

A Graph-Based Framework for Analyzing Bias in Effort Estimation under Conditions of Uncertainty

Massimo de Falco ✉

University of Salerno

Marco Vitolo ✉

University of Salerno

Nicola Mastrandrea ✉

University of Salerno

Abstract

Effort estimation under uncertainty is a decision-support problem affected by recurrent informational, methodological, cognitive, and organizational factors. Although these factors have been extensively documented in the literature, they are typically considered as separate sources of estimation risk. This perspective makes it difficult to understand how estimation problems interact, where distortions accumulate within the overall estimation process, and which points may offer the greatest leverage for intervention.

This paper proposes a framework for *effort estimation problems analysis* that represents these issues as an interconnected system. The framework consists of three components. First, a four-domain taxonomy of 16 recurrent estimation problems, covering problem definition, methods and models, human factors, and organizational context, whose completeness is assessed through a falsifiable generate-and-discard procedure. Second, a directed acyclic graph $G = (V, E)$ connecting the 16 problems, with 19 edges representing 12 literature-documented dependency connections and 7 interactions inferred from interaction analysis; the two types of connections are kept analytically distinct. Third, two asymmetric, relative-normalised centrality indices, I_{inf} (influence) and I_{vuln} (vulnerability), which characterise the role of each problem in the dependency structure and partition the nodes, using a single threshold of 0.5, into four profiles: *generator*, *bridge*, *terminal*, and *base*.

The resulting structure identifies *three qualitatively distinct critical paths* through the estimation problem system: structural immaturity, time pressure, and political pressure. The framework therefore provides a system-level view of estimation failure, shifting the analysis from isolated risks towards dependencies, propagation mechanisms, and potential intervention points. An illustrative case study of a public-sector estimation failure demonstrates the intended use of the framework, followed by a five-step roadmap for future empirical validation and application.

2012 ACM Subject Classification Software and its engineering → Software development process management; Mathematics of computing → Graph theory; General and reference → Estimation

Keywords and phrases effort estimation, uncertainty, systems analysis, network theory, critical paths

1 Introduction

Effort estimation is a critical activity in any project whose effort cannot be derived from price books, tariffs or standardised production processes. The class of *high-uncertainty projects* (R&D, novel product development, digital transformation, and large-scale scientific and public-sector software programmes) combines three structural features that make estimation simultaneously indispensable and intrinsically hard: requirements emerge during execution, technologies are immature, and there is no organizational baseline to calibrate against. The result is a paradox: the projects that most require accurate estimates are those for which achieving accurate estimates is least feasible [22].

The empirical SE literature contains a long catalogue of reasons why estimates go wrong: insufficient information [21], inadequate models [23, 4], cognitive biases [19, 11, 16, 12], external pressure on the estimator [13], and strategic misrepresentation in large public

projects [9]. These accounts are valuable, but they primarily identify and classify sources of estimation failure without explaining how these factors interact, how their combined effects accumulate, or where interventions are most effective. Three questions remain unanswered. (Q1) *How* do the recurrent estimation problems influence each other? (Q2) *Where* in the structure of estimation failure does distortion accumulate? (Q3) *Which* points are highest-leverage for intervention?

The main contribution of this vision paper is a *network-theoretic framework for effort estimation problems analysis* (Sections 3 and 4), illustrated on a public-sector failure (Section 4) and accompanied by a five-step empirical roadmap (Section 5). The framework's limits are made explicit in Section 5. The paper is positioned as a *vision* contribution: the proposed framework is intended as a hypothesis generating instrument rather than a validated diagnostic model.

2 Literature Review

Effort-estimation research has largely focused on the *development of estimation methods and tools*. The parametric tradition begins with COCOMO [2] and matures into COCOMO II [3]; the size-driven tradition includes Function Points [1] and Use Case Points [14]; expert-judgement methods are reviewed in Jørgensen [11] and operationalised in agile settings as Planning Poker and Story Points [5]; data-driven techniques migrate the problem to machine-learning regression [27, 4]. Systematic mappings [4, 21] and reviews [11] catalogue the methodological space in depth. This paper adopts a different perspective, focusing not on the proposal of a new estimation method, but rather on analysing the structure of the *problems* that undermine any method's accuracy under high uncertainty, and asks where in that structure intervention has highest leverage.

The empirical SE literature provides comparative accounts of estimation failures, but these typically focus on selected subsets of problems rather than offering an integrated view. Pasuksmit et al. [21] identify information quality as the dominant determinant of agile estimation accuracy. Trendowicz [23] documents the dependency of parametric models on calibration data that high-uncertainty projects rarely possess. Mohanani et al. [19] systematise the cognitive biases that affect SE judgement, finding that technical experts tend to be *more* optimistic than non-technical ones because the bottom-up reasoning that competence enables amplifies optimism rather than correcting it. Flyvbjerg [9] argues that for large projects with high top-management attention, bias is predominantly *political* (strategic misrepresentation) rather than cognitive (optimism), and that decision-makers actively exclude experts when the stakes are high. Jørgensen and Sjøberg [11, 13] document how customer expectations distort expert estimates even when estimators are aware of the influence.

Network representations of risk propagation and task dependency are by now standard in project management; the present work applies the same logic to a new target: the structure of effort-estimation failure. Fang et al. [6] model risk interactions in large engineering projects as a directed network and identify high-leverage risks through centrality analysis; Marle and Vidal [17] develop the broader programme of network-based analysis of risk interactions in complex projects. These accounts share a common move with the present paper: they replace the enumeration of risks (or problems) with a connected graph and ask where in the graph the cumulative impact concentrates.

3 Methodology

The methodology consists of three sequential phases: a literature-grounded taxonomy of the problems that affect effort estimation, an identification of the inter-problem dependencies reported in the literature, and a centrality analysis on the resulting directed graph.

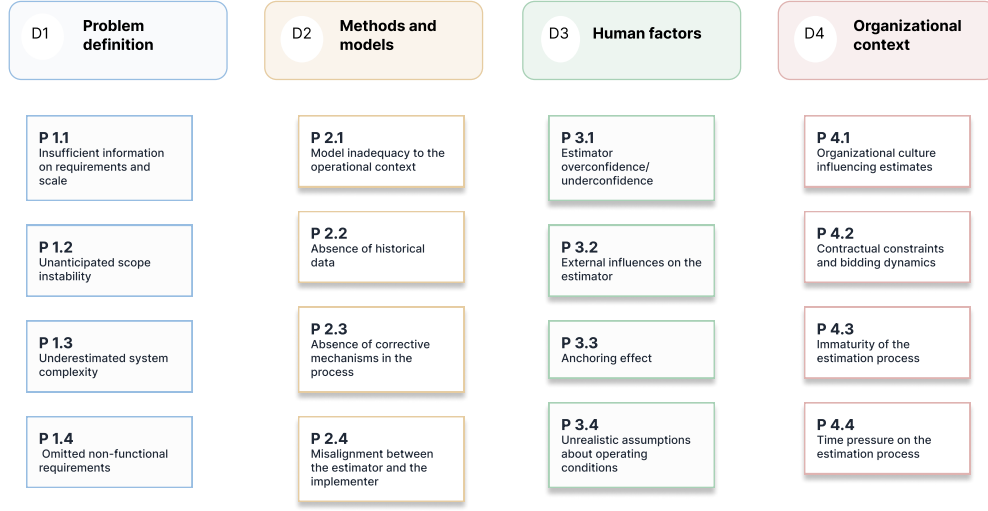
3.1 The Node Set: A Four-Domain Taxonomy of 16 Problems

The taxonomy integrates four primary literatures: Trendowicz and Jeffery [23] for effort factors and estimation principles; Pasuksmit et al. [21] for reasons of agile inaccuracy; Mohanani et al. [19] for cognitive biases; and Flyvbjerg [7, 9] for organizational and political factors and strategic misrepresentation. Sixteen problems are organised into four domains. *Technical and contextual factors* (technology, industry, geography, regulation) are deliberately excluded because they are project characteristics rather than estimation-process problems; their effect on estimation is captured indirectly through the problems they aggravate.

The taxonomy was developed and then subjected to peer debriefing in the sense of Lincoln and Guba [15] with a senior researcher in project management, the discipline that provides the analytic lens of the present paper. The debriefer audited the four-domain partition, the generate-and-discard procedure, and the per-row discard rationale for the 12 candidates; the partition and rationale were revised based on the resulting feedback before the present version was finalised. Peer debriefing is a process check, not a substitute for inter-rater coding with κ statistics, which remains Step 1 of the roadmap (Section 5).

The taxonomy is checked through a *generate-and-discard* procedure: 12 candidate problems were proposed in addition to the 16 retained nodes and were systematically classified into one of two exclusion categories with a discard rationale recorded for each. The two categories are (i) *subcase*, where the candidate reduces to a specific instance of one or two retained problems (e.g., *sunk-cost bias* reduces to P3.2 because the source of the influence is internal, not external, but the mechanism is the same); and (ii) *context variable*, where the candidate is a project characteristic that aggravates existing problems rather than being an estimation-process problem of its own (e.g., *turnover and knowledge loss* is captured by P3.4 once the estimator's assumption of team stability is violated). The distribution of the 12 candidates is 7 subcases and 5 context variables. The complete list, with classification and rationale per row, is shipped in `discarded_candidates.csv` of the replication package (Section 5). Figure 1 summarises the resulting 16-problem partition.

Domain D1 – Problem definition groups problems pertaining to the input of the estimation process: what the estimator knows about what is to be built. P1.1 captures vague, incomplete or contradictory requirements; P1.2 captures the absence of margins to absorb emerging requirements; P1.3 captures the failure to internalise architectural, emergent and interface complexity; P1.4 captures non-functional effort (performance, security, reliability, scalability, compliance) systematically excluded from scope. *Domain D2 – Methods and models* groups problems pertaining to the method used to produce the estimate. P2.1 captures the use of models calibrated on a different domain; P2.2 captures the absence of organizational history that prevents calibration; P2.3 captures the lack of work breakdown, independent review or cross-check; P2.4 captures the production of estimates by people other than those who will perform the work. *Domain D3 – Human factors* groups the cognitive distortions acting on the estimator. P3.1 captures the well-documented miscalibration between subjective confidence and objective accuracy; P3.2 captures the transformation of estimation from an analytical to a political-negotiation activity; P3.3 captures the orientation of judgement around an externally supplied number; P3.4 captures the default assumption of ideal team availability



■ **Figure 1** Four domain taxonomy of 16 effort estimation problems.

and zero overhead. *Domain D4 – organizational conditions of the estimation process* groups conditions of the estimation activity itself, as shaped by the surrounding organisation: culture, contractual structure, process maturity, time pressure on the process. D4 is *distinct* from the *project-context variables* (technology, industry, geography, regulation) excluded above: those describe the artefact being estimated, whereas D4 describes the institutional environment in which the estimate is produced. P4.1 captures a culture that systematically rewards aggressive estimates and penalises prudent ones; P4.2 captures contractual and bidding structures that make the estimate a tactical instrument rather than a forecast; P4.3 captures the absence of templates, role definitions and systematic post-mortem collection; P4.4 captures time pressure on the estimation process itself (distinct from delivery time pressure), which causes the estimator to skip work breakdown and cross-check.

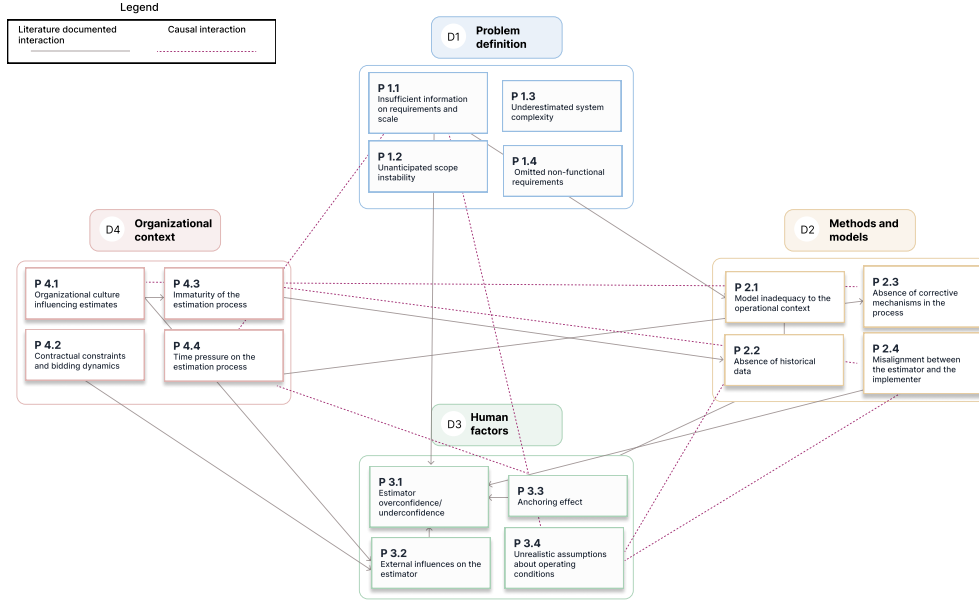
The taxonomy is a partition of the estimation-process problem space at a deliberately chosen level of abstraction, not an exhaustive enumeration of every individual issue.

3.2 The Edge Set and Centrality Definitions

Edge identification. The sixteen problems do not act in isolation. Nineteen inter-problem interactions are identified through two complementary procedures. First, a *literature search* retains an edge $u \rightarrow v$ if at least one primary source explicitly reports a causal influence of u on v ; this yields 12 *documented edges*. Second, a *mechanism-based inference* procedure admits an edge when the underlying causal mechanism is specific, non-trivial and defensible from established SE practice, even if not directly tested in a primary study; this yields 7 *inferred edges*. Inferred edges are visually distinct from documented ones in Figure 2 and are tracked separately throughout, but they enter the centrality computation with the same unit weight as documented edges. The result is a directed acyclic graph $G = (V, E)$ with $|V| = 16$ and $|E| = 19$.

Acyclicity. The graph is acyclic by design, not because the underlying causal mechanisms are acyclic in the real world. Each node represents the occurrence of a problem within a *single estimation episode*; the same problem firing in a subsequent episode would be a distinct phenomenon, requiring a separate node in a longitudinal extension of the model. Cycles in

the causal mechanism are real but operate across episodes (overconfidence today reduces investment in tomorrow’s corrective mechanisms, which feeds tomorrow’s overconfidence); they are out of scope for the present within-episode DAG and are discussed qualitatively in Section 5. The restriction gains analytic tractability (finite critical-path enumeration, well-defined centrality indices, unambiguous source/sink identification) at the cost of longitudinal expressiveness, restored in Step 4 of the roadmap (Section 5).



■ **Figure 2** Inter-problem interaction DAG. Solid arrows denote the 12 literature-documented edges; dashed arrows denote the 7 inferred edges. Nodes are coloured by domain.

Centrality indices. Two centrality indices are computed; they are *asymmetric* because the role of a node in a directed graph is intrinsically asymmetric: a problem that propagates distortion is not necessarily a problem that receives it. Three raw quantities enter the definitions. The *out-degree* $d^+(v)$ is the number of outgoing edges from v (how many problems v directly influences); the *in-degree* $d^-(v)$ is the number of incoming edges (how many problems directly influence v); the *betweenness centrality* $b(v)$ [10] is the fraction of shortest paths between other pairs of nodes that pass through v (how often v acts as a transit hub in the network). These three quantities enter the formal definitions of the Influence Index, I_{inf} (Eq. (1)), and the Vulnerability Index, I_{vuln} (Eq. (2)), introduced below.

Both indices are *relative-normalised* in the sense of Freeman [10], that is, each raw metric is divided by its maximum observed value in the network rather than by a graph-size constant. Absolute normalisation by $|V| - 1$ would compress all values into a narrow range on small graphs ($|V| = 16$); relative normalisation maps every metric into $[0, 1]$ and preserves the comparative structure among nodes, with a direct interpretation: a node with normalised value x is x -as-influential (or x -as-vulnerable) as the most influential (vulnerable) node in the network. The *Influence Index* is

$$I_{\text{inf}}(v) = \frac{1}{2}(\hat{d}^+(v) + \hat{b}(v)), \quad \hat{d}^+(v) = \frac{d^+(v)}{\max_{u \in V} d^+(u)}, \quad \hat{b}(v) = \frac{b(v)}{\max_{u \in V} b(u)}. \quad (1)$$

188 The *Vulnerability Index* is the relative-normalised in-degree:

$$189 \quad I_{\text{vuln}}(v) = \hat{d}^-(v) = \frac{d^-(v)}{\max_{u \in V} d^-(u)}. \quad (2)$$

190 The choice of out-degree and betweenness for I_{inf} , rather than an eigenvector- or PageRank-
 191 style metric, is dictated by the acyclicity of the graph: eigenvector centrality rewards
 192 connections to high-influence nodes, but a DAG has no strongly-connected components
 193 by construction, so eigenvector centrality collapses to a degenerate ordering. Out-degree
 194 captures *direct* propagation capacity; betweenness captures *transit* capacity; their arithmetic
 195 mean gives equal weight to both.

196 *Profile classification.* Nodes are classified by the joint profile of I_{inf} and I_{vuln} using a
 197 single cut at 0.5:

	<i>Bridge</i>	if $I_{\text{inf}} \geq 0.5$ and $I_{\text{vuln}} \geq 0.5$,
	<i>Generator</i>	if $I_{\text{inf}} \geq 0.5$ and $I_{\text{vuln}} < 0.5$,
198	<i>Terminal</i>	if $I_{\text{vuln}} \geq 0.5$ and $I_{\text{inf}} < 0.5$,
	<i>Base</i>	otherwise.

199 Under relative-normalised centrality [10], the cut-point 0.5 identifies nodes that are at least
 200 half as central as the most central node of the network. The classification is not entirely
 201 insensitive to small perturbations of the threshold: a shift to 0.45 leaves all profiles unchanged,
 202 whereas a shift to 0.55 reclassifies the borderline node P4.4 ($I_{\text{inf}} = 0.500$) from *Generator* to
 203 *Base*; no other node changes profile within ± 0.05 .

204 4 Results and Case Study

205 Section 4 reports the centrality and critical-path analysis and then applies the framework to
 206 the Healthcare.gov launch and recovery (2011–2015) as an *illustration* of how the analysis
 207 maps onto a documented failure. The case study is offered as illustrative, not confirmatory.

208 4.1 Centrality Results and Critical Paths

209 The resulting structure is strongly polarised: nodes with high influence have low vulnerability
 210 and vice versa, with one exception: P3.2, which carries both ($I_{\text{inf}} = 0.708$, $I_{\text{vuln}} = 0.50$) and
 211 is examined below. Of 16 nodes, 5 qualify as generators, 1 as a bridge, 4 as terminals, and 6 as
 212 base. Domain D4 contributes 40% of the cumulative network influence ($\sum_{v \in D4} I_{\text{inf}}(v) = 1.999$
 213 out of a network total of 5.039); its four nodes jointly form the dominant generator of
 214 distortion in the network. Three problems in D1 (P1.2, P1.3, P1.4) sit at the origin and the
 215 sink simultaneously: this does not signal irrelevance but autonomy from the inter-problem
 216 dynamics: they are recurrent problems that, when they fire, do so independently of the others.
 217 P3.1 has maximum vulnerability ($I_{\text{vuln}} = 1.000$) and zero influence: it is the terminal where
 218 distortions of informational, methodological, organizational and cognitive origin converge and
 219 become part of the final estimate. P3.2, the only Bridge in the network, receives distortion
 220 from organizational culture (P4.1) and contractual constraints (P4.2) and propagates it to
 221 P3.1 and to P2.3.

222 A *critical path of error amplification* is defined as a propagation tree rooted in a generator
 223 that (i) uses only documented edges, (ii) crosses at least two domains, (iii) terminates at a
 224 *terminal* node. The procedure identifies three paths, illustrated in Figure 3 and summarised
 225 qualitatively below.

■ **Table 1** Centrality scores and node profiles.

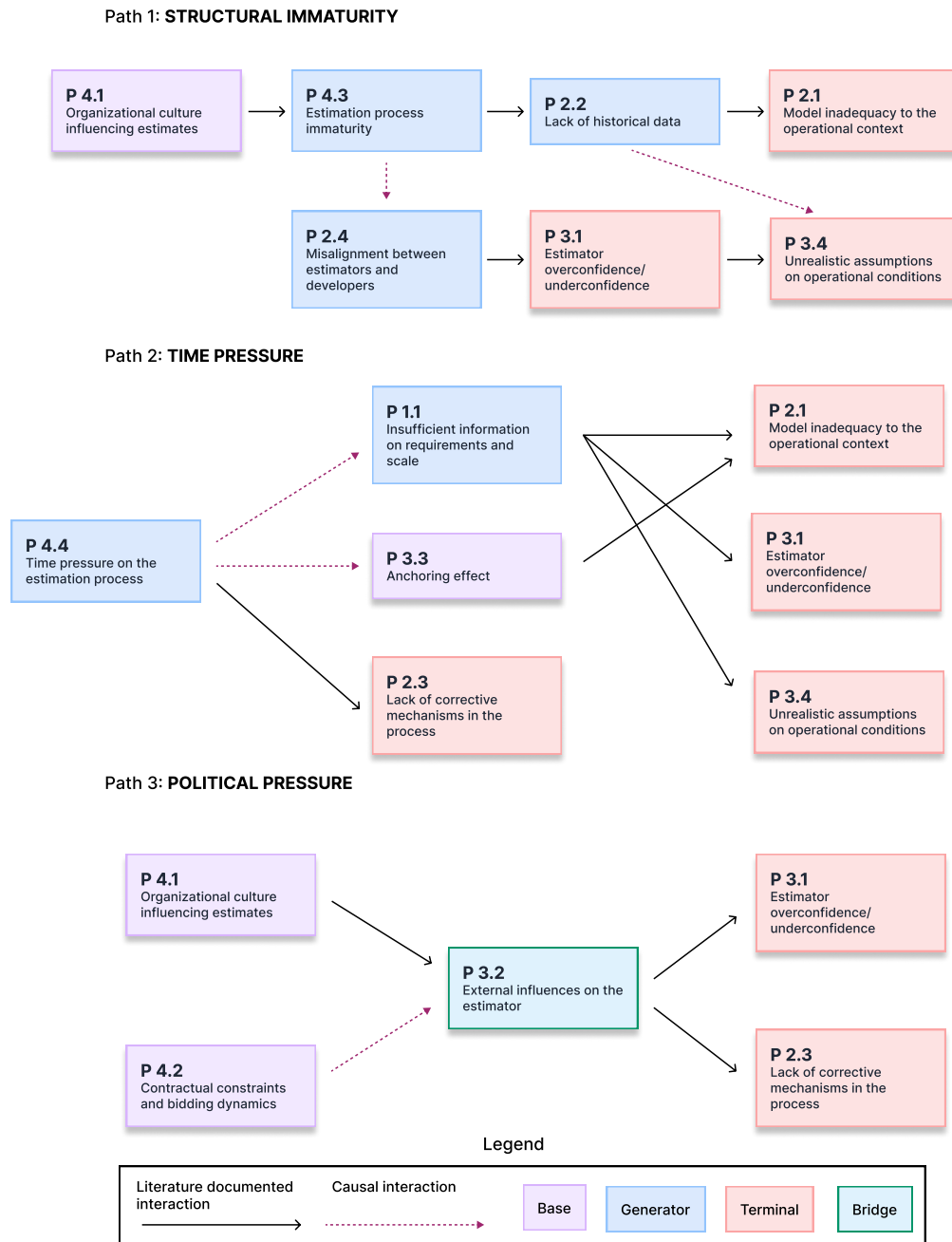
Node	Domain	d^+	d^-	I_{inf}	I_{vuln}	Profile
P4.3	D4	2	1	0.833	0.25	Generator
P1.1	D1	3	1	0.812	0.25	Generator
P2.2	D2	2	1	0.708	0.25	Generator
P3.2	D3	2	2	0.708	0.50	Bridge
P2.4	D2	2	1	0.583	0.25	Generator
P4.4	D4	3	0	0.500	0.00	Generator
P4.1	D4	2	0	0.333	0.00	Base
P4.2	D4	2	0	0.333	0.00	Base
P3.3	D3	1	1	0.229	0.25	Base
P1.2, P1.3, P1.4	D1	0	0	0.000	0.00	Base
P2.1	D2	0	2	0.000	0.50	Terminal
P2.3	D2	0	3	0.000	0.75	Terminal
P3.1	D3	0	4	0.000	1.00	Terminal
P3.4	D3	0	3	0.000	0.75	Terminal

Path 1 is the *structural-immaturity cascade*. It is a *cascade* because every node along the chain creates the precondition for the next: a culture that penalises prudent estimates (P4.1) discourages investment in process maturation (P4.3); without process maturation, no historical data are systematically collected (P2.2); without historical data, parametric models cannot be calibrated and the estimator and the implementing team are no longer systematically aligned (P2.1, P2.4); the resulting under-specification feeds three independent terminal distortions (P3.1, P3.4, and the methodological terminal P2.1). The chain is called *structural* because each step removes a piece of organizational scaffolding that the next would need.

Path 2 is the *time-pressure fan*. It is a *fan* because a single root, time pressure on the estimation process itself (P4.4), simultaneously produces five distinct distortion mechanisms: it cuts short the information-gathering effort (P1.1), it eliminates the cross-check and work-breakdown steps (P2.3), and it forces the estimator to anchor on the first plausible number (P3.3). Each of these in turn propagates to one or more terminals (P2.1, P3.1, P3.4). The chain is called *time pressure* because the only intervention that disables all five branches at once is reducing the time pressure on the estimation activity itself (rather than on project delivery). It is also the only path that crosses all four domains.

Path 3 is the *political-pressure funnel*. It is a *funnel* because two independent organizational sources of pressure (P4.1, organizational culture; P4.2, contractual and bidding dynamics) converge on a single intermediate node (P3.2, external influences on the estimator), which then propagates the combined distortion to two terminals (P3.1, P2.3). The chain is called *political pressure* because both source nodes operate not through the estimator's cognitive limits but through the institutional incentives shaping the estimator's behaviour: a culture that rewards aggressive estimates and a bidding game in which realistic estimates lose.

Three facts emerge across the paths. *First*, all three paths originate from D4, consistent with the 40% cumulative-influence share already noted above. *Second*, all three paths terminate on $P_{3.1}$: any intervention that improves estimator calibration (e.g., decomposition and work-breakdown [23], group estimation [20], or reference-class forecasting [8]) attenuates



■ **Figure 3** Critical paths identified.

all three paths simultaneously. *Third*, $P_{3.2}$ is the single point of maximum *leverage*: breaking the causal chain at $P_{3.2}$ simultaneously protects $P_{3.1}$ (maximum-vulnerability terminal) and $P_{2.3}$ (absence of corrective mechanisms, the point of no return).

4.2 An Illustrative Case Study: Healthcare.gov

The framework is illustrated through a compact case study of a documented public-sector estimation failure: the launch and recovery of the U.S. federal healthcare exchange platform *Healthcare.gov* (2011–2015). The case is documented in two GAO reports [25, 26], one HHS Office of Inspector General case study [24], a McKinsey assessment commissioned by CMS in March 2013 [18], and the *Health Affairs* literature on Affordable Care Act implementation.

Mapped onto the taxonomy, the launch exhibits 12 of 16 problems as PRESENT and 4 as NON-EVALUABLE (the latter clustered in D3, a limit of the evidence base that describes decisions but not cognitive processes). Mapped onto the three critical paths, the case is best read as one in which the three paths are *not independent*: *political pressure (Path 3) is the primary driver, and it activates the other two as second-order effects*. The legislated deadline of the Affordable Care Act (POLITICAL PRESSURE on the estimation calendar) generated the time pressure of Path 2 [25]; the contracting structure chosen to be “expedient” under political constraints ($P_{4.2} \rightarrow P_{3.2}$) reinforced a pre-existing process immaturity (Path 1) by removing the oversight gates that would have slowed delivery [25, 24]. The empirical signature of this reading is the documented *desensitisation* of senior CMS officials to bad news on schedule and cost [24, 18]: a political-cultural phenomenon at $P_{3.2}$ rather than a cognitive miscalibration of individual estimators at $P_{3.1}$.

The recovery, the “tech surge”, is interpretable as a sequence of *leverage-point* interventions targeted precisely at the bridge node of the political-pressure path and at the point-of-no-return terminal: a single decision authority neutralised the bridge ($P_{3.2}$, *external influences on the estimator*); a “badgeless culture” mitigated $P_{2.4}$ (*estimator-doer misalignment*) and $P_{3.4}$ (*unrealistic operating-condition assumptions*); ruthless scope prioritisation protected $P_{1.2}$ (*unanticipated scope instability*) and $P_{2.3}$ (*absence of corrective mechanisms*); short iterative cycles reintroduced $P_{2.3}$ as an active corrective mechanism. Of the 12 problems present at launch, 8 were mitigated, 3 remained unchanged and 1 was structurally superseded. The structure of the intervention is consistent with the framework’s interventionist reading: the recovery did not address the political-pressure root ($P_{4.2}$, *contractual constraints*, remained in place; $P_{4.4}$, *time pressure on the estimation process*, was only partially relieved), it broke the path *at the bridge*. The case *illustrates* but does not *test* the framework; replication on a non-governmental case not selected on outcome is Step 5 of the roadmap (Section 5).

5 Conclusion

This paper set out to address three research questions. **(Q1)** *How* do the recurrent problems influence each other? **(Q2)** *Where* in the structure of estimation failure does distortion accumulate? **(Q3)** *Which* points are highest-leverage for intervention? Each is addressed below as a testable structural prediction rather than a validated diagnostic; the roadmap at the end of this section specifies how to test each.

P1 – The recurrent problems form a connected causal DAG (answers Q1). Effort-estimation failure is not the realisation of an unordered list of problems; the 16 recurrent problems are linked by 19 directed inter-problem edges (Figure 2) that form a directed acyclic graph within a single estimation episode. The presence of this structure is the precondition

for the centrality analysis on which P2 and P3 rest, and is itself the central empirical prediction of the paper: if independent coders cannot recover a comparable DAG from the same primary literature, the framework’s premises are falsified. The corresponding empirical test is Step 1 of the roadmap below (inter-rater coding), and is supplemented by Step 2 (effect-size edge weighting) and Step 4 (longitudinal extension to cross-episode dynamics, which would lift the within-episode acyclicity restriction).

P2 – Cognitive convergence at $P_{3.1}$ (answers Q2). Every identified critical path terminates at $P_{3.1}$ (estimator over-/under-confidence), which holds maximum vulnerability ($I_{\text{vuln}} = 1.000$) and zero influence. Distortion of informational, methodological, organizational and cognitive origin converges on cognitive calibration before becoming part of the final estimate. The corresponding intervention claim is that improving estimator calibration (through decomposition and work-breakdown [23], group estimation [20], or reference-class forecasting [8]) is predicted to attenuate all three critical paths simultaneously, regardless of which path is active in a given organisation.

P3 – Bridge-node bottleneck at $P_{3.2}$ (answers Q3). $P_{3.2}$ (external influences on the estimator) is the unique bridge in the DAG and the single structural point of maximum leverage: it simultaneously protects the maximum-vulnerability terminal ($P_{3.1}$) and the *point-of-no-return* terminal ($P_{2.3}$, absence of corrective mechanisms). The prediction is that interventions targeted at the political-pressure path through $P_{3.2}$ should disproportionately reduce downstream distortion compared with broader D4 interventions, given the structural concentration of leverage at the bridge. The complementary observation is that the four organizational nodes of Domain D4 jointly carry 40% of the cumulative network influence ($\sum I_{\text{inf}} = 1.999$ out of 5.039), so organizational-level interventions remain a high-leverage secondary target.

Threats to validity. Five limits are worth flagging. (*Construct.*) The taxonomy was built without inter-rater agreement. (*Inference.*) The Influence and Vulnerability indices are descriptive; statistical inference on them (bootstrap CIs, permutation tests) is not yet associated with the centrality scores. (*Edges.*) Seven of the nineteen edges are inferred from causal-mechanism analysis rather than directly documented in a primary study (Section 3.2); they enter the centrality computation with the same unit weight as documented edges. (*Scope.*) The choice of DAG over a general digraph is a deliberate scoping decision: the analysis covers the propagation of distortion within a single estimation episode, leaving cross-episode feedback loops (the longitudinal evolution of estimation practice) out of scope. (*Case study.*) The Healthcare.gov case has $n = 1$ and is retrospectively selected on a publicly known failure, exposing the analysis to selection-on-outcome bias. Institutional sources systematically describe decisions but not cognitive processes; the four NON-EVALUABLE assignments in D3 are an artefact of this evidence base, not of the framework.

Research roadmap. A five-step empirical agenda addresses the limits above:

1. *Inter-rater coding of the taxonomy.* Independent coding of the 16-problem partition by three to five empirical SE researchers on a corpus of published estimation post-mortems;
2. *Edge weighting from effect-size meta-analysis.* For each documented edge, extract effect sizes, replace uniform weights and re-compute centrality and re-identify critical paths.
3. *Statistical inference on the centrality scores.* Bootstrap resampling on the weighted edge set; 95% confidence intervals for I_{inf} and I_{vuln} ; statistical tests on the ordering of nodes within each profile.
4. *Longitudinal extension to cross-episode dynamics.* Lift the within-episode scope restriction (Section 3.2): admit cross-episode feedback edges (e.g., overconfidence at episode t reduces investment in corrective mechanisms at episode $t+1$ and feeds overconfidence at episode

$t+2$), producing a possibly cyclic digraph in which eigenvector- and PageRank-style centralities become applicable.

5. *Replication on a second, non-governmental case study not selected on outcome.* Apply the taxonomy-coding procedure to a private-sector estimation post-mortem that was not selected because it failed; report the coding before and after blinding to outcome to bound selection bias.

The network framing trades the comprehensiveness of a list for the predictive structure of a graph. The roadmap above turns that trade-off into an empirical programme.

Data Availability

The replication package supporting this paper, comprising the 16-node taxonomy (`nodes.csv`), the 19-edge adjacency matrix (`edges.csv`; 12 documented + 7 inferred), the 12 discarded candidates with per-row rationale (`discarded_candidates.csv`), the Python/NetworkX analysis script (`compute centrality.py`), and the precomputed outputs (`centrality_results.csv`, `critical_paths.txt`), is archived at <https://figshare.com/s/b7f1c67fc5c4353175f8>. Running `compute centrality.py` fully reproduces Table 1, Figure 2, and Figure 3. The package is released under a dual licence: MIT for code, CC BY 4.0 for data.

References

- 1 Allan J. Albrecht and John E. Gaffney. Software function, source lines of code, and development effort prediction: A software science validation. *IEEE Transactions on Software Engineering*, SE-9(6):639–648, 1983. doi:10.1109/TSE.1983.235271.
- 2 Barry W. Boehm. *Software Engineering Economics*. Prentice-Hall, 1981.
- 3 Barry W. Boehm, Chris Abts, A. Winsor Brown, Sunita Chulani, Bradford K. Clark, Ellis Horowitz, Ray Madachy, Donald J. Reifer, and Bert Steece. *Software Cost Estimation with COCOMO II*. Prentice-Hall, 2000.
- 4 Cleyton E. Carbonera, Kleinner Farias, and Vinicius Bischoff. Software development effort estimation: A systematic mapping study. *IET Software*, 14(4):328–344, 2020. doi:10.1049/iet-sen.2018.5334.
- 5 Mike Cohn. *Agile Estimating and Planning*. Prentice-Hall, 2005.
- 6 Chao Fang, Franck Marle, Enrico Zio, and Jean-Claude Bocquet. Network theory-based analysis of risk interactions in large engineering projects. *Reliability Engineering & System Safety*, 106:1–10, 2012. doi:10.1016/j.ress.2012.04.005.
- 7 Bent Flyvbjerg. Policy and planning for large-infrastructure projects: problems, causes, cures. *Environment and Planning B: Planning and Design*, 34(4):578–597, 2007. doi:10.1068/b32111.
- 8 Bent Flyvbjerg. Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. *European Planning Studies*, 16(1):3–21, 2008. doi:10.1080/09654310701747936.
- 9 Bent Flyvbjerg. Top ten behavioral biases in project management: An overview. *Project Management Journal*, 52(6):531–546, 2021. doi:10.1177/87569728211049046.
- 10 Linton C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979. doi:10.1016/0378-8733(78)90021-7.
- 11 Magne Jørgensen. A review of studies on expert estimation of software development effort. *Journal of Systems and Software*, 70(1–2):37–60, 2004. doi:10.1016/S0164-1212(02)00156-5.
- 12 Magne Jørgensen. Unit effects in software project effort estimation. *Journal of Systems and Software*, 117:366–373, 2016. doi:10.1016/j.jss.2016.04.014.

- 392 13 Magne Jørgensen and Dag I. K. Sjøberg. The impact of customer expectations on software
393 development effort estimates. *International Journal of Project Management*, 22(4):317–325,
394 2004. doi:10.1016/S0263-7863(03)00069-7.
- 395 14 Gustav Karner. Resource estimation for objectory projects. Technical report, Objective
396 Systems SF AB, 1993.
- 397 15 Yvonna S. Lincoln and Egon G. Guba. *Naturalistic Inquiry*. Sage Publications, Newbury Park,
398 CA, 1985.
- 399 16 Erik Løhre and Magne Jørgensen. Numerical anchors and their strong effects on software
400 development effort estimates. *Journal of Systems and Software*, 116:49–56, 2016. doi:
401 10.1016/j.jss.2015.03.015.
- 402 17 Franck Marle and Ludovic-Alexandre Vidal. *Managing Complex, High Risk Projects:
403 A Guide to Basic and Advanced Project Management*. Springer, 2016. doi:10.1007/
404 978-1-4471-6787-7.
- 405 18 McKinsey & Company. CMS Healthcare Exchange (HIX) Pre-Launch Assessment: Findings
406 and Recommendations. Technical report, McKinsey & Company (commissioned by CMS),
407 March 2013.
- 408 19 Rahul Mohanani, Iflaah Salman, Burak Turhan, Pilar Rodríguez, and Paul Ralph. Cognitive
409 biases in software engineering: A systematic mapping study. *IEEE Transactions on Software
410 Engineering*, 46(12):1318–1339, 2020. doi:10.1109/TSE.2018.2877759.
- 411 20 Kjetil Moløkken-Østvold and Magne Jørgensen. Group processes in software effort estimation.
412 *Empirical Software Engineering*, 9(4):315–334, 2004. doi:10.1023/B:EMSE.0000039882.39206.
413 5a.
- 414 21 Jirat Pasukmit, Patanamon Thongtanunam, and Shanika Karunasekera. A systematic
415 literature review on reasons and approaches for accurate effort estimations in agile. *ACM
416 Computing Surveys*, 56(11), 2024. doi:10.1145/3664597.
- 417 22 Michael T. Pich, Christoph H. Loch, and Arnoud De Meyer. On uncertainty, ambiguity,
418 and complexity in project management. *Management Science*, 48(8):1008–1023, 2002. doi:
419 10.1287/mnsc.48.8.1008.163.
- 420 23 Adam Trendowicz and Ross Jeffery. *Software Project Effort Estimation: Foundations and
421 Best Practice Guidelines for Success*. Springer, 2014. doi:10.1007/978-3-319-03629-8.
- 422 24 U.S. Department of Health and Human Services Office of Inspector General. CMS Management
423 of the Federal Marketplace: A Case Study. Technical Report OEI-06-14-00350, U.S. Department
424 of Health and Human Services, Office of Inspector General, 2016. URL: [https://oig.hhs.
425 gov/oei/reports/oei-06-14-00350.pdf](https://oig.hhs.gov/oei/reports/oei-06-14-00350.pdf).
- 426 25 U.S. Government Accountability Office. HealthCare.gov: Ineffective Planning and Oversight
427 Practices Underscore the Need for Improved Contract Management. Technical Report GAO-
428 14-694, U.S. Government Accountability Office, July 2014. URL: [https://www.gao.gov/
429 products/gao-14-694](https://www.gao.gov/products/gao-14-694).
- 430 26 U.S. Government Accountability Office. Healthcare.gov: CMS Has Taken Steps to Address
431 Problems, but Needs to Further Implement Systems Development Best Practices. Technical
432 Report GAO-15-238, U.S. Government Accountability Office, March 2015. URL: [https:
433 //www.gao.gov/products/gao-15-238](https://www.gao.gov/products/gao-15-238).
- 434 27 Jianfeng Wen, Shixian Li, Zhiyong Lin, Yong Hu, and Changqin Huang. Systematic literature
435 review of machine learning based software development effort estimation models. *Information
436 and Software Technology*, 54(1):41–59, 2012. doi:10.1016/j.infsof.2011.09.002.