONSHIPS, AND STRUCTURE 10
LINEAR, SUBLINEAR, AND SUPERLINEAR GROWTH 11
WHAT A POWER LAW TELLS US 11
A PLAIN-LANGUAGE GUIDE TO THE SYMBOLS 12
THE FOUR LAYERS OF THE SCALING PROBLEM 14
DEMAND, SUSTAINABLE CAPACITY, PRESSURE, BACKLOG, AND LATENCY 15
WORKLOAD, EXPOSURE, AND RISK ARE DIFFERENT 16
THE DATA LANGUAGE USED IN THIS PAPER 17
WHAT THE EVIDENCE CAN SUPPORT 18
COMMON MISTAKES TO AVOID 19
A SHORT IDENTITY EXAMPLE 20
WHAT CAN BE DONE WITHOUT QUANTITATIVE ANALYSIS 21
WHEN QUANTITATIVE SUPPORT BECOMES NECESSARY 22
WHAT THIS PAPER CAN AND CANNOT PROVIDE 22
HOW TO USE THIS PAPER 22

1. FROM HEARTBEATS TO CYBER SYSTEMS 24

2. SCALING AND POWER LAWS IN PLAIN LANGUAGE 25

2.1 BEGIN BY DEFINING SCALE 25
2.2 THE BASIC POWER LAW 25
2.3 THE DOUBLING RULE 26
2.4 TOTAL, PER-UNIT, AND SCALE-ADJUSTED VIEWS 27
2.5 A SCALING RELATIONSHIP IS DIFFERENT FROM A POWER-LAW DISTRIBUTION 27
2.6 RELATIONSHIPS AND THE COORDINATION CEILING 28

3. WHAT OTHER FIELDS CAN TEACH US 29

3.1 BIOLOGICAL NETWORKS: EFFICIENCY THROUGH DISTRIBUTION 29
3.2 CITIES: ECONOMIES AND ACCELERATION TOGETHER 29
3.3 HUMAN ATTENTION AND SOCIAL LAYERS 29

4. WHY ORGANIZATIONS CHANGE AS THEY GROW 30

5. THE CYBERSECURITY SCALING MODEL 31

5.1 Four linked layers 31
5.2 Demand and sustainable effective capacity 31
5.3 Pressure, backlog, and latency 33
5.4 Estimating an operating limit 34
5.5 What an intervention changes 35
5.6 A model that can fail usefully 37

6. APPLYING THE MODEL ACROSS CYBERSECURITY 38

6.1 Attack surface and asset management 38
6.2 Identity and access 38
6.3 Third-party and supply-chain risk 40
6.4 Cloud, software, and artificial intelligence 40
6.5 Operational technology and critical infrastructure 41
6.6 Incident response and crisis coordination 42
6.7 Vulnerability and exposure management 42
6.8 Security operations and alerting 42
6.9 Phishing and human-targeted attacks 43

7. CYBER RISK AT SCALE 44

7.1 Workload, exposure, and risk are different variables 44
7.2 Aggregation, dependence, and common causes 44
7.3 Concentration can create efficiency and fragility 45
7.4 Attack paths require semantics 46
7.5 Integrating scaling with cyber risk quantification 46

8. LEADERSHIP AND ORGANIZATIONAL BEHAVIOR 47

8.1 Direct leadership does not scale automatically 47
8.2 Distributed authority changes effective capacity 47
8.3 Trust is a network property 47
8.4 Culture and repeated local signals 48
8.5 Feedback loops make adaptation possible 48

9. HOW TO RUN A SCALING REVIEW 49

9.1 Step 1: state the decision 49
9.2 Step 2: draw the four-layer chain 49
9.3 Step 3: define *N* and *Y* 49
9.4 Step 4: measure relationships before compressing them 50
9.5 Step 5: measure demand and sustainable capacity 50
9.6 Step 6: inspect the data 50
9.7 Step 7: estimate and challenge 50
9.8 Step 8: connect the intervention to a model term 50
9.9 Step 9: decide how the result will be verified 51

10. MEASURING SCALING IN A REAL ORGANIZATION 53

10.1 START WITH A NARROW HYPOTHESIS 53
10.2 BUILD A PANEL INSTEAD OF RELYING ON ONE SNAPSHOT 53
10.3 BUILD A DATA DICTIONARY BEFORE A MODEL 53
10.4 INSPECT BEFORE ESTIMATING 54
10.5 ESTIMATE WITH THE ASSUMPTIONS VISIBLE 54
10.6 ACCOUNT FOR DEPENDENCE, HIERARCHY, AND MEASUREMENT ERROR 55
10.7 CHALLENGE THE FUNCTIONAL FORM 55
10.8 TREAT DEVIATIONS AS INFORMATION 55
10.9 EVALUATE INTERVENTIONS WITH A DESIGN THAT CAN SUPPORT THE CLAIM 56
10.10 KNOW WHEN TO STOP 56

11. A WORKED IDENTITY CASE 57

11.1 THE DECISION 57
11.2 THE DATA 57
11.3 FIRST INSPECTION 58
11.4 MODEL COMPARISON 58
11.5 DEMAND, CAPACITY, AND LOCAL PRESSURE 59
11.6 THE INTERVENTION PLAN 60
11.7 VERIFICATION 60

12. LIMITS AND RESPONSIBLE USE 62

13. BUILDING A SCALING CAPABILITY 63

13.1 STAGE ONE: A STRUCTURED REVIEW 63
13.2 STAGE TWO: DESCRIPTIVE ANALYSIS 63
13.3 STAGE THREE: MODEL COMPARISON 63
13.4 STAGE FOUR: INTERVENTION LEARNING 63
13.5 STAGE FIVE: PORTFOLIO AND CROSS-ORGANIZATION RESEARCH 63
13.6 TOOLING 63

14. RESEARCH AGENDA 64

CONCLUSION 65

APPENDIX A. WORKED MATHEMATICAL EXAMPLES 66
A.1 DOUBLING UNDER DIFFERENT EXPONENTS 66
A.2 ESTIMATING AN EXPONENT FROM TWO POINTS 66
A.3 TENFOLD GROWTH AND THE SCALING GAP 66
A.4 ESTIMATING THE OPERATING LIMIT 66
A.5 CONVERTING A SCALE-ADJUSTED RESIDUAL 66
A.6 RELATIONSHIP GROWTH 66
A.7 CANDIDATE ATTACK PATHS 67

A.8 CORRELATED LOSS 67

APPENDIX B. CANDIDATE METRICS 68
APPENDIX C. ONE-PAGE SCALING REVIEW CANVAS 70
APPENDIX D. ANALYSIS AND VALIDATION CHECKLIST 71
REFERENCES 72
ABOUT THE AUTHOR 74

Introduction

This introduction is written for cybersecurity practitioners who have little or no experience with scaling analysis, power laws, network concepts, queue behavior, or statistical modeling. Readers already comfortable with those ideas can move directly to Section 1. Everyone else can use these pages as a foundation for understanding what the paper is trying to measure, why the mathematical language appears, and how the framework can support practical decisions.

The paper asks a deceptively simple question: what changes when a cybersecurity system becomes larger? Many plans assume that growth is proportional. If the organization adds 20 percent more employees, applications, suppliers, or locations, the security program is expected to need about 20 percent more effort. That assumption can be reasonable for some activities. It can fail when growth adds relationships, dependencies, coordination, exceptions, and shared points of failure faster than it adds the components being counted.

The purpose of this paper is to help a practitioner detect that change before it appears only as backlog, delayed decisions, control exceptions, unmanageable privileges, incident confusion, or concentrated risk. The method begins with operational questions and plain-language definitions. Mathematics is introduced later because it can describe the relationship more precisely, compare possible explanations, and test whether an intervention changed the way the system behaves.

> *You do not need prior experience with power laws or cyber risk quantification to use the core framework in this paper. Begin with a decision, define what is growing, map the relationships it creates, compare operating demand with sustainable capacity, and connect the result to a risk scenario. Statistical analysis becomes necessary when the organization wants to estimate the relationship, compare competing models, forecast future demand, or support a causal claim.*

Why scale deserves attention

Security teams already work with size. They count endpoints, identities, vulnerabilities, alerts, suppliers, applications, cloud resources, policies, assessments, and incidents. Those counts are useful for inventory, coverage, workload planning, and reporting. The difficulty is that a count of things does not automatically describe what those things can do together.

Consider an application portfolio. Adding one application increases the inventory by one. It may also add several identity integrations, service accounts, APIs, data flows, supplier connections, logging requirements, recovery dependencies, and approval paths. The security work created by that application depends on those relationships. The business exposure also depends on where the relationships lead, what privileges they carry, which controls protect them, and what happens if they fail or are abused.

This is why a large environment can behave differently from a small environment. A process that works well for 50 applications may struggle at 500 even after the organization adds people and tools. The number of unique cases can grow. More teams may need to coordinate. Shared services may become

bottlenecks. Exceptions can accumulate faster than they are retired. A platform that creates efficiency can also become a common dependency for many important services.

Scale analysis gives the practitioner a disciplined way to ask whether growth is simply adding volume or changing the structure of the system. That distinction affects capacity plans, architecture, automation, governance, sourcing, resilience, and the timing of investment.

What scale means in cybersecurity

In ordinary conversation, scale may refer to company size, global reach, transaction volume, or technical complexity. An analysis needs a narrower definition. Scale is the measurable quantity whose growth may explain a change in an outcome. The right quantity depends on the decision.

Employee count may be useful when studying workforce identity processes. Application count may be more useful when studying integration demand. Supplier count can explain part of assessment volume, while supplier-to-service relationships may explain concentration. Endpoint count can help describe scanning volume, while reachable attack paths may be more informative for exposure. Incident participant count may matter less than the number of workstreams, affected services, jurisdictions, and decisions that must be coordinated.

Choosing scale is therefore part of the hypothesis. A statement such as "cybersecurity gets harder as organizations grow" cannot be tested in a useful way. A statement such as "monthly manual access-exception hours increase as active workforce identities and connected applications grow" defines something that can be observed, challenged, and connected to a decision.

Every scale measure also needs a boundary. An identity count may include employees only, or it may include contractors, customers, service accounts, robots, and AI agents. An application count may include software-as-a-service products, scripts, low-code workflows, and operational technology. A supplier count may stop at contracted providers or include recurring fourth parties. Changing the boundary changes the number and can change the relationship. Stable definitions are often more valuable than sophisticated calculations built on inconsistent data.

Components, relationships, and structure

The paper frequently distinguishes nodes from edges. A node is a component in a network, such as an identity, application, device, supplier, dataset, cloud service, team, or business process. An edge is a relationship between components, such as a permission, network route, API call, data transfer, trust, administrative path, supplier dependency, approval, or handoff.

Two environments can contain the same number of nodes and have very different security properties. One may use standard roles, bounded trust zones, reusable interfaces, and a small number of well-governed dependencies. Another may contain broad privilege, custom integrations, unmanaged service accounts, overlapping suppliers, and many paths between low-trust and high-consequence systems. The inventory count is the same. The structure is not.

Relationship count alone is still incomplete. Some relationships carry little consequence. Others provide privileged access, connect trust zones, support safety-sensitive operations, or concentrate recovery through one provider. Structure includes the type, direction, privilege, reachability, centrality, and concentration of the relationships. In cybersecurity, one consequential path can matter more than thousands of ordinary connections.

This network view is useful even when no statistical model is fitted. Mapping which components are connected, where connections converge, and which paths carry consequence can reveal opportunities for segmentation, role redesign, simplification, local decision-making, supplier alternatives, and stronger recovery design.

Linear, sublinear, and superlinear growth

The words linear, sublinear, and superlinear describe how an outcome changes as scale changes. They do not describe whether the outcome is good or bad. Interpretation depends on what is being measured.

- **Linear growth.** The outcome grows in the same proportion as scale. If the number of identities doubles, a linear relationship means the expected outcome also doubles. The amount per identity remains stable.
- **Sublinear growth.** The outcome grows more slowly than scale. If identities double, the expected outcome rises by less than two times. This can indicate economies from reuse, standardization, automation, or shared infrastructure when the outcome is cost or effort. It can indicate a growing gap when the outcome is control capacity and demand is growing faster.
- **Superlinear growth.** The outcome grows faster than scale. If identities double, the expected outcome rises by more than two times. This can describe increasing relationship density, coordination demand, exceptions, candidate paths, or loss. It can also be desirable when the outcome is useful capacity, detection coverage, or another benefit.

A cybersecurity program can contain all three patterns at once. A shared logging platform may serve additional systems with declining effort per system. Asset inventory obligations may remain close to proportional. Permissions, integrations, and review exceptions may grow faster than the number of identities or applications. The objective is to identify which pattern applies to a defined outcome across a defined range.

What a power law tells us

A power law is one candidate model for describing a scaling relationship. It connects proportional changes in a scale variable with proportional changes in an outcome. The exponent is the number that describes how responsive the outcome is. An exponent below one describes sublinear growth, one describes proportional growth, and an exponent above one describes superlinear growth. Section 2 introduces the equation and shows how to translate the exponent into practical multipliers.

The word law can sound stronger than the evidence deserves. In this paper, a power law is treated as a model to test within a defined population, period, and operating range. It is not assumed to be a universal property of cybersecurity. An organization can redesign roles, automate work, segment networks, replace a platform, acquire a business, change its measurement process, or face a new threat environment. Any of those changes can alter the observed relationship.

A fitted power law also does not explain itself. If manual access work grows faster than identities, the exponent describes the pattern. The practitioner still needs to determine whether application diversity, non-human access, exceptions, poor role design, incomplete integration, data quality, or another mechanism produced it. A mathematical relationship becomes useful when it directs attention toward a mechanism and a decision.

Other models may fit better. Growth may be linear. It may slow as a process reaches a natural limit. It may change abruptly after a threshold. Several variables may contribute separately. There may be no stable relationship because architecture, maturity, or local context explains more than size. Those are valid results. The analysis should select the simplest explanation that fits the data and improves the decision.

> *The purpose is not to demonstrate that cybersecurity always follows a power law. It is to determine how a particular cybersecurity outcome changes as an organization grows, identify the mechanism producing that change, and use the result to improve a decision. A linear relationship, threshold, saturation point, or absence of a stable relationship may be the correct conclusion.*

A plain-language guide to the symbols

The equations use a small set of recurring symbols. A symbol is simply a short label for a defined quantity. Its meaning comes from the study design, so N might represent identities in one analysis and applications in another. The table below provides the operational meaning needed to read the paper. Additional symbols used in specialized sections are defined where they appear.

Scaling relationship symbols. These symbols describe the relationship between scale and an outcome.

Symbol	Meaning in this paper	Practical reading
N	Defined measure of scale	What growing population or quantity are we studying?
Y	Outcome associated with scale	What result are we trying to explain, compare, or forecast?
a	Coefficient in the short power-law form	What baseline level follows from the chosen units?
N0	Reference scale	What baseline size are later values compared with?

Symbol	Meaning in this paper	Practical reading
Y0	Expected outcome at the reference scale	What outcome is expected at the baseline size?
β	Scaling exponent	How responsive is Y to a proportional change in N?
c	Factor by which scale changes	Are we doubling scale, increasing it tenfold, or using another multiplier?

Operating symbols. These symbols describe incoming work, sustainable capacity, and the pressure between them.

Symbol	Meaning in this paper	Practical reading
D	Incoming security demand per period	How much quality-adjusted work arrives?
K	Sustainable effective capacity per period	How much work can the current operating model sustain at the intended quality?
βD, βK	Exponents for demand and capacity	How do demand and sustainable capacity respond separately to scale?
ρ	Operating pressure, based on demand relative to capacity	How much of sustainable capacity is being consumed?

Pressure, structural, and diagnostic symbols. These symbols describe the scaling gap, accumulated work, relationships, and deviations from the expected pattern.

Symbol	Meaning in this paper	Practical reading
Δβ	Difference between the demand and capacity exponents	Does operating pressure rise, remain stable, or fall as scale grows?
B	Backlog	How much unfinished work has accumulated?
N*	Estimated operating limit	At what scale would demand reach sustainable capacity under the stated assumptions?
E	Number of relationships or edges	How many connections exist among the defined components?
k	Average degree	How many relationships does a typical node have?

Symbol	Meaning in this paper	Practical reading
δ	Exponent describing how average degree changes with scale	Does relationship density increase as the environment grows?
r	Scale-adjusted residual	How far is an observation above or below the model's expected value?

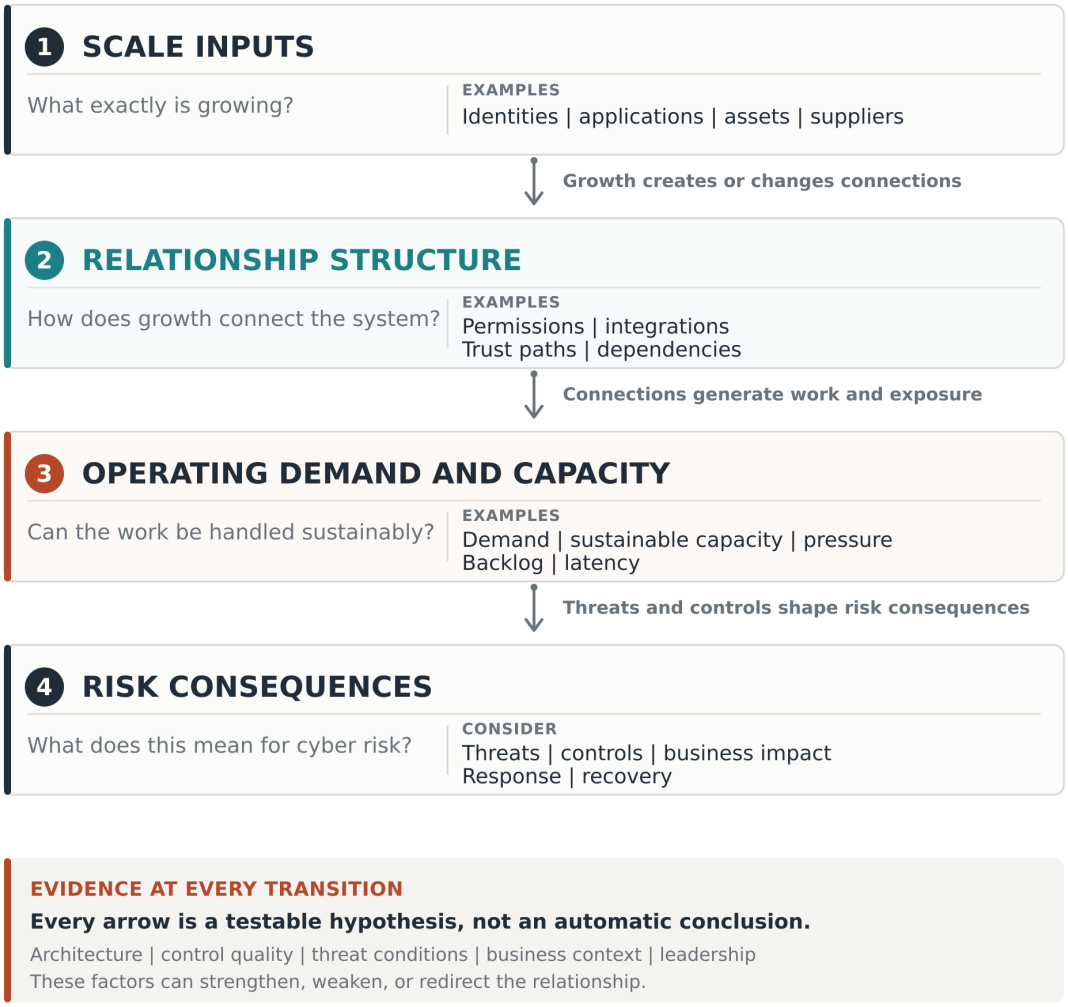
The symbols keep the mathematical statements compact. They do not replace the definitions. Before interpreting a result, translate each symbol back into the population, unit, period, source, and business decision it represents.

The four layers of the scaling problem

A common analytical mistake is to move directly from organizational size to cyber risk. Too many different processes sit between those two ideas. This paper separates the problem into four layers so that a result measured in one layer is not presented as proof in another.

- **Scale inputs.** These are the populations that are growing, such as identities, applications, assets, suppliers, locations, data sources, business services, external connections, AI agents, or incident workstreams.
- **Relationship structure.** These are the permissions, integrations, trust paths, data flows, dependencies, administrative routes, handoffs, and decision interfaces created by the growing population.
- **Operating demand and capacity.** The structure creates work. The security operating model has a finite ability to process that work while meeting its intended quality and service expectations. Queues and delays appear when demand approaches or exceeds sustainable capacity.
- **Risk consequences.** Threat conditions, control strength, reachability, privilege, business dependence, detection, response, and recovery determine what the structure and operating condition mean for the probability and impact of a defined scenario.

The layers are connected, and each connection requires evidence. Entitlement relationships may grow faster than identities while automation keeps manual effort stable. Manual effort may grow quickly without expected loss following the same pattern. A shared platform may lower operating effort while increasing concentration. Keeping the layers separate allows the practitioner to explain each link and avoid turning several different variables into one unsupported conclusion.



Orientation map. The four-layer cybersecurity scaling model. Each transition requires evidence, and contextual factors can alter the relationship.

Demand, sustainable capacity, pressure, backlog, and latency

The operating part of the framework uses several terms that can sound similar. Demand is the work arriving during a period. Capacity is the maximum work the current operating model can sustain during the same period at the intended quality. Meeting a response-time target usually requires operating below that processing limit. Pressure describes how much of that sustainable capacity the demand consumes.

The units must be compatible. If demand is measured as investigation hours per week, capacity must also be expressed as sustainable investigation hours per week. Case counts work when cases are reasonably similar. When one case takes ten minutes and another requires several days of architecture review and business consultation, a raw case count can hide the operating reality. Hours, case types, complexity classes, or separate workflows may give a more honest comparison.

Completed work is not the same as capacity. A team that receives 80 cases and could sustainably complete 100 will normally complete 80. Calling 80 its capacity would hide the reserve. A team that receives 130 cases and completes 100 may appear productive while backlog grows by 30. Capacity has to be estimated from productive time, service rates, work sampling, observed throughput under load, or another defensible method.

Backlog is accumulated unfinished work. Latency is the time work waits or takes to move through the system. Both can worsen before average demand visibly exceeds estimated capacity. Security work arrives unevenly, cases vary in difficulty, urgent incidents interrupt routine tasks, and specialized knowledge may be scarce. A process operating close to its limit has little room for that variation. Waiting time can then rise much faster than workload [27].

This distinction changes the response. More people, external services, or tools can increase capacity and reduce immediate pressure. Simplification can remove demand. Standard interfaces and reusable controls can prevent every new component from creating unique work. Segmentation can reduce the consequence of relationships. Distributed authority can remove approval queues when competence, information, boundaries, and accountability are in place. The intervention should be connected to the part of the system it is expected to change.

Workload, exposure, and risk are different

Workload measures the effort required to operate controls, investigate conditions, make decisions, and complete treatments. Exposure describes conditions that a threat or failure could use, such as reachability, privilege, vulnerable services, weak recovery, or dependency concentration. Risk describes the probability and consequence of a defined scenario after considering threats, controls, response, and recovery.

These variables influence one another, although none can automatically stand in for the others. A rise in vulnerability findings may reflect broader scanning coverage. A rise in control workload may result from better review of important services. A stable incident count may coexist with increasing severity or weaker detection. A rapid increase in entitlement relationships does not establish the same increase in expected loss.

The practical sequence is to measure each layer with a variable suited to that layer. Count and map the exposure condition. Measure the work it creates. Determine whether the operating model can process that work. Then connect the consequential relationships and operating delays to scenarios that matter to the business. This approach preserves the context needed for cyber risk analysis.

Concentration deserves special attention. Central identity, cloud, logging, recovery, or supplier services can create real efficiency. The same design can allow several losses or disruptions to occur through one common cause [16, 17, 18]. Aggregating scenarios as if they were independent can understate this enterprise effect. The paper introduces dependence and common-cause risk later, while the foundation is simple: ask which important services rely on the same control point, provider, technology, or recovery capability.

The data language used in this paper

A few data terms appear throughout the paper. Understanding them prevents many analytical errors, even when the practitioner never builds a statistical model.

Term	Meaning and use
Variable and observation	A variable is something measured, such as active identities, privileged relationships, manual review hours, backlog age, or expected loss. An observation is a recorded value for a particular unit and time, such as one business unit in one month.
Population and boundary	The population identifies what the analysis covers. The boundary defines which items are included and excluded. Results from comparable business units may not apply to customers, operational technology, acquired companies, or environments with different architecture.
Unit and period	Every measure needs a unit and a time basis. Hours per month, cases per week, month-end identities, and annual loss describe different quantities. Mixing them without a clear model produces a result that is difficult to interpret.
Stock, flow, rate, proportion, and duration	A stock is an amount at a point in time, such as open exceptions. A flow is an amount entering or leaving during a period, such as new exceptions per month. A rate relates an amount to time or another measure. A proportion is a share of a defined total. A duration measures elapsed time.
Cross-sectional, longitudinal, and panel data	Cross-sectional data compare several units at one time. Longitudinal data follow the same unit over time. Panel data follow several identified units over time. A relationship across large and small business units today may differ from the way one unit changes as it grows.
Total and per-unit measures	A total describes the full amount. A per-unit measure divides that amount by a denominator, such as exceptions per application. Per-unit measures can hide a nonlinear relationship and can make a larger unit look worse even when it performs exactly as the scaling model predicts.
Expected value and residual	The expected value is what the fitted model predicts for an observation at its scale. A residual is the difference between what was observed and what the model expected. A large residual is a question to investigate, not an automatic performance grade.

Term	Meaning and use
Range, outlier, and breakpoint	The range is the span of scale represented by the data. An outlier is an observation far from the general pattern. A breakpoint is a scale or time at which the relationship changes. Predictions outside the observed range require stronger assumptions.
Observed, estimated, illustrative, and synthetic values	Observed values come from a defined source. Estimated values come from a model and carry assumptions and uncertainty. Illustrative values explain an idea. Synthetic data are generated to demonstrate a method and do not describe a real organization.

What the evidence can support

An observed relationship is an association. It means two measured quantities changed together across the available data. It does not, by itself, show that one caused the other. Application count may rise at the same time as manual review hours, while the actual driver is application diversity, acquisition history, weak integration, or a change in how work is recorded.

A causal claim requires a stronger design. The team needs a clear mechanism, timing, alternative explanations, and evidence that the expected change followed an intervention. Comparison groups, staggered implementation, interrupted time series, or other evaluation designs can strengthen the claim [29, 30]. A simple before-and-after comparison can be useful for monitoring while remaining vulnerable to unrelated changes.

Uncertainty is part of the result. Data can be incomplete, definitions can change, business units can differ, and a small number of observations can support several plausible explanations. An exponent should therefore be reported with its uncertainty, useful range, sensitivity to definitions, and comparison with alternatives. A precise number produced by weak data is still weak evidence.

Prediction also needs restraint. A relationship fitted across the current organization may support a bounded workload scenario within a similar range. It may fail after an acquisition, platform change, control redesign, regulatory shift, or change in attacker behavior. Forecasts should remain local to the population and conditions that produced the data unless there is evidence that the relationship transfers.

The following evidence levels are a practical reading aid rather than a formal standard. They help match the wording of a conclusion to the work that has actually been completed.

Evidence level	What the work contains	What it can reasonably support
1. Structured review	Defined decision, boundary, scale, outcome, relationship map, and proposed mechanism	A plausible operating hypothesis and actions worth investigating; no quantified scaling claim

Evidence level	What the work contains	What it can reasonably support
2. Descriptive evidence	Stable observed values, trends, levels, per-unit views, and carefully labeled two-point comparisons	A statement about what happened in the observed interval; no stable exponent or causal claim
3. Estimated relationship	Sufficient variation, fitted models, uncertainty, diagnostics, and comparison with plausible alternatives	An association for the defined population, period, and useful range
4. Predictive evidence	Held-out testing, stable measurement, forecast error, and evidence that the relationship persists under similar conditions	A bounded prediction within conditions and ranges supported by the validation
5. Intervention or causal evidence	Predefined mechanism, timing, comparison design, alternative explanations, and verification of mechanism and outcome	A conclusion that the intervention contributed to the observed change, subject to the design's assumptions

Common mistakes to avoid

Most weak scaling analyses fail because the question, variable, or claim is poorly defined, not because the calculation is difficult. The following checks address the errors a new practitioner is most likely to encounter.

Common mistake	Why it misleads	Better practice
Treating an asset, alert, finding, or incident count as risk	Counts can represent inventory, visibility, or workload while saying little about scenario probability and consequence	Keep workload, exposure, and risk measures separate, then explain how they connect
Using employee count as the default measure of scale	Headcount may be distant from the mechanism creating integrations, privileges, supplier dependencies, or decision work	Choose the scale variable that matches the mechanism and decision
Treating all relationships as equivalent	A routine connection and a privileged path into a high-consequence system carry different meanings	Preserve type, direction, privilege, reachability, concentration, and business consequence

Common mistake	Why it misleads	Better practice
Calling completed work capacity	Completions are limited by demand when demand is low and can hide backlog when demand is high	Estimate sustainable capacity independently using compatible quality and time units
Calling a two-point comparison a stable law	Two points provide no uncertainty, alternative-model comparison, breakpoint detection, or protection against a measurement change	Label it as descriptive and collect wider, repeated observations before making a scaling claim
Treating a straight log-log pattern as proof	Several functions can look similar across a limited range, and log axes compress deviations	Inspect the original scale, compare plausible models, and examine residuals and held-out performance
Presenting association as causation	Two quantities can move together because of architecture, maturity, measurement, or another shared driver	State the result as an association until a stronger evaluation design supports a causal claim
Using a per-unit metric as an automatic performance grade	The expected amount per unit can change with scale, and visibility or business exposure can differ	Use scale-adjusted comparisons as questions and investigate the context behind deviations
Forecasting far outside the observed range	Acquisitions, thresholds, redesigns, regulation, and attacker behavior can change the relationship	Keep forecasts within supported conditions and show the assumptions and uncertainty
Combining data collected under changing definitions	A dashboard can keep the same label while its population, coverage, or collection method changes	Version definitions, mark structural breaks, and avoid fitting one relationship across incompatible periods

A short identity example

Suppose a business unit grows from 5,000 to 10,000 workforce identities while active entitlement relationships grow from 120,000 to 300,000. Identities doubled. Entitlements increased by more than two times, and the average number of entitlements per identity rose from 24 to 30. Across that interval, relationship growth was faster than identity growth.

That observation is useful because it challenges a proportional planning assumption. It does not establish a stable power law. Two points cannot show uncertainty, identify a breakpoint, separate the effects of application mix, or determine whether a measurement change contributed to the increase. The next step is to investigate the structure: which applications and roles added relationships, which relationships are privileged or dormant, and whether a small number of identities or resources became highly connected.

The operating layer then asks whether certification effort, manual lifecycle work, exception volume, approval latency, and backlog also changed. Automation may have absorbed much of the additional work. If demand grew faster than sustainable capacity, delayed removals or unresolved exceptions may have increased exposure. The risk layer finally asks which access relationships create credible paths to important resources and what threats, controls, response, and recovery mean for those scenarios.

The decision might involve role engineering, application integration, stale-access removal, added processing capacity, tighter privilege boundaries, or acceptance of a temporary backlog. The example shows the full reasoning chain: observe growth, inspect relationships, measure operating demand and capacity, connect consequential conditions to risk, and choose an intervention whose effect can be verified.

What can be done without quantitative analysis

A practitioner can use much of the framework before estimating an exponent. Start by stating the decision that needs to be made. Define the population and the measure of scale. Draw the components and consequential relationships. Describe in plain language how those relationships create work, delay, exposure, or concentration. Measure demand, estimated sustainable capacity, backlog, and latency with stable definitions.

This structured review can reveal that the selected denominator is wrong, the process boundary is inconsistent, a shared dependency supports too many services, manual exceptions are accumulating, or every local decision is being sent to one central forum. Those findings can support immediate design work even when the available data are too limited for statistical inference.

The practitioner can also define the expected mechanism of an intervention. Simplification should remove work or relationships. Automation should reduce processing effort while keeping exceptions and quality visible. Segmentation should reduce reachability or propagation. A reusable control pattern should reduce unique interpretation as services grow. Distributed authority should reduce decision latency inside clear boundaries. Measuring these expected effects creates a learning process.

The nine-step review in Section 9 can therefore begin as a structured practitioner workflow. Steps 1 through 6 can produce useful findings before model estimation begins. Appendix C provides the one-page canvas; the full sequence is presented once in Section 9.

When quantitative support becomes necessary

Quantitative analysis becomes more important when the organization wants to estimate how strongly an outcome responds to scale, compare a power law with other models, calculate a bounded forecast, identify statistically supported breakpoints, compare units after adjusting for scale, or determine whether an intervention changed the baseline or the growth pattern.

The work then requires enough variation in scale, repeated observations when possible, a data dictionary, attention to missing values and zeros, and methods suited to the type of outcome. Counts, proportions, durations, and losses do not always behave the same way. Observations from the same business unit or period may also be related, which affects uncertainty.

A statistician, data scientist, economist, actuary, or quantitative risk specialist can help when the result will influence a material investment, risk acceptance, operating limit, or causal claim. The cybersecurity practitioner remains responsible for the question, definitions, mechanism, data quality, business context, and interpretation. Statistical expertise cannot repair a poorly defined decision or an outcome that does not represent the problem.

What this paper can and cannot provide

The framework can support bounded workload forecasts, capacity scenarios, scale-adjusted comparisons, relationship and concentration analysis, identification of thresholds, and tests of whether a design change improved the operating model. It can help leaders see where growth is changing the system and where effort added to the current design will provide only temporary relief.

The framework cannot predict when an attacker will act, convert an asset count directly into expected loss, prove causation from a fitted curve, or produce credible rare-loss forecasts from a small internal sample. It cannot supply one cybersecurity exponent that applies across organizations, domains, architectures, and periods. Those limitations are part of responsible use.

A model can also fail usefully. A weak fit may reveal a measurement change, a process threshold, a business unit with unusual architecture, a dominant dependency, or a simpler linear relationship. Stopping or narrowing a claim when the evidence is weak protects the decision and often identifies the next question worth investigating.

How to use this paper

Readers interested primarily in strategy, architecture, leadership, and operating decisions can begin with Sections 1 through 9. These sections explain why organizations change as they grow, introduce the four-layer cybersecurity scaling model, apply it across cybersecurity domains, distinguish workload from risk, examine leadership and distributed authority, and provide a nine-step scaling review. The one-page canvas in Appendix C can be used without fitting a statistical model.

Readers conducting quantitative analysis should also study Sections 10 and 11 and the mathematical and validation appendices. Those sections address data structure, model assumptions, dependence,

measurement error, alternative functional forms, intervention evaluation, and a worked synthetic identity case. Appendix D provides checks that help keep the analysis tied to evidence and a decision.

If an equation is unfamiliar, first identify what each symbol represents in operational language. Ask what is growing, what outcome is being measured, what period and population apply, and what the result would change. The worked examples translate the equations into doubling, tenfold growth, operating pressure, relationship count, and correlated loss. The objective is interpretation rather than memorization.

The paper is best used iteratively. Start with a narrow decision and the information available today. Improve the definitions, relationship map, and operating measures. Add quantitative analysis when it will change the decision. Record what the intervention is expected to change, then return to the data and verify whether the mechanism and outcome moved as expected.

1. From Heartbeats to Cyber Systems

The idea that a mouse, an elephant, and many other mammals may experience a broadly similar order of magnitude of lifetime heartbeats is memorable because it makes scale visible [23]. The popular version is simplified. Heart rate and lifespan vary across species, ecological conditions, and taxonomic groups, and the product is not a universal constant. The more defensible lesson is that a larger organism is not a smaller organism multiplied. Its rates, dimensions, and time scales change in patterned ways.

Max Kleiber connected metabolic rate with body mass through an allometric relationship close to a three-quarter exponent [9]. West, Brown, and Enquist later proposed that resource-distribution networks could explain families of quarter-power relationships in biology [24]. The exponent, mechanism, and degree of universality remain debated. More recent work has proposed life-history optimization as an alternative explanation [25]. That uncertainty is useful here. Scaling theory earns its place by producing testable relationships and disciplined questions. An attractive exponent should never be treated as destiny.

Geoffrey West extended this way of thinking to cities and companies [23]. The move from organisms to human systems is attractive because organizations also distribute resources, information, authority, and attention through networks. The comparison has a limit that matters greatly in cybersecurity: organizations redesign themselves. They automate work, change incentives, outsource functions, retire technology, divide into business units, integrate acquisitions, and reorganize teams. Their observed scaling behavior is partly a result of design.

Cybersecurity is a useful setting for this question because a security program sits inside a network of people, technologies, identities, suppliers, data, decisions, and adversaries. Many of those elements can be counted. A large part of the exposure emerges from their connections. An application may be well protected in isolation and still create a path through an integration. A supplier may appear manageable on its own and still become a common dependency for several important services. An identity may be legitimate while its accumulated permissions create an unexpected route to a sensitive system.

That gives us the bridge from scaling theory to cyber risk:

> *Scale changes the number of things. Structure changes how those things interact. Threats, controls, and business consequence determine what those interactions mean for risk.*

The mathematics in this paper is less important than the discipline behind that sentence. The practitioner must define what is growing, map the relationships that growth creates, measure the work imposed on the security operating model, and then connect the result to scenarios that matter to the business.

2. Scaling and Power Laws in Plain Language

2.1 Begin by defining scale

The word *scale* can mean size, rate, scope, or level of aggregation. A mathematical analysis needs something more precise. N must represent a positive and measurable quantity. In one study it may be workforce identities. In another it may be applications, internet-facing services, supplier connections, privileged paths, locations, business services, tool-enabled agents, or incident workstreams.

The choice of N is part of the hypothesis. “Cybersecurity effort grows with company size” is too broad to test. “Monthly manual access-exception hours grow with the number of active workforce identities across comparable business units” can be tested. It identifies a population, scale variable, outcome, and period.

Scale also needs a boundary. Does an identity count include contractors, service accounts, customers, and agents? Does an application include software-as-a-service products, scripts, and low-code workflows? Does a supplier count include subcontractors and shared fourth parties? Boundary changes can materially alter the estimated relationship. Consistent definitions will usually contribute more than complicated mathematics.

2.2 The basic power law

A scaling relationship is commonly written as:

$$Y = aN^{\beta}$$

A reference form is easier to interpret:

$$Y(N) = Y_0 \left(\frac{N}{N_0}\right)^{\beta}$$

N_0 is a reference scale and Y_0 is the expected outcome at that scale. The ratio N/N_0 has no units. The coefficient a in the shorter equation changes when the unit of N changes. Measuring workforce in people or thousands of people changes a even when β remains the same. The model is intended for positive values inside a defined useful range. It does not give a meaningful organizational interpretation at $N = 0$.

The exponent is an elasticity. Near a given point, a 1 percent change in N is associated with approximately a β percent change in Y. For finite changes, use the exact ratio:

$$\frac{Y_2}{Y_1} = \left(\frac{N_2}{N_1}\right)^{\beta}$$

Sublinear and superlinear are descriptions. They are not judgments. A sublinear cost can indicate efficiency. Sublinear control capacity can become dangerous when demand grows faster. Superlinear capacity may be desirable. Superlinear loss can be a warning. Interpretation depends on what Y represents.

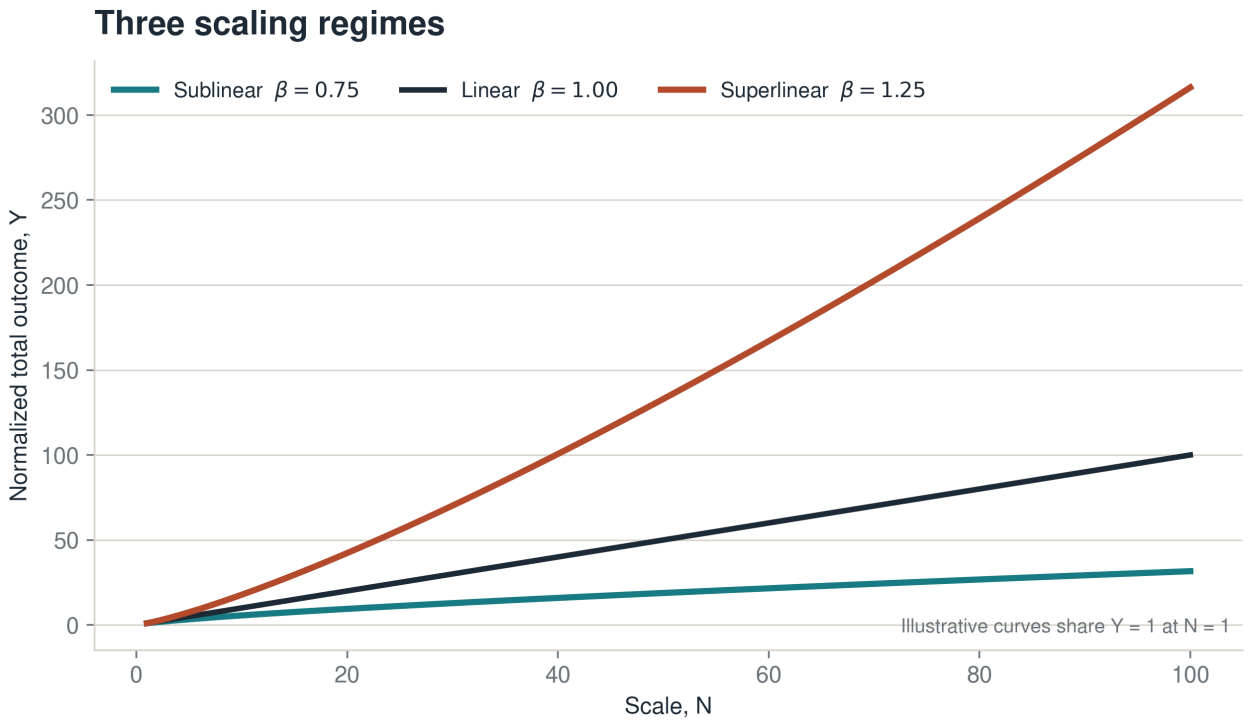


Figure 1. Sublinear, linear, and superlinear scaling. The curves are normalized illustrations, not empirical cybersecurity estimates.

2.3 The doubling rule

Power laws are convenient because proportional changes in scale have a consistent interpretation. If N is multiplied by a factor c, the expected outcome is multiplied by c^{β}:

$$\frac{Y(cN)}{Y(N)} = c^{\beta}$$

For a doubling:

$$\frac{Y(2N)}{Y(N)} = 2^{\beta}$$

If $\beta = 0.80$, doubling N multiplies total Y by about 1.74. Total Y still increases, while the amount per unit falls by about 13 percent. If $\beta = 1.20$, doubling N multiplies total Y by about 2.30, and the amount per unit rises by about 15 percent.

β	Total multiplier when N doubles	Change in Y per unit
0.75	1.68x	16% lower
0.80	1.74x	13% lower
1.00	2.00x	No change
1.15	2.22x	11% higher
1.25	2.38x	19% higher

This translation is often more useful in a meeting than the exponent. A practitioner can say what doubling identities, applications, suppliers, or workstreams would mean under the fitted relationship.

2.4 Total, per-unit, and scale-adjusted views

Dividing total Y by N gives the per-unit relationship:

$$\frac{Y}{N} = \frac{Y_0}{N_0}\left(\frac{N}{N_0}\right)^{\beta-1}$$

Per-unit metrics are familiar. We use cost per employee, incidents per asset, exceptions per application, or loss per unit of revenue. They help, and they can also hide nonlinear structure. If β is above one, a larger unit will naturally have a higher Y/N even when it performs exactly as the model predicts. Comparing raw per-unit values can then penalize scale.

A scale-adjusted residual asks how far an observation sits above or below the expected relationship:

$$r = \ln(Y_{observed}) - \ln(Y_{expected})$$

The residual is a log ratio, so $e^r = Y_{observed}/Y_{expected}$. A residual of 0.20 means the observation is about 22 percent above the expected value after accounting for scale. That result does not yet mean good or bad. Architecture, business exposure, measurement quality, maturity, and local practice may explain the deviation. The next question is why.

2.5 A scaling relationship is different from a power-law distribution

Two ideas are often called a power law:

Idea	Question	Cybersecurity example
Power-law scaling relationship	How does outcome Y change with scale N?	How do entitlement edges change as identities grow?
Power-law probability distribution	How frequently do values of different sizes occur?	Do a small number of incidents account for a large share of loss?

This paper focuses mainly on the first. The statistical methods used to test whether a distribution has a heavy power-law tail are not automatically the right methods for estimating a relationship between N and Y. Clauset, Shalizi, and Newman provide well-known guidance for testing power-law distributions and comparing heavy-tailed alternatives. Their warning against declaring a power law from a straight log-log line remains useful, while the model and error assumptions still need to match the question being asked [5].

Power-law scaling also differs from exponential growth. In a power law, Y grows as a power of N. In an exponential relationship, Y grows as a power of a constant, such as e^{cN}. A superlinear power law can become large, although an exponential function eventually grows faster. Cybersecurity discussions

sometimes use *exponential* as a synonym for fast. The distinction matters when we make a mathematical claim.

2.6 Relationships and the coordination ceiling

A group of N people has at most $N(N-1)/2$ unique one-to-one pairs. Five people have 10 possible pairs. Ten have 45. One hundred have 4,950. This complete-network calculation grows approximately as N^2.

Real organizations and technical environments are sparse. Teams, hierarchy, segmentation, platforms, and interfaces prevent every node from connecting to every other node. A more useful expression connects edge count E to average degree $\overline{k}$ in an undirected network:

$$E = \frac{N\overline{k}}{2}$$

If average degree remains constant, edge count grows linearly. If average degree grows as N^{δ}, then:

$$E \propto N^{1+\delta}$$

Relationship growth becomes superlinear when each additional node also connects, on average, to more of the existing environment. Employees may use more applications. Applications may expose more APIs. Suppliers may support more services. Agents may receive access to more tools and data sources. The complete-network formula provides a ceiling. Forecasting requires the observed network. The practical question is whether average degree, privilege, concentration, or reachability is increasing as the environment grows.

Directed, bipartite, and typed graphs require formulas that match their structure. An identity-to-resource graph should not be treated as an undirected social network. Edge semantics become especially important in cybersecurity because one privileged, reachable relationship can carry more consequence than thousands of ordinary edges.

3. What Other Fields Can Teach Us

3.1 Biological networks: efficiency through distribution

Biological scaling research asks how organisms deliver energy and materials across very different sizes. One influential explanation focuses on branching networks and transport constraints [24]. Other work examines life-history optimization [25]. The organizational lesson is architectural: growth can be supported by nested distribution networks and local processes instead of global coordination.

An enterprise is not an organism. People interpret instructions, negotiate priorities, resist controls, and redesign networks; adversaries adapt as well. The transferable question is which distribution architecture allows a security system to grow without making every decision global.

3.2 Cities: economies and acceleration together

Research by Bettencourt, West, and colleagues reported two broad urban patterns. Shared infrastructure often scaled sublinearly, while many social and economic quantities scaled superlinearly. Large cities appeared to share physical networks efficiently while creating more opportunities for interaction [3, 4].

Cybersecurity can show the same combination. Central identity, logging, cloud foundations, and shared tooling can reduce marginal cost. Software reuse, data exchange, supplier connectivity, and shared platforms can add dependencies and opportunities for compromise. One platform may create an economy of scale and a common mode of failure.

Urban scaling also warns us that results change with city definition, outcome, error model, and whether units are compared at one point or one unit is followed over time [2, 6, 10]. Cybersecurity has the same problem. A cross-sectional model may not predict growth through acquisitions, architecture change, regulation, or major incidents.

3.3 Human attention and social layers

Dunbar connected primate neocortex measures with social group size and proposed a human relationship scale near 150 [7]. Later work found layered personal networks, with an inner group receiving more time [19, 21]. The values vary; 150 is not a hard organizational limit. The durable point is that attention and relationship quality are constrained, so networks organize into layers.

This helps explain why informal leadership loses reach as groups expand. One leader may know ten people and their work directly; that relationship cannot extend to hundreds. Teams, managers, routines, and shared language preserve context while filtering information. The design question is what remains local, what is standardized, and what must reach enterprise leadership without losing meaning.

4. Why Organizations Change as They Grow

Organizations coordinate specialized work. As specialization increases, the organization must process more information to align tasks, resolve uncertainty, and manage interdependence. Galbraith described organization design as an information-processing problem: greater uncertainty increases the information that must be processed during task execution [8]. Herbert Simon's work on bounded rationality adds the human constraint. Decision-makers have limited information, time, and cognitive capacity, so they use structures and rules to reach workable decisions [20].

Growth creates opposing effects. Specialization creates capability, while integration creates coordination demand. Shared services create efficiency, while shared dependencies create common-cause exposure. Standardization increases consistency, while excessive standardization can slow local response. Hierarchy reduces direct relationships, while too many layers can increase latency and filter weak signals. Scale changes the trade-offs that organization design must manage.

Network structure matters as much as network size. Small-world research shows that a network can combine local clustering with short paths across the system [22]. Research on heterogeneous networks has shown that some networks tolerate random failures and remain highly exposed to targeted disruption of central hubs [1]. In an enterprise, an identity provider, remote-access service, cloud tenant, domain service, integration platform, or supplier can become such a hub. Consolidation can improve control and cost while increasing the consequences of failure or compromise.

Organizations also experience thresholds. A process can appear stable until volume, coupling, or decision latency crosses a point where queues grow, exceptions accumulate, and workarounds become normal. Added effort may no longer restore performance. The operating model needs structural change. The analogy to a phase transition should remain modest. Organizational thresholds are neither universal nor precisely predictable.

The question for a practitioner is direct:

> *As scale increases, which relationships should remain local, which should be standardized, which should be automated, and which require enterprise judgment?*

5. The Cybersecurity Scaling Model

5.1 Four linked layers

Cybersecurity discussions often jump from organizational growth directly to cyber risk. The jump is too large. A scaling-aware analysis separates four linked layers.

- **Scale inputs** are the populations that are growing: identities, applications, assets, suppliers, locations, data sources, business services, external connections, agents, and incident workstreams.
- **Relationship structure** contains the permissions, integrations, trust paths, data flows, shared dependencies, administrative routes, and decision interfaces created by those populations.
- **The operating system** contains the demand created by that structure, the sustainable capacity available to process it, the queues that form, and the decisions that must be made.
- **Risk consequences** depend on scenario probability and impact after considering threats, controls, reachability, business dependence, detection, response, and recovery.



Figure 2. The scaling-aware cybersecurity model. Each transition requires evidence, and moderators can alter the relationship.

The exponent at one layer does not transfer automatically to the next. Entitlement edges may grow superlinearly while automated lifecycle controls keep manual work close to linear. Manual work may grow superlinearly without expected loss doing the same. A common platform may reduce operating effort and increase concentration. Each transition needs its own measure and explanation.

This is the discipline at the center of the paper. Count components for inventory. Map relationships for propagation. Measure demand and sustainable capacity for operating performance. Model scenarios for risk. Keep the layers connected without compressing them into one number.

5.2 Demand and sustainable effective capacity

Let security demand D represent incoming work per period. Examples include quality-adjusted assessment hours per month, investigation effort per week, access decisions per quarter, or incident tasks per day. The phrase *quality-adjusted* matters. Ten routine cases are not necessarily equivalent to ten cases requiring business consultation, architecture analysis, or privileged-access review.

Let sustainable effective capacity K represent the maximum work the current operating model can process per period at the intended quality under sustainable working conditions. Demand and capacity must use compatible units and the same period.

$$D(N) = D_0 \left(\frac{N}{N_0}\right)^{\beta_D}$$

$$K(N) = K_0 \left(\frac{N}{N_0}\right)^{\beta_K}$$

Their ratio is operating pressure:

$$\rho(N) = \frac{D(N)}{K(N)} = \rho_0 \left(\frac{N}{N_0}\right)^{\Delta\beta}$$

$$\Delta\beta = \beta_D - \beta_K$$

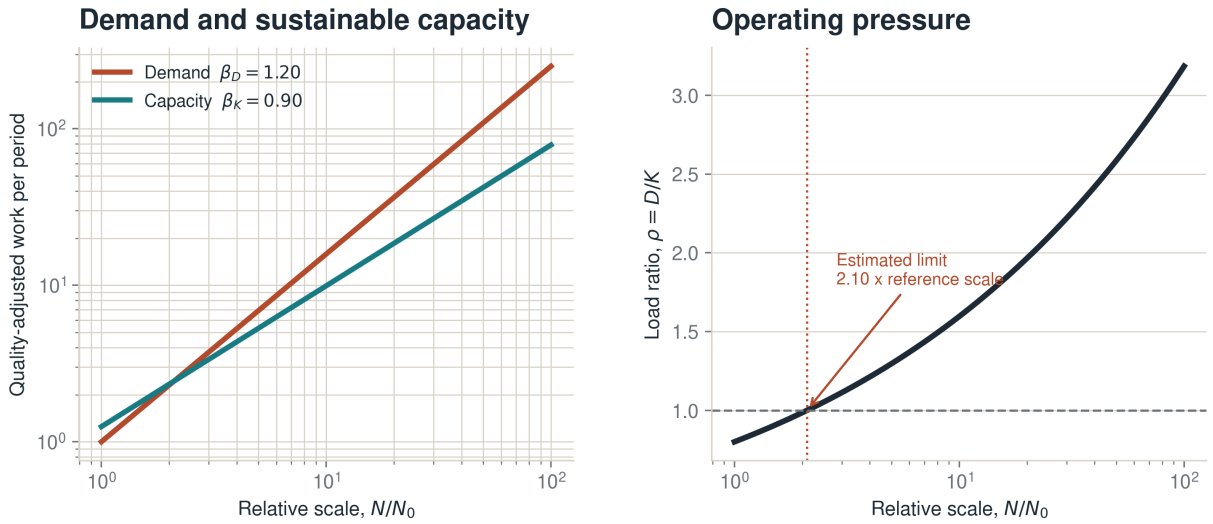
Here, $\rho_0 = D_0/K_0$ is pressure at the reference scale. A value below one indicates theoretical reserve. A value near one indicates little room for demand variation, rework, urgent cases, leave, or service disruption. A value above one means incoming work exceeds sustainable capacity and backlog must accumulate unless demand is deferred, rejected, or processed below the intended quality level.

Capacity should not be measured as cases completed. If demand is 80 cases and a team could sustainably process 100, observed completions will usually be 80. Using completions as capacity would produce a ratio of one and hide the reserve. Capacity must be estimated from available productive hours, work sampling, case-type service rates, stable throughput under load, controlled capacity tests, or another defensible method.

Raw headcount is also weak. Analysts differ in skill, context, tooling, interruption, and available time. A team of ten can have less effective capacity than a team of eight if rework, fragmentation, or coordination consumes the difference. A stronger measure converts the work and the available service capacity into common units.

Processing capacity and a service target are different boundaries. A team may be able to sustain a given service rate while needing reserve to meet an agreed response time. In the equations below, capacity means sustainable processing potential. The organization should also choose a lower operating threshold from its arrival patterns, case mix, scarce skills, and service commitments. There is no universal safe load ratio.

Demand must measure work arriving, including work that remains unprocessed. Logged effort alone can understate demand when the team is overloaded. Estimate incoming work from arrivals and stable case-type effort weights, and record rejected, deferred, and reworked cases separately. Keep the accounting boundary consistent when evaluating automation so that the same saved effort is not counted as both lower demand and higher capacity.



Illustrative exponents and baseline pressure. Capacity means maximum sustainable service capacity, not observed completions.

Figure 3. Demand, sustainable processing capacity, and the illustrative saturation boundary. A service-target operating limit may be lower.

5.3 Pressure, backlog, and latency

Pressure describes the relationship between incoming demand and capacity at a point. Backlog is a stock that accumulates over time. For a simple fluid approximation with compatible work units and one-period time steps:

$$B_{t+1} = \max[0, B_t + D_t - K_t]$$

This approximation assumes that available capacity can serve the waiting work. In an actual workflow, backlog changes with incoming work and realized completions, with cancellations, rejected work, and rework recorded explicitly. Skill bottlenecks, blocked cases, and interruptions can prevent the team from using all of its processing capacity.

Under stable long-run conditions and a consistent system boundary, Little's Law relates average work in the system, arrival rate, and average time in the system [11]. It is an accounting relationship, not the source of the congestion curve below.

$$L = \lambda W$$

Here, L is the long-run average number of items in the system, λ is the average arrival rate, and W is the average time in the system. All three measures must use the same boundary and compatible units. When demand repeatedly exceeds capacity, queues grow. Decision latency rises, evidence becomes stale, exceptions remain open, and teams invent workarounds. Backlog can remain hidden until an audit threshold, service commitment, or incident exposes it.

The operating system often degrades before ρ reaches one. Security work does not arrive at a perfectly constant rate, and cases do not take the same amount of time. Priority incidents interrupt ordinary work. Difficult cases require rare expertise. Rework returns completed items to the queue. Near full load, the team has little ability to absorb that variation.

Queueing theory gives a useful warning. In the M/M/1 teaching model, arrivals are Poisson, service times are independent and exponentially distributed, and one server works at a fixed average service

rate. Time in the system, normalized by mean service time, rises as $1/(1-\rho)$. At 50 percent load, the normalized value is 2. At 80 percent it is 5. At 90 percent it is 10 [27]. Real security workflows have multiple people, priorities, dependencies, and service-time distributions, so this curve should not be used as a direct forecast. The mechanism still matters: high average load can produce a disproportionate increase in waiting time.

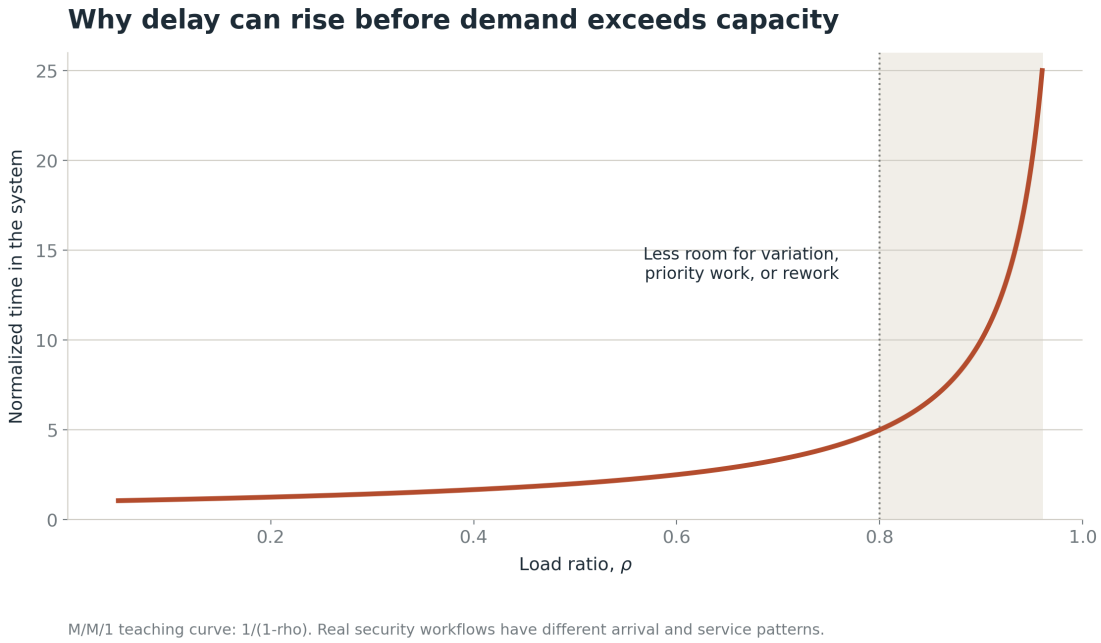


Figure 4. An M/M/1 teaching queue shows why latency can rise sharply as processing utilization approaches one. Based on [27].

This explains a familiar pattern. A risk-assessment team may appear busy but stable, then suddenly miss every commitment after an acquisition. A SOC may hold alert volume steady while investigation time expands. An access-governance process may complete routine requests and leave complex exceptions aging in the queue. The visible backlog is the end of the story. The structural gap started earlier.

5.4 Estimating an operating limit

When $\Delta\beta > 0$, the model can estimate the scale at which pressure reaches one. Set $\rho(N^*) = 1$ and solve for N^*:

$$\frac{N^*}{N_0} = \left(\frac{1}{\rho_0}\right)^{1/\Delta\beta}$$

Suppose current pressure is 0.80 and demand grows with $\beta_D = 1.20$ while capacity grows with $\beta_K = 0.90$. The difference is 0.30. The estimated operating limit is:

$$\frac{N^*}{N_0} = \left(\frac{1}{0.80}\right)^{1/0.30} \approx 2.10$$

Under those assumptions, incoming work reaches sustainable processing capacity when scale becomes about 2.1 times the reference level. This is a saturation boundary, not a promise that service targets remain acceptable until that point. This is not a deadline carved into the organization. The

estimate inherits uncertainty from the exponents, baseline pressure, measurement quality, and the possibility that the process changes regime. Its value is in scenario planning. It gives leaders a defensible reason to redesign before overload becomes the only visible signal.

A service-target limit should be estimated before saturation. If the reference pressure is 0.80 and the organization has evidence that its service target requires pressure no higher than 0.90, the same illustrative scaling gap of 0.30 puts that threshold at about 1.48 times the reference scale. The threshold must come from the workflow, not from a universal percentage. When the scaling gap is close to zero or its uncertainty includes zero, a single operating-limit estimate can be unstable or misleading.

5.5 What an intervention changes

An intervention can move the baseline, change the scaling relationship, move a breakpoint, reduce variance, or alter risk concentration. These effects should not be treated as equivalent.

Hiring analysts, purchasing capacity, or outsourcing work can increase K_0. The immediate queue may improve while $\beta_D - \beta_K$ remains unchanged. The organization has bought time.

Automation can increase K_0 by processing current work faster. It may also increase β_K when the same automated capability can serve additional scale with low marginal effort. Automation can have the opposite effect if it creates alerts, exceptions, integrations, or manual assurance work.

Simplification reduces N or the demand baseline. Application retirement, stale-identity removal, duplicate-control elimination, and deletion of unused data relationships can remove work from the system.

Architecture can reduce edge density, privilege, or reachability. Segmentation may leave the node count unchanged while reducing propagation. Standard interfaces and reusable control patterns can prevent each new component from requiring unique interpretation.

Distributed authority can increase parallel decision capacity. It works when authority is accompanied by competence, information, boundaries, and accountability. Delegating decisions without context redistributes risk. Retaining decisions that local leaders can safely make creates queues.

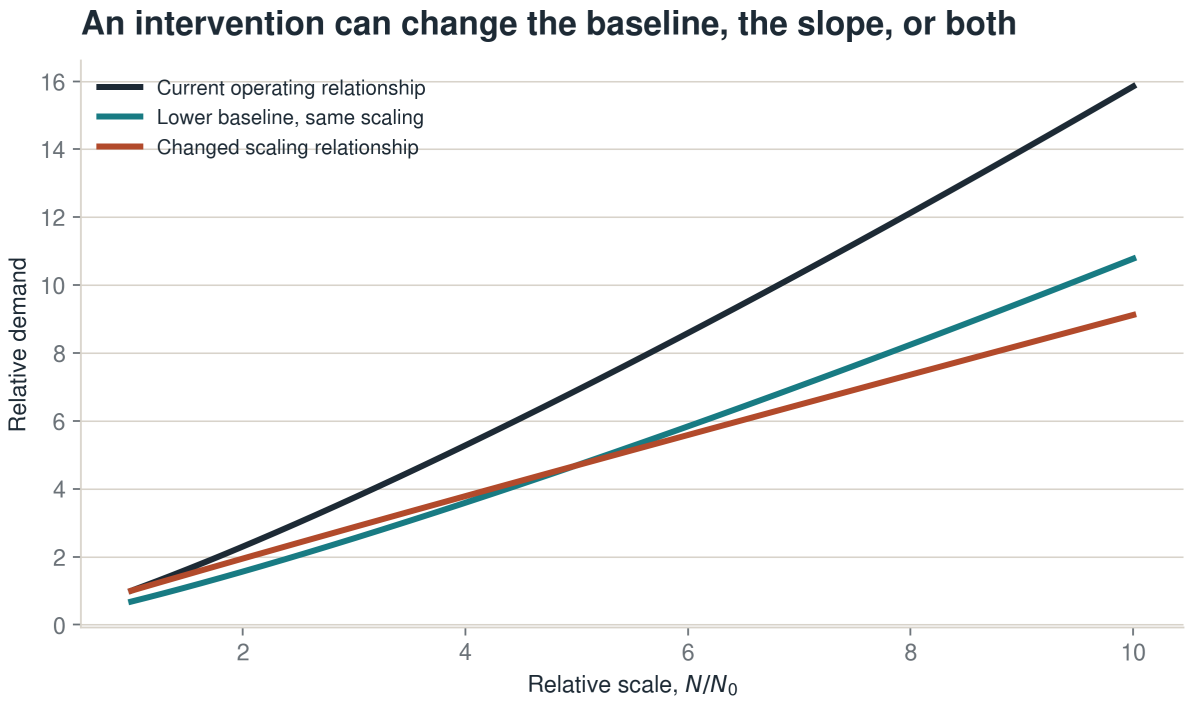


Figure 5. A lower baseline provides immediate relief. A changed slope improves how the operating model behaves as scale continues to grow.

Intervention	Primary model effect	Practical example
Simplify the environment	Reduce N or D_0	Retire applications, stale identities, and unused data flows
Reduce relationship density	Reduce edge growth or β_D	Constrain integrations, roles, privilege, and trust paths
Increase immediate capacity	Increase K_0	Add effective hours, external service capacity, or improved tooling
Build reusable capacity	Increase K_0 and potentially β_K	Automate lifecycle events, standardize evidence, use policy as code
Improve prioritization	Reduce avoidable demand	Suppress duplicate alerts and focus on consequential exposures
Distribute judgment	Increase parallel capacity and reduce queues	Define decision rights and bounded local authority
Reduce common modes	Change concentration and tail exposure	Create independent recovery paths and supplier exit options

Changing an exponent is harder to demonstrate than shifting a baseline. A six-month improvement after implementation may show lower demand or higher capacity. It may not provide enough variation in scale to show that the slope changed. Practitioners should state which result they can actually observe.

5.6 A model that can fail usefully

Organizations adapt, adversaries respond, and measurement changes behavior. One exponent is unlikely to hold across every business unit, architecture, or period. A merger, platform migration, regulatory change, control redesign, or major incident can create a breakpoint.

That does not remove the value of the analysis. A poor fit can identify the reason the model fails. A business unit far above the demand curve may carry a different architecture or business exposure. A relationship that changes after a certain size may indicate a process threshold. A linear model that performs as well as a power law tells the team that a nonlinear story is unnecessary.

The analysis needs stopping rules. Abandon or narrow the scaling claim when the exponent is highly uncertain, a plausible alternative predicts better, one unit drives the result, definitions changed, the observed range is too narrow, residuals show missing structure, held-out performance is poor, or the mechanism cannot be explained. Use the simplest model that survives contact with the data and improves a decision.

6. Applying the Model Across Cybersecurity

The framework is not a single enterprise equation. Each domain needs its own scale variable, outcome, mechanism, and risk connection. The examples below are starting points rather than prescribed metrics.

6.1 Attack surface and asset management

Asset counts are a beginning. Exposure depends on connectivity, privilege, reachability, data sensitivity, control coverage, and business consequence. Two environments with the same number of servers can have very different attack surfaces because one is segmented and standardized while the other contains broad trust paths, exceptions, and external connections.

A scaling-aware asset model tracks nodes and edges. Nodes include applications, devices, accounts, datasets, locations, and services. Edges include authentication relationships, network routes, APIs, data transfers, administrative paths, and supplier access. The practitioner asks whether edge count is growing faster than node count, whether average degree is increasing, whether a few nodes are becoming more central, whether paths between low-trust and high-consequence zones are shortening, and whether control coverage is reusable across the added nodes.

Security architecture can change the relationship by making growth modular. Stable interfaces, bounded trust zones, standard identity patterns, and explicit data flows allow the environment to expand without allowing every new component to connect broadly.

6.2 Identity and access

Identity provides a clear example of relationship growth. Headcount may rise by 20 percent while entitlements rise faster because employees use more applications, non-human identities multiply, cloud roles accumulate, and exceptions remain after job changes. The relevant structure is the identity-resource graph.

The total number of active entitlement edges is a useful starting measure. More informative measures include privileged edges, cross-domain trusts, dormant relationships, toxic combinations, service-account ownership gaps, and access concentrated through highly connected identities. Identity governance can create an economy of scale when common lifecycle processes serve many systems. That benefit disappears when every application requires unique manual interpretation.

Consider a descriptive example in which a business unit grows from 5,000 to 10,000 workforce identities while active entitlement edges grow from 120,000 to 300,000:

$$\beta = \frac{\ln(300{,}000/120{,}000)}{\ln(10{,}000/5{,}000)} \approx 1.32$$

Entitlements per identity rise from 24 to 30. The observation suggests superlinear relationship growth across that interval. It does not establish a stable law. Two points provide no uncertainty estimate, and

application mix, non-human access, or data quality may explain the difference. The calculation still has practical value because it directs attention toward role design, application breadth, lifecycle integration, and accumulated privilege.

Identity growth through system accumulation

As an illustrative scenario, consider an organization implementing a new enterprise resource planning system while keeping its previous ERP systems operational. Employees may need access to both environments during migration. Existing roles may be recreated in the new platform, temporary permissions may remain longer than planned, and integrations must be built between old and new applications. The workforce may remain almost the same size, yet the number of entitlements, service accounts, interfaces, approval paths, and access exceptions can increase substantially.

The organization may also introduce AI systems that use service principals and automation accounts, authenticate through API keys or other credentials, and connect agents or retrieval services to business resources. These non-human identities can access applications, data sources, development environments, and business processes without following the same lifecycle as employees. Some may have broad permissions, unclear ownership, weak credential-rotation practices, or no defined process for reviewing whether their access is still required.

A third source of growth appears when vendors add AI capabilities to products already used by the organization. A familiar supplier may begin processing organizational data through an external model provider, create new service identities, connect additional APIs, or rely on subcontractors that were not part of the original assessment. The product may still appear as one application and one supplier in the inventory, while its identity relationships, data flows, dependencies, and possible failure paths have changed.

Employee count would provide a weak measure of scale in this situation. More useful variables include the number of applications operating in parallel, human and non-human identities, active entitlement edges, privileged relationships, vendor-managed identities, external trust connections, and identities with access to both legacy and replacement systems. These variables should be examined separately before deciding whether a multivariable model is needed. Migration phases and retirement dates should also be recorded. A temporary period of dual access can create a workload peak; it should not be treated as evidence of a persistent scaling relationship unless that pattern remains beyond the transition.

The relationship structure then creates operating demand. Identity teams must integrate new applications, reconcile old and new roles, identify owners for non-human identities, certify vendor access, rotate credentials, investigate exceptions, and remove permissions from systems approaching retirement. Third-party risk, architecture, privacy, and security operations teams may also need to review new data flows, model providers, monitoring capabilities, contractual responsibilities, and incident-response arrangements.

If this demand grows faster than sustainable capacity, access reviews may be delayed, temporary permissions may persist, non-human identities may remain without accountable owners, and vendor AI connections may operate without the intended controls. The resulting risk depends on which identities are privileged, what data and processes they can reach, how activity is monitored, whether access can be removed quickly, and how dependent the organization has become on the connected systems and providers.

This example shows why growth should not be measured only through employees or applications. The organization may appear stable when viewed through inventory counts while its identity and dependency structure is expanding. A scaling review can reveal which relationships are increasing, what work they create, and whether the operating model can handle that work. Practical decisions include retiring legacy access, setting expiry dates for temporary permissions, assigning accountable owners to non-human identities, and reassessing material vendor AI changes before system overlap becomes permanent complexity.

The analysis should then move to the operating layer. Measure manual lifecycle hours, access exceptions, certification effort, approval latency, rework, and backlog. Finally, connect high-consequence relationships to scenarios such as privilege escalation, persistent orphaned access, cross-domain compromise, or inability to remove access quickly during an incident.

6.3 Third-party and supply-chain risk

Supplier risk is often managed as a list of vendors. A scaling view treats it as a dependency network. One supplier may support several important services. Several suppliers may depend on the same cloud provider, identity platform, software component, telecommunications carrier, or managed service. Aggregate exposure is shaped by convergence and substitutability as well as vendor count.

The NIST Cybersecurity Framework 2.0 places cybersecurity supply-chain risk management within the Govern function and emphasizes strategy, roles, integration, monitoring, and improvement [13]. A scaling-aware program extends that work by asking where relationships converge. What share of important services depends on one supplier or platform? Which dependencies lack tested alternatives? Which fourth parties appear repeatedly? Does assessment demand grow with vendor count, service count, external connection count, or concentration? Which controls can be inherited reliably, and which still require service-specific evidence?

A supplier can have strong controls and still represent substantial concentration. The leadership question is larger than whether each provider passed an assessment. Leaders need to know whether several important services can fail through one dependency and whether the organization can recover, substitute, or exit when it does.

6.4 Cloud, software, and artificial intelligence

Cloud and software platforms can create strong economies. A common landing zone, policy as code, shared logging, standardized identity, and repeatable deployment patterns can protect additional workloads with lower marginal effort. The same platforms create highly connected control points. A

configuration error, compromised administrator, insecure automation token, or failed shared service can affect many workloads at once.

Modern AI-enabled systems make relationship complexity visible. The NIST Generative Artificial Intelligence Profile examines risks across the AI lifecycle and its system context [14]. For a scaling review, the relevant relationships can include prompts, enterprise data, retrieval sources, vector stores, tools, APIs, identities, workflows, memory, monitoring, and downstream decisions. NIST's discussion of tool use in agent systems distinguishes the model, tool, and surrounding system [15]. For the scaling framework, the agent's permissions and orchestration relationships belong inside the security boundary.

The useful scale variable depends on the decision. Model count matters for inventory. Tool relationships matter for agency. Context sources matter for poisoning and leakage. Non-human identities and permission breadth matter for unauthorized action. Shared orchestration matters for concentration. Business-process dependence matters for resilience.

AI can affect both sides of the scaling gap. It can increase capacity by summarizing evidence, automating repetitive analysis, translating queries, and assisting triage. It can increase demand by producing more software, creating more agents and tokens, accelerating adversary activity, and adding new paths between instructions, data, and actions. The result needs to be measured rather than assumed.

For an agent program, a useful study might ask whether tool-permission edges grow faster than agent count and whether assurance hours grow faster than the team's sustainable review capacity. The interventions proposed here include purpose-specific identities, tool allowlists, permission boundaries, separated execution tiers, human authorization for consequential actions, and retirement of unused tokens and context sources. These are design options to test against the agent's purpose and consequence, informed by the broader AI risk and tool-use considerations in [14, 15].

6.5 Operational technology and critical infrastructure

In operational technology, scale may be driven less by device count and more by coupling and consequence. A small number of systems can support a large physical process. Long lifecycles, safety dependencies, remote access, shared engineering workstations, historians, and vendor-support connections can create concentrated exposure. One relationship can matter more than hundreds of low-consequence assets.

A composite scale score can become misleading. Combining device counts, trust-zone crossings, safety dependencies, remote-access paths, and recovery constraints requires defensible weights. It is often better to analyze several variables separately or use a multivariable model.

Useful measures include remote-access routes, trust-zone crossings, shared engineering resources, common credentials, safety dependencies, and important processes relying on the same recovery capability. Modularity and segmentation can reduce propagation, while local operating knowledge

remains necessary for safe decisions. The scaling lens supports tailoring. It should not force enterprise IT metrics onto operational environments.

6.6 Incident response and crisis coordination

Incidents increase information demand when time and attention are constrained. More participants can add expertise and authority. They also add communication paths, handoffs, and opportunities for conflicting reports. Participant count is a weak measure on its own. Workstreams, decisions, jurisdictions, affected services, and technical domains may represent scale better.

The response is structured coordination: defined roles, a stable incident-command model, explicit decision rights, shared facts, controlled channels, and delegated workstreams. A response model that depends on one leader understanding and approving everything may work in a small event and collapse in a multi-business, multi-jurisdiction IT and OT crisis.

Exercises can measure decision latency, handoffs, unresolved information conflicts, escalation volume, and rework as scenario scope expands. A sharp change after a certain number of workstreams may indicate a span-of-control or information-synthesis breakpoint. The team does not need to force a power-law fit. Finding the breakpoint and redesigning the command structure may be enough.

6.7 Vulnerability and exposure management

Vulnerability counts grow with assets, scanning coverage, software diversity, and disclosure volume. Findings may rise after visibility improves, so growth can reflect better measurement as much as worse exposure. Business consequence does not scale with counts alone. Exposure concentrates where vulnerabilities combine with reachability, privilege, exploitation, business importance, weak detection, and limited recovery.

A scaling-aware analysis separates discovery, decision demand, and risk reduction. Discovery measures findings generated per asset or product. Decision demand measures the findings requiring human context, acceptance, or coordinated remediation. Risk reduction measures consequential paths or expected loss reduced.

Automation is well suited to repeated, low-context work. Human judgment can focus on exceptions, business trade-offs, and pathways that connect an exploitable condition to meaningful consequence. MITRE ATT&CK can support analysis of adversary behaviors and paths [12]. Technique counts should not be treated as a direct measure of risk.

6.8 Security operations and alerting

Security operations shows why counts, rates, coverage, and capacity need to remain separate. Telemetry may grow with endpoints, cloud services, identities, data retention, and detection coverage. Alert volume may grow at a different rate because logic, duplication, and suppression change. Investigation demand may follow another relationship because alerts vary in context and complexity.

If events grow superlinearly while high-fidelity investigations grow sublinearly and coverage remains dependable, the program may be gaining leverage. If alerts remain flat while telemetry coverage falls, the apparent efficiency is misleading. A useful analysis connects coverage, signal quality, investigation effort, time to decision, confirmed outcomes, and backlog.

This domain especially benefits from queue analysis. A SOC can maintain average throughput and still suffer rapidly increasing investigation age as load approaches sustainable capacity. Separate routine and high-complexity cases if one weighted measure hides too much variation.

6.9 Phishing and human-targeted attacks

Phishing can be studied with scaling, although employee count is often the wrong denominator. Threat opportunity changes over time. A better scale variable may be malicious messages delivered, unique users exposed, user-campaign interactions, confirmed credential-harvesting attempts, or exposed privileged identities.

Possible outcomes include verified credential submissions, token theft, account takeover, analyst investigation hours, remediation effort, and business-impacting compromise. Because successes are bounded by opportunities and often rare, a binomial, Poisson, or negative-binomial model may fit better than a simple log-log regression. The scaling review still helps because it requires the team to compare alternatives instead of choosing a power law in advance.

Scale-adjusted comparisons can help determine whether one business unit has more successful events than expected given exposure. External benchmarking requires compatible definitions, attack volume, campaign difficulty, control coverage, and validation. An organization with stronger detection may appear worse if it simply finds more events.

7. Cyber Risk at Scale

7.1 Workload, exposure, and risk are different variables

Cyber risk is usually expressed through scenarios that combine likelihood and impact [17, 18]. Workload measures the effort required to operate controls and make decisions. Exposure describes conditions an adversary or failure can exploit. These variables influence one another, and none is a direct substitute for another.

A superlinear increase in vulnerability findings does not prove superlinear risk. The findings may be low consequence, duplicates, or the result of improved coverage. A sublinear increase in incidents does not prove an economy if severity rises. Growing control workload may reduce risk when the work is well targeted. The outcome must match the decision.

When risk is modeled directly, define it clearly. Candidate measures include annualized expected loss, frequency of a defined scenario, loss exceedance at a chosen percentile, downtime, safety consequence, or a validated proxy. Qualitative risk scores are ordinal in many methods and should not be treated as though equal numerical differences have consistent meaning [17, 18, 31].

The scaling framework can supply inputs to risk analysis. It can identify how exposure relationships, operating delays, and concentration change with scale. Scenario analysis must still estimate threat frequency, control effect, consequence, response, and recovery.

7.2 Aggregation, dependence, and common causes

Risk registers often present scenarios as separate entries. Enterprise exposure depends on how those scenarios move together. For losses L_1 through L_n:

$$\mathrm{Var}\left(\sum_i L_i\right) = \sum_i \mathrm{Var}\,(L_i) + 2\sum_{i<j} \mathrm{Cov}\,(L_i, L_j)$$

If losses are independent, the covariance terms are zero. Common causes and coupling matter when cyber scenarios are aggregated for enterprise decisions [16, 17, 18]. In a cybersecurity review, candidate shared dependencies include an identity platform, cloud provider, administrator group, supplier, software component, network service, or recovery weakness. Positive dependence means several losses may occur together.

For n losses with equal variance σ^2 and equal pairwise correlation r:

$$\mathrm{Var}(\textstyle\sum L_i) = n\sigma^2[1 + (n-1)r]$$

With 100 standardized losses, variance 1, and pairwise correlation 0.01, aggregate variance is 199 rather than 100. Aggregate standard deviation rises from 10 to about 14.1. Expected loss does not double because of the covariance term. Dependence changes dispersion and tail behavior unless the common cause also changes the marginal frequencies or impacts.

The scaling connection becomes clearer for large n. If positive pairwise dependence remains roughly constant, the $n(n-1)r$ term can cause aggregate variance to approach quadratic growth. This equal-correlation model is too simple for real cyber portfolios, yet it makes one point visible: small shared dependence across many scenarios can materially change the enterprise tail.

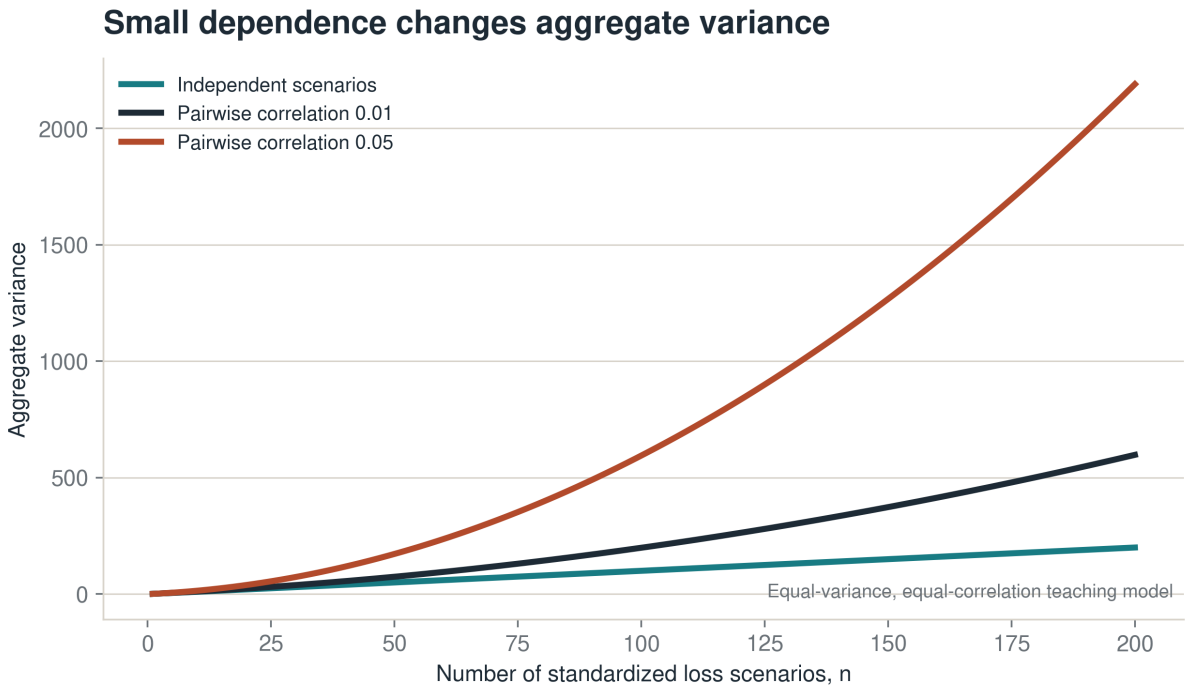


Figure 6. Small positive dependence across many standardized scenarios can increase aggregate variance much faster than an independence assumption.

Real analysis may require common-cause scenarios, dependency graphs, Bayesian networks, factor models, copulas, or Monte Carlo aggregation. Pairwise correlation provides an educational bridge. A complete model of cyber tail dependence needs more structure.

7.3 Concentration can create efficiency and fragility

Centralization often improves consistency and reduces marginal effort. The same decision can increase concentration. A common identity platform can improve access control across thousands of applications and become a major dependency. A cloud foundation can improve configuration and logging while increasing the blast radius of a privileged compromise. A managed service can create expertise and coverage while concentrating access and operational knowledge.

The answer is rarely to avoid shared services. The trade-off needs to be visible and managed. Useful measures include the share of important services dependent on one platform, the percentage of privileged paths crossing one control point, recovery resources shared across services, tested substitution time, and the number of independent barriers remaining after a common dependency fails.

NIST IR 8286 and IR 8286C describe how cybersecurity risk information can be integrated into enterprise risk management and governance oversight [17, 18]. Scaling analysis adds questions about

convergence, propagation, and whether aggregation assumes independence where common causes exist.

7.4 Attack paths require semantics

Graph thinking explains why modest connectivity changes can produce a much larger analysis problem. If each compromised state opens an average of b new reachable options and the analysis explores to depth d, a simple branching-tree approximation is:

$$b + b^2 + \cdots + b^d = \frac{b(b^d - 1)}{b - 1}, \quad b \neq 1$$

With $b = 3$ and $d = 5$, the tree contains 363 path prefixes across five levels. Real attack graphs contain cycles, repeated nodes, conditional permissions, defensive controls, and incomplete data. The calculation is a simplified teaching model, not a guaranteed upper bound. It shows why reducing branching, removing unnecessary transitive privilege, and creating choke points can matter more than reducing node count alone.

Edge count is insufficient. Cyber edges have direction, privilege, reachability, authentication conditions, trust boundaries, temporal availability, and business consequence. One persistent vendor path into an engineering environment can matter more than thousands of ordinary user-to-application relationships. A scaling study can reveal that relationships are accelerating. Threat modeling and graph semantics determine which relationships deserve action.

7.5 Integrating scaling with cyber risk quantification

Scaling analysis complements scenario-based quantification. It can help estimate how exposure drivers change, how many scenarios share a dependency, how control workload affects decision latency, and how concentration influences aggregate dispersion. It should not turn β into an annual loss number.

A practical integration follows this sequence:

1. Define a business-loss scenario.
2. Identify the relevant scale inputs and dependency structure.
3. Measure how those exposure drivers change with scale.
4. Estimate threat frequency and control performance using appropriate evidence.
5. Model business impact and recovery.
6. Aggregate scenarios with explicit dependence assumptions.
7. Test how structural interventions change exposure, control performance, or concentration.

This sequence keeps the scaling result in the role it can support. It informs the risk model without replacing it.

8. Leadership and Organizational Behavior

The leadership ideas in this section are organizational hypotheses and design implications. They are not consequences proven by the power-law equations. Some can be measured, and each organization still needs to test whether the relationship holds in its context.

8.1 Direct leadership does not scale automatically

Research on bounded rationality and social networks provides a foundation for thinking about limits on attention and relationships [7, 20, 21]. The leadership implication is practical: one person's time, memory, and context cannot expand with every new team, system, and decision. Personal involvement can improve quality in a small organization. As scale grows, the same involvement can create queues, dependency, and learned helplessness. The leader can become a concentrated dependency and a source of delay.

Leadership must then be carried by principles, decision rights, talent, trusted information, and feedback. Sound judgment needs to become available throughout the system instead of depending on one person's direct involvement.

Candidate measures include decisions awaiting senior approval, escalation rate, median decision latency, rework after delegation, and the share of decisions completed inside defined risk boundaries. The objective is appropriate escalation. Local decisions should be made where the necessary information and accountability exist.

8.2 Distributed authority changes effective capacity

Delegation is often described as a personal skill. At scale, it becomes an architectural requirement. Authority must be matched with competence, information, boundaries, and accountability. Delegating without context creates exposure. Retaining a decision when local leaders have the context creates delay.

Distributed authority may increase effective capacity by enabling parallel action and reducing approval delays. Whether capacity grows faster than headcount is an empirical question. It can fragment standards when shared principles and feedback are weak. The design challenge is bounded autonomy: local decisions inside clear risk limits, supported by enterprise capabilities and visible through common evidence.

A governance forum adds reusable capacity when a decision produces a principle, pattern, or boundary that resolves future cases. If every similar case returns for approval, the forum is processing transactions instead of building capacity.

8.3 Trust is a network property

Trust is often discussed as a relationship between two people. In an organization, it also depends on the network. People need confidence that information is reliable, commitments are visible, exceptions

are handled fairly, and leaders respond consistently. A culture based entirely on personal trust cannot scale because new employees and distant teams cannot inherit relationships they never experienced. Scalable trust uses transparent decision processes, clear ownership, dependable follow-through, and the ability to surface bad news without punishment. Documentation and governance support trust when they reduce uncertainty. They weaken it when they create control without context.

8.4 Culture and repeated local signals

Culture spreads through local signals: which risks receive attention, what happens after someone raises a concern, whether deadlines override stated principles, and how leaders respond under pressure. A formal policy is one signal among many. Informal interpretation moves through teams and can create wider effects.

A small number of visible leadership decisions can influence behavior across a large organization. The same mechanism can spread unhealthy norms. If teams learn that evidence is optional, exceptions are routine, or operational friction will be ignored, those expectations propagate. This is a design hypothesis worth observing rather than a claim that culture has a universal exponent.

8.5 Feedback loops make adaptation possible

A growing organization cannot rely on leaders seeing every problem directly. It needs feedback loops that detect a difference between expectation and reality, assign an owned adjustment, implement the change, and gather evidence that it worked. Incidents, assessments, audits, exceptions, architecture reviews, exercises, near misses, and service trends all provide signals.

Feedback quality matters more than report volume. Leaders need information that preserves weak signals, dependencies, local variation, and uncertainty. Metrics that aggregate only upward can hide the conditions that produced them. Local teams also need enterprise context so that optimization in one area does not create exposure elsewhere. A mature system moves information in both directions.

At small scale, leadership can be carried by personal knowledge. At larger scale, it must be carried by an operating system of principles, roles, evidence, relationships, and feedback.

9. How to Run a Scaling Review

Scaling analysis should earn its place by improving a real decision. It can support strategy, transformation, merger integration, cloud adoption, AI deployment, operating-model design, incident readiness, or a period of rapid growth. The work can begin before the organization has enough data to estimate an exponent. A clear boundary, relationship map, and operating hypothesis can already expose structural problems.

9.1 Step 1: state the decision

Begin with the action, investment, design, or governance choice at stake. “Does cybersecurity follow a power law?” is an interesting research question and a poor operating question. A better question is:

> *Should we integrate more applications into identity lifecycle automation, add processing capacity, redesign roles, or accept the current exception backlog?*

If a result of $\beta = 1.20$ rather than 1.00 would not change the decision, stop. The analysis has no operational purpose yet.

9.2 Step 2: draw the four-layer chain

Write the path from scale to risk in plain language. For identity:

> *Identities and applications create entitlement and privilege edges. Those relationships create certification, exception, and lifecycle work. Delayed or weak decisions can leave access active and create paths into important resources. Threat, control, business consequence, and recovery determine the risk.*

This sentence is the causal hypothesis. It prevents a relationship measured at one layer from being presented as a result at another.

9.3 Step 3: define *N* and *Y*

Document the population, boundary, unit, period, data source, inclusion rules, and known collection changes. Identify whether each measure is a stock, flow, rate, proportion, duration, cost, or loss. Do not mix a month-end inventory with annual work unless the model and period make the relationship meaningful.

The scale variable must fit the mechanism. Employee count may explain some identity demand. Application diversity or entitlement edges may explain more. Vendor count may explain assessment volume. Supplier-service edges may explain concentration. Asset count may explain scanning volume. Reachable paths may explain exposure.

9.4 Step 4: measure relationships before compressing them

Map edges, degree, privilege, reachability, centrality, and concentration when they matter. A single composite score can make the analysis easier to present and harder to defend. Analyze variables separately until a composite has evidence-based weights and a clear interpretation.

Ask which edges carry consequence. A supplier contract, an API integration, a trust relationship, a standing privilege, and an emergency access path are all connections. They are not interchangeable.

9.5 Step 5: measure demand and sustainable capacity

Convert work into compatible units. Case count may be sufficient when cases are similar. When they are not, use hours, complexity classes, case-type service rates, or separate models. Estimate sustainable capacity using the same quality definition and time period.

Record backlog and latency as outcomes of the operating system. A flat completion rate can hide rising demand when the team is capacity constrained. A low backlog can hide rejected work or reduced quality. The measure needs to reflect the service the organization intends to provide.

9.6 Step 6: inspect the data

Examine levels, per-unit values, original-scale plots, log-scale plots, missing data, zeros, outliers, and known breaks. Look at variation by business unit, architecture, and period. Determine whether one event or one very large unit drives the result.

Inspection often finds the real issue before regression begins. A sudden jump may be a scanner deployment. A lower alert rate may be a lost data source. A business unit above the curve may be the only one with an acquisition-era identity model.

9.7 Step 7: estimate and challenge

Fit the simplest plausible model and compare it with alternatives. Report the exponent with uncertainty, the useful range of N, residual diagnostics, sensitivity to definitions, held-out performance when available, and evidence of breakpoints.

The power-law model should compete with linear, saturating, exponential, piecewise, and multivariable explanations. Statistical significance does not establish operational importance. A very large dataset can make $\beta = 1.04$ statistically different from one while the expected difference across the organization's actual range remains minor.

9.8 Step 8: connect the intervention to a model term

State whether the proposed action should reduce N, reduce edge density, lower D_0, reduce β_D, increase K_0, increase β_K, move a breakpoint, reduce variance, or reduce common-mode concentration. This forces the proposal to explain its mechanism.

A tool purchase describes the intervention. "Reduce manual evidence gathering by reusing standardized control evidence across additional services" describes the mechanism. Hiring may be

exactly what an insufficient baseline needs. It becomes a temporary answer when demand continues to outscale the operating model.

9.9 Step 9: decide how the result will be verified

Before implementation, define the expected effect, owner, review date, data source, quality threshold, and conditions that would disprove the hypothesis. Measure the mechanism and the outcome. An automation project may reduce processing hours while creating more exceptions. A centralized platform may lower operating cost while increasing concentration. A complete verification plan needs to observe both.

The nine steps can be summarized as follows:

Step	Question	Output
1	What decision are we trying to improve?	Decision, owner, and horizon
2	How should scale connect to risk?	Four-layer causal hypothesis
3	What exactly are N and Y?	Definitions, units, boundary, and period
4	Which relationships carry the effect?	Typed dependency or graph view
5	What work arrives and what can be sustained?	Demand, capacity, pressure, backlog, and latency
6	What do the data show before modeling?	Level, per-unit, log, quality, and break inspection
7	Which model survives comparison?	Estimate, uncertainty, alternatives, and useful range
8	Which lever changes the mechanism?	Intervention tied to a model term
9	How will we know it worked?	Expected result, owner, evidence, and review date

10. Measuring Scaling in a Real Organization

10.1 Start with a narrow hypothesis

The first study should avoid explaining the entire enterprise. Choose one domain with reliable data and a plausible mechanism. Identity is a strong candidate because nodes and relationships can be measured. Third-party concentration, vulnerability workload, incident coordination, cloud-policy coverage, alert investigation, and risk-treatment capacity are other options.

A useful hypothesis identifies the population, scale variable, outcome, mechanism, and period. For example:

> *Across comparable business units, monthly manual access-exception hours grow faster than active workforce identities because application diversity, non-human access, and incomplete lifecycle integration increase entitlement complexity.*

The wording makes the claim challengeable. It also tells the team which additional variables may explain the relationship.

10.2 Build a panel instead of relying on one snapshot

Cross-sectional data compare units of different sizes. Longitudinal data show how the same unit changes. Both are useful, and they answer different questions. A relationship across business units today may not predict the history of one unit as it grows [6].

Preserve business-unit or system identity and collect repeated observations. Monthly data for 24 to 36 months across several comparable units can be a practical starting point. This is a planning guideline rather than a statistical rule. Required sample size depends on variation, noise, effect size, model complexity, and data quality.

A wider range of N is often more useful than many observations clustered at almost the same size. If N barely changes inside a unit, a fixed-effects model has little information about the within-unit relationship. Repeated monthly observations are also correlated. Three years of data from eight units is not equivalent to 288 independent organizations [28].

Mark mergers, platform changes, new regulations, major incidents, measurement changes, and control redesigns. They can create structural breaks and should not be absorbed into one exponent without inspection.

10.3 Build a data dictionary before a model

The data dictionary should record:

Field	What to document
Population	Business units, systems, periods, and object definitions included

Field	What to document
Scale variable N	Positive quantity, unit, source, and reason it represents scale
Outcome Y	Stock, flow, count, rate, proportion, duration, cost, or loss
Demand	Incoming quality-adjusted work per period
Capacity	Maximum sustainable service capacity using the same work definition
Relationships	Edge type, direction, privilege, reachability, and consequence labels
Moderators	Architecture, control coverage, threat conditions, business mix, maturity
Break indicators	Acquisitions, migrations, incidents, policy or measurement changes
Quality	Missingness, duplicates, known blind spots, and definition changes

Version the definitions. If an identity changes from "employee account" to "all workforce and non-human identities," the series has changed even if the dashboard name has not.

10.4 Inspect before estimating

Review definitions, missing values, duplicates, and changes in collection. Plot the distributions and ranges of N and Y. Compare levels and per-unit values. Use original and logarithmic axes. Examine units, architectures, periods, zeros, bounds, and possible thresholds.

A straight line on log-log axes is evidence worth examining. It is not proof. Log axes compress deviations, and several functions can appear similar over a limited range [5, 10, 26].

10.5 Estimate with the assumptions visible

For positive N and Y, a basic panel model is:

$$\ln(Y_{it}) = \alpha + \beta \ln(N_{it}) + \gamma' X_{it} + u_i + \lambda_t + \varepsilon_{it}$$

i identifies the business unit or system, t identifies time, X contains observed moderators, u_i captures persistent unit differences, λ_t captures shared time shocks, and ε_{it} is unexplained variation.

Ordinary least squares on logarithms estimates a conditional relationship on the log scale and implies multiplicative error on the original scale. Nonlinear regression on the original scale implies a different error structure. Neither is automatically correct. The fitting method should follow how variability behaves, and predictions transformed to the original scale may require a retransformation correction [26].

Counts, proportions, durations, and losses may need different models. A log transform cannot accept zero. Adding an arbitrary constant changes the question and may distort the estimate. Count models can handle zeros directly. Bounded proportions may need binomial or beta models. Heavy-tailed losses require robust or distribution-specific methods.

10.6 Account for dependence, hierarchy, and measurement error

Panel observations are rarely independent. Months from the same business unit share architecture, leadership, and data processes. Standard errors should account for clustering or serial correlation when inference matters. With few units, uncertainty can remain large even when the row count looks impressive [28].

Organizations also respond to demand. A team may hire after backlog rises. Capacity may therefore be partly caused by observed demand, which complicates causal interpretation. Architecture maturity can influence both scale and outcomes. Measurement improvements can raise findings while lowering actual exposure. These are not reasons to avoid modeling. They are reasons to describe the result as an association until a stronger design supports a causal claim [29].

Measurement error deserves particular attention in cybersecurity. CMDB completeness, identity reconciliation, supplier ownership, log coverage, and incident classification are seldom perfect. If larger units have better measurement, apparent superlinearity may reflect visibility. If smaller units underreport work, scale-adjusted comparisons can become unfair.

10.7 Challenge the functional form

Compare the candidate model with alternatives that match plausible mechanisms:

- **Linear.** Each additional unit creates a constant amount of Y.
- **Logarithmic or saturating.** Growth slows as reuse or a natural limit increases.
- **Exponential.** The growth rate in Y depends on the level of Y as N changes.
- **Piecewise.** The relationship changes after a threshold or operating redesign.
- **Multivariable.** Several scale drivers contribute separately.
- **No stable relationship.** Architecture, maturity, or context explains more than scale.

Model comparison should consider performance on the original business quantity, interpretability, and held-out prediction alongside fit. A statistically elegant model that predicts monthly work poorly is not useful for capacity planning.

10.8 Treat deviations as information

The average relationship is a baseline. A unit below the expected demand curve may have stronger modularity, automation, leadership, or architecture. A unit above it may carry legacy complexity, unusual exposure, or a data problem. The deviation can be more useful than the exponent because it points toward a condition that can be investigated.

Interpretation requires context. A high finding rate may reflect stronger visibility. A high control workload may reflect greater business exposure. A low incident count may reflect weak detection. Treat residuals as questions. Using them as automatic performance grades removes the context the analysis is trying to preserve.

10.9 Evaluate interventions with a design that can support the claim

Pre- and post-implementation comparisons are easy to produce and vulnerable to false attribution. Demand may have changed for unrelated reasons. A major incident may alter behavior. A broader measurement program may increase the visible workload.

Where practical, use staggered rollouts, comparison units, interrupted time-series designs, or difference-in-differences analysis [29, 30]. Record the mechanism before the change. Measure whether the baseline, slope, breakpoint, queue, or concentration moved. Short evaluations may support a level-change claim while remaining unable to show that an exponent changed.

10.10 Know when to stop

Do not continue fitting until a preferred story appears. Stop or narrow the claim when:

- Definitions are unstable.
- The scale range is too narrow.
- One unit or event drives the result.
- Uncertainty spans materially different operating regimes.
- A simpler model predicts as well.
- Residuals reveal missing structure.
- Held-out performance is poor.
- The fitted result has no decision attached to it.

The objective is a better decision. The exponent earns attention only when it contributes to one.

11. A Worked Identity Case

This case is synthetic. It illustrates the reasoning from a decision through model comparison to an intervention and verification plan. The reported values do not describe a real company and should not be used as a benchmark. The dataset and analysis code are not included with this paper, so the reported estimates should be read as teaching outputs rather than independently reproducible research findings.

11.1 The decision

An enterprise identity team is deciding whether to add staff, accelerate application integration with identity governance, redesign roles, and place additional controls around non-human identities. Leaders know that manual work is increasing. They do not know whether growth comes mainly from the identity population, application diversity, limited automation, or local operating differences.

The decision statement is:

> *Which changes will keep manual identity-governance demand inside sustainable capacity as the organization grows?*

11.2 The data

The synthetic dataset contains 288 monthly observations from eight business units over 36 months. It includes active identities, applications, identity-governance integration coverage, non-human identity share, manual work hours, and sustainable capacity hours. The final six months from every unit are held out for model comparison.

Variable	Definition
Active identities	Workforce identities active at month end
Applications	Applications in the unit's identity-governance boundary
IGA coverage	Share of applications using the standard lifecycle process
Non-human share	Non-human identities as a share of total managed identities
Demand hours	Monthly manual identity work, adjusted for case type
Capacity hours	Maximum sustainable monthly service capacity at the quality threshold

The data were generated with known relationships and noise. Units differ in application density and initial automation. Two acquired units begin with lower IGA coverage. The case therefore contains a scaling signal and structural differences that a single-variable model should struggle to explain.

11.3 First inspection

The original and log-scale plots show demand rising with active identities. Several units form visible bands. Industrial and acquired units sit above the overall relationship, while retail and central units sit below it. That pattern already suggests that scale is not the whole explanation.

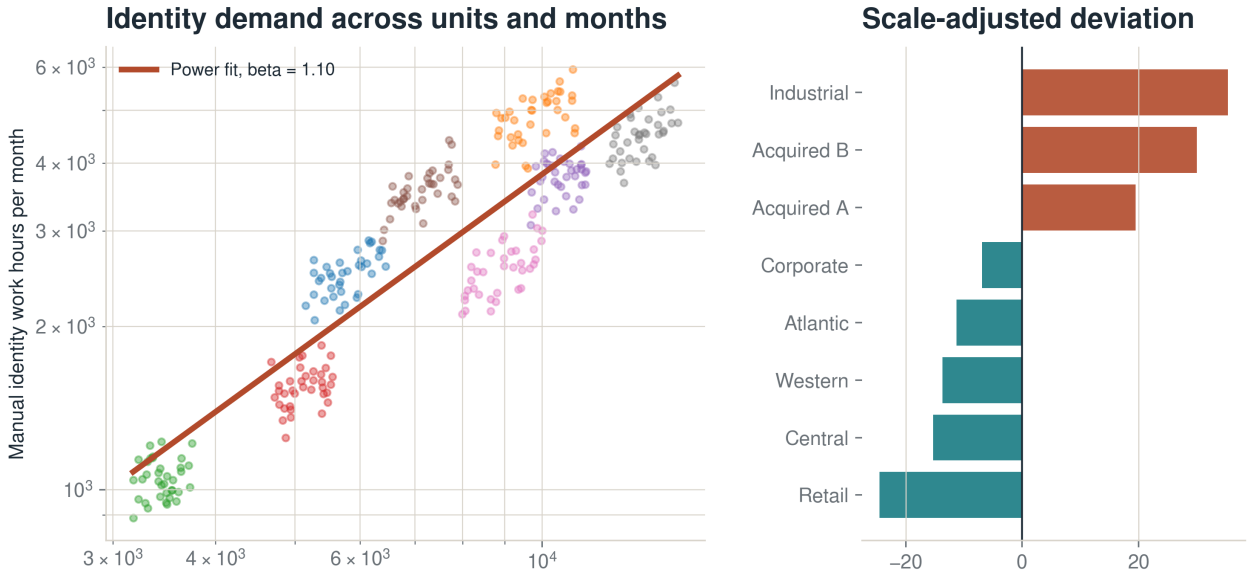


Figure 7. Synthetic identity demand and average scale-adjusted deviations. Left horizontal axis: active identities. Right horizontal axis: average difference from expected demand (%). Positive deviations indicate more manual demand than the simple scale model expects.

The simple unadjusted power model estimates $\beta_D = 1.10$. A row count of 240 training observations might make that estimate look stable. A bootstrap that resamples whole business units produces a much wider teaching interval, approximately 0.58 to 1.42. With eight business units, the study cannot confidently distinguish sublinear, linear, and superlinear behavior across the units represented. It cannot establish a relationship across independent organizations.

That is an important result. Repeated months do not create hundreds of independent units [28].

11.4 Model comparison

Three models are compared on the final six months:

Candidate	Held-out RMSE in monthly demand hours	Reading
Simple linear model	694	Reasonable baseline
Simple power model	724	About 4% worse than the linear model
Adjusted log model	286	About 59% better than the linear model

The simple power law does not earn adoption. Its estimate is mildly superlinear, and its held-out prediction is worse than the linear baseline. The adjusted model includes application density, IGA coverage, and non-human identity share. It predicts much better in the reported comparison after adding variables that represent mechanisms left out of the scale-only model. This comparison changes both the predictors and the functional form; it does not establish that a log model is better than an adjusted linear model using the same information.

The reported adjusted coefficient for identities is about 1.11. Its interpretation depends on the exact definition and transformation of application density and the other moderators. If application density means applications per identity, holding it fixed while identities grow also implies more applications. The model still describes association. The synthetic data were generated with a relationship, while a real organization would need stronger evidence before treating the coefficient as causal.

11.5 Demand, capacity, and local pressure

The unadjusted capacity relationship is estimated at $\beta_K = 0.95$. The difference between the simple demand and capacity estimates is approximately 0.15. If both unadjusted power relationships were valid, the point estimates would imply a pressure multiplier per doubling of:

$$2^{0.15} \approx 1.11$$

The conditional calculation is about 11 percent per doubling. The demand model performed worse than the linear baseline and its reported exponent interval is wide, so this is not a supported pressure forecast. The arithmetic is less dramatic than the earlier illustrative $\Delta\beta = 0.30$ case, but the uncertainty matters more than the difference between those point estimates. A planning forecast should use the demand and capacity models that survive comparison and carry their joint uncertainty into the pressure ratio.

Average pressure across all synthetic observations is about 0.81. The average hides meaningful unit differences. Retail, central, and Atlantic units retain material reserve. Industrial and Acquired B are already above sustainable capacity. Acquired A is close to the limit.

The residual pattern and pressure measure point in the same direction. The first intervention should not be an enterprise-wide staffing ratio. The acquired and industrial units need investigation into application density, lifecycle coverage, role design, and non-human access. A general capacity increase may still be warranted, although it should not be the only action.

A conditional risk scenario makes the operating result more useful. Suppose the growing queue includes removal of privileged access after a person changes role or leaves the organization. Longer delay can leave a consequential access relationship active for longer. That is an exposure hypothesis; the additional work hours do not, by themselves, establish a higher probability of compromise.

The review would connect the removal request, completion time, identity, privilege, reachable resource, and business service. It would then examine whether credentials or sessions remain usable, which preventive and detective controls apply, and whether the access provides a credible route to a consequential action. A scenario could involve misuse of still-active privileged access to alter an important service. Threat opportunity, control effectiveness, business consequence, response, and recovery determine what the delay means for that scenario.

Verification should therefore track both the queue and the access condition: time to remove access, aged privileged removals, orphaned access, quality failures, and tested reachability to important resources. A lower demand baseline is useful only if the consequential access conditions also improve or remain within the agreed boundary. No loss frequency or financial impact is estimated here because the synthetic case does not supply evidence for those quantities.

11.6 The intervention plan

The team chooses four linked actions:

1. Integrate the highest-volume acquired applications with the standard lifecycle process.
2. Redesign roles for the application clusters producing repeated manual exceptions.
3. Establish ownership and lifecycle controls for non-human identities.
4. Add temporary processing capacity to reduce the current backlog while the structural work is completed.

The actions target different model terms. Temporary staffing increases K_0. Application integration lowers D_0 and may improve β_K if the standard process can serve additional scale. Role design reduces edge density and exception demand. Non-human identity governance reduces unmanaged relationship growth and high-consequence exposure.

11.7 Verification

The evaluation records pre-intervention definitions and uses unaffected units as comparisons where possible. The team measures manual hours, exception volume, backlog age, quality failures, entitlement density, application coverage, and local pressure.

After six months, the team may be able to support a claim that the demand baseline changed. Showing that the scaling exponent improved will require more variation and time. If demand hours fall while quality failures or orphaned access rise, the intervention has shifted work out of view rather than improving the system.

The case demonstrates the intended behavior of the method. The exponent begins the investigation. Model comparison prevents the team from forcing a nonlinear story. Residuals locate operating differences. Mechanism variables explain more than scale alone. The final plan combines immediate capacity with structural redesign.

To make the case reproducible, a companion analysis would need to publish the synthetic data or generator, random seed, generation equations and noise assumptions, data dictionary, exact model formulas, coefficients, and training and test assignments. It would also need to state how original-scale predictions were recovered and how the unit-level bootstrap interval was calculated. The reported estimates should not be used as a benchmark until that package is available and the results can be reproduced.

A stronger comparison would give the competing models the same predictors, report forecast error by unit and distinguish prediction for future months in known units from prediction for a new unit. Any intervention evaluation would then test the stated mechanism and service quality. The verification plan above describes that next step; it does not report a completed intervention or a demonstrated change in the exponent.

12. Limits and Responsible Use

Power laws are a lens for detecting structure. They are not permission to replace evidence, context, or judgment with a curve.

Use four labels consistently. **Observed** values come from a defined source. **Estimated** values come from a stated model with uncertainty and assumptions. **Illustrative** values explain a mechanism. **Synthetic** data are generated to demonstrate a method and do not describe a real organization.

Language should match the evidence. “The outcome increased superlinearly across this sample and range” is more defensible than “cyber risk follows a universal power law.” State the population, period, units, range, uncertainty, and plausible alternatives.

Limit	Why it matters
No universal cyber exponent	Domains, architectures, businesses, and adversaries differ. Each relationship must be measured.
Scale depends on definitions	Changing an identity, application, supplier, or incident boundary can change the result.
Association is not mechanism	A curve can summarize data without explaining why N and Y move together.
The useful regime can change	Comparing large and small units does not predict how one unit will evolve; acquisitions, thresholds, and redesigns can create breakpoints.
Fitting and inference have assumptions	Zeros, bounds, error structure, outliers, limited range, and dependent observations can mislead.
Workload is not risk	Findings, alerts, exceptions, and effort do not equal expected loss or tail exposure.
Capacity is not completed work	Observed throughput can hide reserve, rejected demand, or reduced quality.
The system adapts	Adversaries respond to controls, while leadership, incentives, architecture, and learning change the relationship.
Metrics change behavior	Measures used as targets can be optimized in ways that weaken the objective.

The framework cannot predict attack timing, turn asset count directly into expected loss, prove causation from an observed change, or produce credible rare-loss forecasts from a small internal sample. It can support bounded workload forecasts, pressure scenarios, scale-adjusted comparisons, breakpoint and concentration analysis, and intervention tests. Keep prediction local to the defined population and observed range. Acquisition, architecture, attacker, or regulatory shifts can invalidate the relationship.

13. Building a Scaling Capability

An organization does not need to begin with a statistical program. The capability can mature in stages.

13.1 Stage one: a structured review

Use the one-page canvas in Appendix C. Define the decision, boundary, scale, outcome, mechanism, dependencies, and intended intervention. Map nodes and consequential relationships. This stage can improve architecture and governance decisions without estimating β.

13.2 Stage two: descriptive analysis

Build level and per-unit trends. Calculate descriptive two-point slopes only when they help frame a question, and label them accordingly. Track demand, sustainable capacity, pressure, backlog, and latency with stable definitions.

13.3 Stage three: model comparison

Collect a panel across units and time. Fit power, linear, piecewise, and multivariable alternatives. Report uncertainty, useful range, residuals, and held-out performance. Bring a statistician or quantitative analyst into the work when inference affects material investment or risk decisions.

13.4 Stage four: intervention learning

Use comparison units, staggered implementation, or interrupted time-series designs. Determine whether the baseline, slope, breakpoint, queue, or concentration changed. Preserve negative results. A tool that did not improve scaling is still useful evidence.

13.5 Stage five: portfolio and cross-organization research

Connect dependency maps with risk scenarios, common causes, and recovery concentration. Replicate relationships across organizations using common definitions and confidentiality protections. Expect families of relationships rather than one cyber exponent.

13.6 Tooling

A practical workbook or web tool can accept unit and time observations, preserve the data dictionary, show level, per-unit, and log views, compare plausible models, calculate multipliers, display residuals, simulate backlog, record interventions, and flag weak ranges. It should distinguish a two-point slope from a statistical estimate and expose assumptions and uncertainty. The tool supports judgment; it should not collapse the result into a traffic-light score.

14. Research Agenda

The framework creates questions that can be tested inside one enterprise and later compared across organizations. The first objective should be practical learning.

The initial questions are empirical. Which cybersecurity variables show repeatable scaling relationships? Identity-entitlement edges, manual exceptions, privileged paths, alert investigation, vulnerability workload, treatment decisions, supplier dependencies, and incident coordination are candidates. Which scale variables explain them best? Workforce, application diversity, external connections, privilege edges, and business-service dependencies may tell different stories.

The next questions concern architecture. Do standardized platforms, segmentation, role engineering, policy as code, automation, and federated governance change the exponent or mainly change the baseline? Where do breakpoints appear? Does a process scale well until geographic spread, application diversity, regulatory complexity, or coupling crosses a threshold?

Leadership creates another research area. Organizations can compare decision latency, escalation volume, rework, and risk outcomes before and after changes in delegated authority, team structure, governance, and information flow. This would test whether distributed judgment increases effective capacity without fragmenting control quality.

Risk aggregation needs its own evidence. Dependency maps can be connected to scenarios to find concentration hidden by traditional registers. Research can compare expected loss, variance, tail loss, recovery time, and common-cause exposure rather than compressing them into one score.
Artificial intelligence may change demand and capacity simultaneously. Agentic systems can reduce coordination cost in some workflows while increasing tool, data, permission, and non-human identity relationships in others. Research should measure assurance, monitoring, response, and human-review cost as part of the result.

A multi-organization program would need common definitions, confidentiality protections, sector controls, and separation between operating metrics and loss outcomes. The result may be a family of exponents, mechanisms, breakpoints, and failure conditions. That would help practitioners see which parts of cybersecurity gain efficiency with scale, which become harder, and which designs change the relationship. The empirical principle is straightforward: search for repeatable relationships, mechanisms, deviations, and breakpoints that help redesign the system before overload becomes failure.

Conclusion

Scaling theory is attractive because it shows how much can change when size changes. A mouse and an elephant do not run the same biological clock. A city is not a town with more residents. A large cybersecurity program cannot be operated as a small program with more people, meetings, and tools.

Some security capabilities become more efficient as organizations grow. Shared platforms, automation, and standardized services can reduce marginal effort. Other demands accelerate. Relationships, dependencies, privileges, coordination, and common-mode exposure may grow faster than the components being counted. Human attention and centralized judgment do not expand at the same rate.

The cybersecurity scaling model separates these effects. Scale inputs shape relationship structure. Relationship structure creates operating demand. Sustainable capacity determines whether work becomes a queue and whether latency rises. Threats, controls, business consequence, response, and recovery determine what the structure means for risk. This sequence keeps an appealing mathematical result in one layer from being presented as proof in another.

When demand grows faster than capacity, effort alone will not close the gap for long. The organization may need to simplify, reduce unnecessary relationships, create modular boundaries, automate repeatable work, distribute authority, and build feedback that preserves context. Shared platforms and central services can remain part of the answer when their concentration is understood and managed.

Power laws will not fit every environment, period, or outcome. That is expected. The model remains useful when it reveals a linear relationship, a breakpoint, a dominant hub, a measurement change, or evidence too weak to support a claim. Learning where the model fails is part of learning how the system works.

The practical question is larger than whether the program performs today. It is whether the architecture of performance can survive the next order of magnitude.

Appendix A. Worked Mathematical Examples

A.1 Doubling under different exponents

For $Y = Y_0(N/N_0)^{\beta}$, doubling N multiplies Y by 2^{β}. With $\beta = 0.80$, the multiplier is 1.74 and Y per unit falls by about 13 percent. With $\beta = 1.20$, the multiplier is 2.30 and Y per unit rises by about 15 percent.

A.2 Estimating an exponent from two points

For two observations:

$$\beta = \frac{\ln(Y_2/Y_1)}{\ln(N_2/N_1)}$$

If N doubles and Y increases from 120,000 to 300,000, $\beta = \ln(2.5)/\ln(2)$, or approximately 1.32. This is a descriptive two-point slope. It has no confidence interval and cannot show whether the exponent is stable.

A.3 Tenfold growth and the scaling gap

Assume demand scales with $\beta_D = 1.25$ and capacity with $\beta_K = 0.90$. The difference $\Delta\beta$ is 0.35. A tenfold increase in N multiplies operating pressure by $10^{0.35}$, or about 2.24. A hundredfold increase multiplies it by $100^{0.35}$, or about 5.01.

If the original pressure ratio is 0.60 at N_0, tenfold growth raises it to about 1.34 under the illustrative exponents. Demand then exceeds sustainable capacity and backlog accumulates unless the relationship changes.

A.4 Estimating the operating limit

If $\rho_0 = 0.80$ and $\Delta\beta = 0.30$:

$$\frac{N^*}{N_0} = \left(\frac{1}{0.80}\right)^{1/0.30} \approx 2.10$$

The estimate means pressure reaches one at about 2.1 times the reference scale if the baseline, exponents, and operating regime remain valid.

A.5 Converting a scale-adjusted residual

If the observed value is 1,220 and the model expects 1,000:

$$r = \ln(1{,}220) - \ln(1{,}000) = \ln(1.22) \approx 0.20$$

The percentage difference is $e^{0.20} - 1$, or about 22 percent above the scale-adjusted expectation.

A.6 Relationship growth

A complete network of 100 nodes has 4,950 possible pairs. A complete network of 1,000 nodes has 499,500, slightly more than 100 times as many pairs even though node count increased tenfold.

For a sparse undirected network, use $E = N\bar{k}/2$. If 100 nodes have average degree 8, $E = 400$. If 1,000 nodes have average degree 20, $E = 10{,}000$. Growth comes from more nodes and increasing average degree. Identity graphs are usually bipartite or directed, so their edge formula and interpretation must match that structure.

A.7 Candidate attack paths

With branching factor $b = 3$ and depth $d = 5$:

$$3 + 3^2 + 3^3 + 3^4 + 3^5 = 363$$

The value counts path prefixes across levels in a tree with no cycles or repeated nodes. It is explanatory rather than predictive.

A.8 Correlated loss

With $n = 100$ standardized losses, $\sigma^2 = 1$, and equal correlation $r = 0.01$:

$$\mathrm{Var}(\textstyle\sum L_i) = 100[1 + 99(0.01)] = 199$$

Standard deviation is $\sqrt{199} \approx 14.1$, compared with 10 under independence. Expected aggregate loss is unchanged by the covariance term alone.

Appendix B. Candidate Metrics

Domain	Metric	Definition
Complexity	Applications per business service	Applications directly supporting each important service
Complexity	Integration edge density	Active technical relationships divided by possible or relevant relationships
Identity	Entitlements per identity	Active identity-resource edges, separated by privilege and identity type
Identity	Orphaned non-human identities	Service accounts, keys, agents, and tokens without accountable ownership
Identity	Lifecycle automation coverage	Share of applications using the standard joiner, mover, and leaver process
Third party	Dependency concentration	Share of important services dependent on the same supplier or platform
Third party	Tested substitution time	Time required to replace or operate without a common provider
Architecture	Trust-zone crossings	Approved pathways crossing security or operational boundaries
Architecture	High-consequence reachability	Paths from lower-trust origins to important services or processes
Operations	Exception rate	Control exceptions per defined unit and period
Operations	Decision latency	Time from identified decision need to accountable disposition
Operations	Operating pressure	Incoming quality-adjusted work divided by sustainable capacity
Operations	Backlog age	Age distribution of work awaiting decision or treatment
Incident response	Coordination load	Workstreams, handoffs, decision forums, and unresolved information conflicts
Risk	Common-cause concentration	Scenarios sharing suppliers, platforms, identities, controls, or recovery dependencies

Domain	Metric	Definition
Capacity	Reusable-control coverage	Share of the environment governed through standardized, automated controls
Leadership	Escalation dependency	Decisions requiring approval above the accountable operating level
Learning	Closed-loop improvement rate	Lessons implemented and verified for effectiveness
Resilience	Recovery concentration	Important services sharing recovery resources, platforms, or personnel
AI	Agent-tool edge density	Authorized tool relationships per agent, separated by consequence and privilege

Appendix C. One-Page Scaling Review Canvas

Field	Practitioner entry
Decision to improve	What action, investment, design, or governance choice is at stake?
Population and boundary	Which units, systems, periods, and object definitions are included?
Scale variable N	What positive quantity represents scale, and why?
Outcome Y	Is it a stock, flow, count, rate, proportion, duration, cost, or loss?
Four-layer chain	How should scale create relationships, work, and risk consequence?
Relationship semantics	Which edges, directions, privileges, paths, or dependencies matter?
Demand	What quality-adjusted work arrives per period?
Sustainable capacity	What can the current system process at the intended quality, and what reserve is needed for the service target?
Moderators	Architecture, controls, threat conditions, business mix, maturity, and measurement changes
Current evidence	Data range, frequency, missingness, sources, and known breaks
Candidate models	Power, linear, saturating, exponential, piecewise, multivariable, or no stable relationship
Estimated relationship	β, uncertainty, useful range, alternatives, and held-out performance
Deviations and breakpoints	Which units or periods differ, and what might explain them?
Intervention	Which model term or mechanism should change?
Verification	Expected effect, owner, quality threshold, review date, and disconfirming evidence

Appendix D. Analysis and Validation Checklist

Before presenting a scaling result, confirm:

- The decision and hypothesis are stated.
- N and Y have clear definitions, units, boundaries, and periods.
- Stocks, flows, rates, proportions, costs, and losses are not mixed without a reason.
- Demand and sustainable capacity use compatible quality-adjusted work units.
- Capacity is not inferred only from observed completions.
- Level, per-unit, original-scale, and log-scale views have been inspected.
- Architecture, maturity, threat conditions, coverage, and measurement changes have been considered.
- Edge direction, privilege, reachability, and consequence are preserved where relevant.
- Linear, nonlinear, multivariable, and piecewise alternatives have been compared.
- β is reported with uncertainty and a useful range.
- Dependence among repeated observations is reflected in the uncertainty analysis.
- Cross-sectional results are not presented as longitudinal predictions.
- Workload, exposure, expected loss, variance, and tail risk remain distinct.
- Observed, estimated, illustrative, and synthetic values are labeled.
- Residuals are investigated as questions rather than used as automatic rankings.
- The proposed intervention identifies the mechanism and model term it should change.
- The evaluation can distinguish a baseline change from a changed scaling relationship.
- A date, owner, quality threshold, and evidence standard exist to verify the result.
- Conditions that would cause the team to abandon or narrow the claim are recorded.

References

1. Albert, R., Jeong, H., and Barabási, A.-L. “Error and Attack Tolerance of Complex Networks.” *Nature* 406 (2000): 378-382. https://doi.org/10.1038/35019019.
2. Arcaute, E., Hatna, E., Ferguson, P., Youn, H., Johansson, A., and Batty, M. “Constructing Cities, Deconstructing Scaling Laws.” *Journal of the Royal Society Interface* 12 (2015): 20140745. https://doi.org/10.1098/rsif.2014.0745.
3. Bettencourt, L. M. A. “The Origins of Scaling in Cities.” *Science* 340, no. 6139 (2013): 1438-1441. https://doi.org/10.1126/science.1235823.
4. Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C., and West, G. B. “Growth, Innovation, Scaling, and the Pace of Life in Cities.” *Proceedings of the National Academy of Sciences* 104, no. 17 (2007): 7301-7306. https://doi.org/10.1073/pnas.0610172104.
5. Clauset, A., Shalizi, C. R., and Newman, M. E. J. “Power-Law Distributions in Empirical Data.” *SIAM Review* 51, no. 4 (2009): 661-703. https://doi.org/10.1137/070710111.
6. Depersin, J., and Barthelemy, M. “From Global Scaling to the Dynamics of Individual Cities.” *Proceedings of the National Academy of Sciences* 115, no. 10 (2018): 2317-2322. https://doi.org/10.1073/pnas.1718690115.
7. Dunbar, R. I. M. “Neocortex Size as a Constraint on Group Size in Primates.” *Journal of Human Evolution* 22, no. 6 (1992): 469-493. https://doi.org/10.1016/0047-2484(92)90081-J.
8. Galbraith, J. R. “Organization Design: An Information Processing View.” *Interfaces* 4, no. 3 (1974): 28-36. https://doi.org/10.1287/inte.4.3.28.
9. Kleiber, M. “Body Size and Metabolism.” *Hilgardia* 6, no. 11 (1932): 315-353. https://doi.org/10.3733/hilg.v06n11p315.
10. Leitão, J. C., Miotto, J. M., Gerlach, M., and Altmann, E. G. “Is This Scaling Nonlinear?” *Royal Society Open Science* 3 (2016): 150649. https://doi.org/10.1098/rsos.150649.
11. Little, J. D. C. “A Proof for the Queuing Formula: $L = \lambda W$.” *Operations Research* 9, no. 3 (1961): 383-387. https://doi.org/10.1287/opre.9.3.383.
12. MITRE. “MITRE ATT&CK.” Accessed July 30, 2026. https://attack.mitre.org/.
13. National Institute of Standards and Technology. *The NIST Cybersecurity Framework (CSF) 2.0*. NIST CSWP 29, 2024. https://doi.org/10.6028/NIST.CSWP.29.
14. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1, 2024. https://doi.org/10.6028/NIST.AI.600-1.
15. National Institute of Standards and Technology. “Lessons Learned from the Consortium: Tool Use in Agent Systems.” August 5, 2025. https://www.nist.gov/news-events/news/2025/08/lessons-learned-consortium-tool-use-agent-systems.
16. Perrow, C. *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books, 1984.
17. Quinn, S. D., et al. *Integrating Cybersecurity and Enterprise Risk Management*. NIST IR 8286 Revision 1, 2025. https://doi.org/10.6028/NIST.IR.8286r1.
18. Quinn, S. D., et al. *Staging Cybersecurity Risks for Enterprise Risk Management and Governance Oversight*. NIST IR 8286C Revision 1, 2025. https://doi.org/10.6028/NIST.IR.8286Cr1.
19. Saramäki, J., et al. “Persistence of Social Signatures in Human Communication.” *Proceedings of the National Academy of Sciences* 111, no. 3 (2014): 942-947. https://doi.org/10.1073/pnas.1308540110.

20. Simon, H. A. “A Behavioral Model of Rational Choice.” *Quarterly Journal of Economics* 69, no. 1 (1955): 99-118. https://doi.org/10.2307/1884852.
21. Tamarit, I., Cuesta, J. A., Dunbar, R. I. M., and Sánchez, A. “Cognitive Resource Allocation Determines the Organization of Personal Networks.” *Proceedings of the National Academy of Sciences* 115, no. 33 (2018): 8316-8321. https://doi.org/10.1073/pnas.1719233115.
22. Watts, D. J., and Strogatz, S. H. “Collective Dynamics of ‘Small-World’ Networks.” *Nature* 393 (1998): 440-442. https://doi.org/10.1038/30918.
23. West, G. B. *Scale: The Universal Laws of Life, Growth, and Death in Organisms, Cities, and Companies*. New York: Penguin Press, 2017.
24. West, G. B., Brown, J. H., and Enquist, B. J. “A General Model for the Origin of Allometric Scaling Laws in Biology.” *Science* 276, no. 5309 (1997): 122-126. https://doi.org/10.1126/science.276.5309.122.
25. White, C. R., et al. “Metabolic Scaling Is the Product of Life-History Optimization.” *Science* 377, no. 6608 (2022): 834-839. https://doi.org/10.1126/science.abm7649.
26. Xiao, X., White, E. P., Hooten, M. B., and Durham, S. L. “On the Use of Log-Transformation vs. Nonlinear Regression for Analyzing Biological Power Laws.” *Ecology* 92, no. 10 (2011): 1887-1894. https://doi.org/10.1890/11-0538.1.
27. Shortle, J. F., Thompson, J. M., Gross, D., and Harris, C. M. *Fundamentals of Queueing Theory*. 5th ed. Hoboken, NJ: John Wiley & Sons, 2018. https://doi.org/10.1002/9781119453765.
28. Cameron, A. C., and Miller, D. L. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources* 50, no. 2 (2015): 317-372. https://doi.org/10.3368/jhr.50.2.317.
29. Shadish, W. R., Cook, T. D., and Campbell, D. T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin, 2002.
30. Lopez Bernal, J., Cummins, S., and Gasparrini, A. “Interrupted Time Series Regression for the Evaluation of Public Health Interventions: A Tutorial.” *International Journal of Epidemiology* 46, no. 1 (2017): 348-355. https://doi.org/10.1093/ije/dyw098.
31. Cox, L. A., Jr. “What’s Wrong with Risk Matrices?” *Risk Analysis* 28, no. 2 (2008): 497-512. https://doi.org/10.1111/j.1539-6924.2008.01030.x.

About the Author

Carlos Morales is a cybersecurity executive, risk-management practitioner, educator, and author. His work focuses on translating technical complexity into business decisions, developing adaptive cybersecurity leadership, and building security programs that connect governance, risk, architecture, resilience, and operations. He is the author of *Cyber Risk Management in Practice* and co-author of forthcoming books on adaptive cybersecurity leadership and strategic cybersecurity for business leaders.

Suggested citation

Morales, C. (2026). Cybersecurity at Scale: Why Risk, Coordination, and Leadership Stop Growing Linearly. Practitioner field paper, Version 4.0, November 2026.