Learning When to Quit: An Empirical Model of Experimentation in Standards Development

Online Appendix

Bernhard Ganglmair*
University of Mannheim & ZEW Mannheim

Timothy Simcoe†
Boston University & NBER

Emanuele Tarantino‡
LUISS, EIEF & CEPR

June 2024

Abstract

This is the Online Appendix for "Learning When to Quit: An Empirical Model of Experimentation in Standards Development," published in the *American Economic Journal: Microeconomics*. The data is used for the analysis in the main text, and this appendix is available at https://doi.org/10.5281/zenodo.11180268.

*ZEW Mannheim, L7, 1, 68161 Mannheim, Germany; Email: b.ganglmair@gmail.com.

†Boston University, Questrom School of Business, 595 Commonwealth Ave, Boston, MA 02215, USA; Email: tsimcoe@bu.edu.

‡LUISS and EIEF, Viale Romania 32, 00124 Rome, Italy; Email: etarantino@luiss.it.

Contents

A Additional Tables and Figures

Table A.1: Examples of IETF Internet Standards

	Description	RFC	Year
RTP	Real-time Transport Protocol	3550	2003
SIP	Session Initiation Protocol	3261	2002
HTTP	Hypertext Transfer Protocol – HTTP/1.1	2616	1999
IPV6	Internet Protocol, Version 6 (IPv6) Specification	2460	1998
DHCP	Dynamic Host Configuration Protocol	2131	1997
POP3	Post Office Protocol – Version 3	1939	1996
NAT	Network Address Translator	1631	1994
FTP	File Transfer Protocol	959	1985
TCP	Transmission Control Protocol	793	1981
IP	Internet Protocol	791	1981

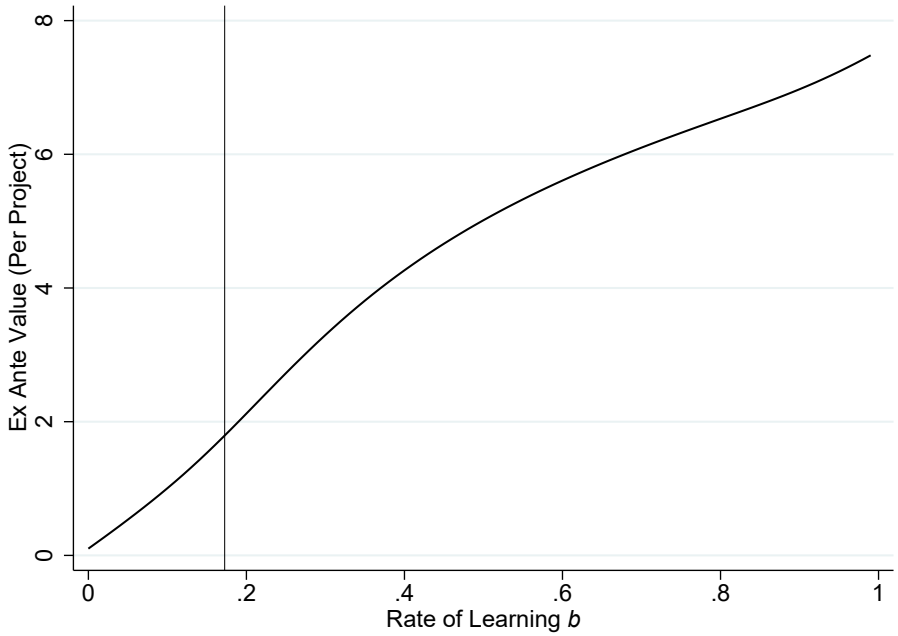
Notes: RFC means "Request for Comments," the respective column provides the number of the RFC; Year is the year in which the standard was certified by the IETF.

Table A.2: More Robustness Results

	Area Specific (1)	Censored Projects (2)	RFC Citations (3)	Deadline $\overline{T} = 50$ (4)	First Project (5)
Parameter Estimates b and p					
Rate of Learning: b	0.10	0.23	0.20	0.18	0.29
	(0.003)	(0.004)	(0.005)	(0.003)	(0.021)
Quality Prior: p	0.59	0.51	0.45	0.58	0.72
	(0.008)	(0.003)	(0.006)	(0.004)	(0.015)
Cost Estimates					
Costs at $t = 1$	1.34	2.39	1.74	2.47	3.97
	(0.015)	(0.026)	(0.021)	(0.024)	(0.241)
Costs at $t = 10$	1.08	0.90	0.67	1.21	1.39
	(0.010)	(0.006)	(0.006)	(0.006)	(0.034)
Costs at $t = 20$	0.46	0.46	0.38	0.68	0.74
	(0.015)	(0.010)	(0.012)	(0.005)	(0.058)
Costs at $t = 30$				0.47	
				(0.011)	
log-likelihood/Project	-2.070	-1.913	-2.096	-2.150	-3.328
Projects	14,444	21,779	14,444	14,444	226

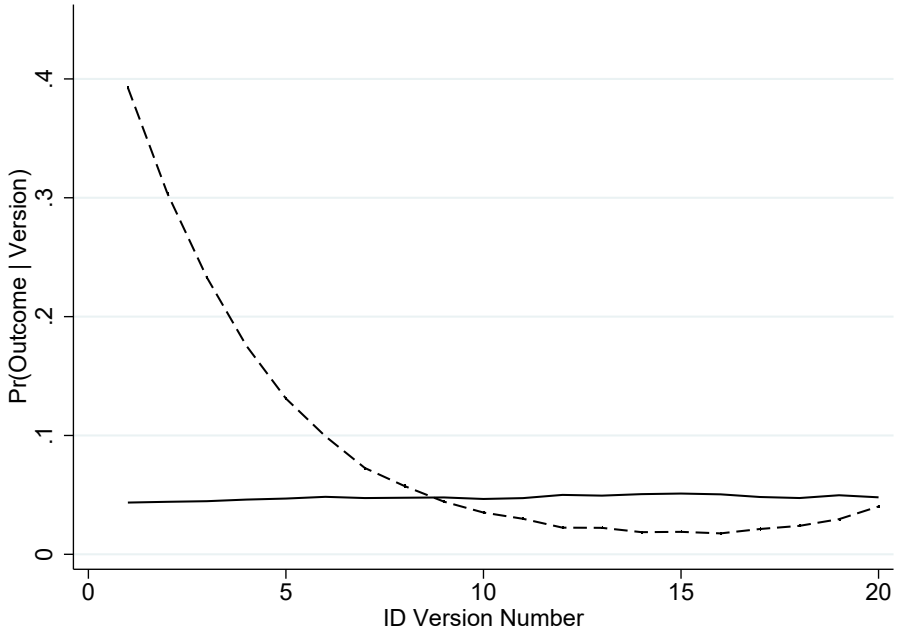
Notes: This table presents ML estimates from further robustness checks. In (1) ("Area Specific"), we allow project values $\hat{\pi}(t)$ to differ across the seven broad Technology Areas defined by the IETF; in (2) ("Censored Projects"), we include active projects and those completed projects that were initiated in or after 2010; in (3) ("RFC Citations"), we use citations by other RFCs for project values estimates $\hat{\pi}(t)$; in (4) ("Deadline $\overline{T} = 50$"), we extend the forced deadline to $\overline{T} = 50$; in (5) ("First Project"), we consider only working group projects that were the first project of their respective working group. In model (2), the estimates for project values $\hat{\pi}(t)$ are based on the extended sample, including RFCs initiated in year 2010 or after; in model (5), the estimates for project values $\hat{\pi}(t)$ are based on all working group projects (as in model (2) in Table 2). Standard errors for the parameter estimates are reported in parentheses. Standard errors for cost estimates are calculated using the delta method.

Figure A.1: Ex-Ante Project Value (Varying with Rate of Learning b)



Notes: This figure depicts the ex-ante project value (in $t = 0$ before authors are endowed with their initial version) as a function of the rate of learning b. It illustrates the benefits of faster learning. We plot the ex-ante value for the parameter estimates from the model in Column (1) of Table 2. The vertical line depicts $b^* = 0.172$.

Figure A.2: Hazard Rates Without Post-Breakthrough Phase



Notes: This figure depicts the hazard rates obtained using the estimates of the empirical model in which we turn off the post-breakthrough phase.

B Model Identification

In this appendix, we show the conditions under which we achieve point-identification of the parameters in our empirical model. We proceed in two steps. We first show that, without further restrictions, the statistical features of the data, combined with our model, are not enough to uniquely identify all our parameters. We then make an assumption on the fixed costs that renders point-identification feasible.

B.1 No Further Restrictions

We consider a recursive solution of the empirical model with finite-time horizon $\overline{T}$, given $p \in (0,1)$, $b \in (0,1)$, $G(\varepsilon)$ strictly monotone, and $\hat{\pi}(t) > \hat{\pi}(t-1)$ and $\hat{\pi}(1) > 0$. Moreover, to ease notation, we let $\hat{p}(t) \equiv \hat{p}(t,0)$ represent posterior beliefs in period t (before a breakthrough). The number of parameters to identify is $2 \times \overline{T}$: p, b and $F(t, \sigma_t)$, with $t = 1, \ldots, \overline{T}-1$, and $\sigma_t = 0, 1$. We start by constructing the expressions for the choice probabilities (of publication and abandonment) in each t. Then, we analyze whether these expressions are sufficient to identify all the parameters in our model.

For the construction of the choice probabilities, we compute the number of projects in the pre- and post-breakthrough phase in each stage t of the model. We start with the count of projects in the pre-breakthrough phase. Recall that we have normalized the total number of projects to one so that $N_t^0 + N_t^1 = 1$, where N_t^0 and N_t^1 have been defined in Section II.D.

Claim B.1. *The number of projects in the pre-breakthrough phase in any $t \in \left[1, \overline{T}-1\right]$ is given by*

$$N_t^0 = A_t + \sum_{j=t}^{\overline{T}-1} \frac{A_{j+1}}{\prod_{k=t}^{j} (1 - \hat{p}(k)b)}. \tag{B.1}$$

Proof. To derive (B.1), we proceed recursively, starting from $\overline{T}-1$:

$$\begin{aligned} N_{\overline{T}-1}^0 &= A_{\overline{T}-1} + A_{\overline{T}} \frac{1}{1 - \hat{p}(\overline{T}-1)b} \\ N_{\overline{T}-2}^0 &= A_{\overline{T}-2} + N_0^{\overline{T}-1} \frac{1}{1 - \hat{p}(\overline{T}-2)b} \\ &= A_{\overline{T}-2} + \frac{A_{\overline{T}-1}}{1 - \hat{p}(\overline{T}-2)b} + \frac{A_{\overline{T}}}{\left[1 - \hat{p}(\overline{T}-2)b\right]\left[1 - \hat{p}(\overline{T}-1)b\right]}. \end{aligned}$$

By iteration of this rule, we obtain Equation (B.1) in the claim. Q.E.D.

We now define the number of projects in the post-breakthrough phase in each stage t.

Claim B.2. *The number of projects in the post-breakthrough phase in any $t \in \left[1, \overline{T}-1\right]$ is given by*

$$N_t^1 = \sum_{j=t}^{\overline{T}} S_j - \sum_{j=t}^{\overline{T}-2} N_{j+1}^0 \frac{\hat{p}(j)b}{1 - \hat{p}(j)b} - A_{\overline{T}} \frac{\hat{p}(\overline{T}-1)b}{1 - \hat{p}(\overline{T}-1)b}. \tag{B.2}$$

Proof. For the derivation of (B.2), we proceed recursively, starting from $\overline{T}-1$:

$$\begin{aligned} N^1_{\overline{T}-1} &= S_{\overline{T}} + S_{\overline{T}-1} - A_{\overline{T}}\frac{\hat{p}(\overline{T}-1)b}{1-\hat{p}(\overline{T}-1)b} \\ N^1_{\overline{T}-2} &= S_{\overline{T}-2} + N^1_{\overline{T}-1} - N^0_{\overline{T}-1}\frac{\hat{p}(\overline{T}-2)b}{1-\hat{p}(\overline{T}-2)b} \\ N^1_{\overline{T}-3} &= S_{\overline{T}-3} + N^1_{\overline{T}-2} - N^0_{\overline{T}-2}\frac{\hat{p}(\overline{T}-3)b}{1-\hat{p}(\overline{T}-3)b}, \end{aligned}$$

which, after rearranging, yields

$$N^1_{\overline{T}-3} = \sum_{j=\overline{T}-3}^{\overline{T}} S_j - \sum_{j=\overline{T}-3}^{\overline{T}-2} N^0_{j+1}\frac{\hat{p}(j)b}{1-\hat{p}(j)b} - A_T\frac{\hat{p}(T-1)b}{1-\hat{p}(T-1)b}.$$

By iteration of this process, we obtain Equation (B.2) in the claim. Q.E.D.

The formulas for $N^{\sigma_t}_t$, $\sigma_t = 0, 1$, allow us to construct the $2\left(\overline{T}-1\right)$ expressions for the conditional choice probabilities in each t:

$$\frac{S_t}{N^1_t} = 1 - G(\hat{\pi}(t+1) - \hat{\pi}(t) - F(t,1)) \tag{B.3}$$

$$\frac{A_t}{N^0_t} = 1 - G(\hat{p}(t)b\hat{\pi}(t+1) - F(t,0)), \tag{B.4}$$

with $t = 1, \ldots, \overline{T}-1$. Given any guess for b and p and the strict monotonicity assumption on $G(\cdot)$, the expressions in (B.3) and (B.4) can be inverted to solve for the costs in each t. That is, for each t, $F(t,1)$ and $F(t,0)$ are uniquely identified by (B.3) and (B.4), yielding expressions that depend on p, b and data.

To explore identification of p and b, we start exploiting the additional condition stemming from the property that, by construction, $pb = N^1_1$. We then use (B.2) to compute the number of post-breakthrough phase projects in $t = 1$, so that

$$pb = N^1_1 = \sum_{j=1}^{\overline{T}} S_j - \sum_{j=1}^{\overline{T}-2} N^0_{j+1}\frac{\hat{p}(j)b}{1-\hat{p}(j)b} - A_{\overline{T}}\frac{\hat{p}(\overline{T}-1)b}{1-\hat{p}(\overline{T}-1)b}. \tag{B.5}$$

Claim B.3. *If $b \geq \underline{b} \in (0,1)$, there exists a unique function $p(b)$ that satisfies Equation (B.5), with $p(b) \in (\sum_t S_t, 1)$ and $t = 1, ..., \overline{T}$.*

Proof. To prove the claim, we proceed by induction. First, we show the claim for $\overline{T} = 2$. For this case, let

$$\beta_2(p,b) \equiv pb + \frac{\hat{p}(1)b}{1-\hat{p}(1)b}A_2 = S_1 + S_2. \tag{B.6}$$

The right-hand side takes values within the unit interval. At the left-hand side of the equation, function $\beta_2(p,b)$ is continuous for all $p, b \in (0,1)$. Moreover, it is strictly increasing

in p. This is because $\hat{p}(1)b/\left(1-\hat{p}(1)b\right)$ is increasing in $\hat{p}(1)$, and $\hat{p}(1)$ and pb are both increasing in p. Finally, $\beta_2(p,1) \to p$, $\beta_2(0,b) \to 0$ and $\beta_2(1,b) \to (A_2 b)/(1-b)+b$, increasing in b, with

$$\lim_{b\to 0} \beta_2(1,b) = 0 \quad \text{and} \quad \lim_{b\to 1} \beta_2(1,b) = \infty.$$

Thus, there exists a lower bound for b, $\underline{b}_2 \in (0,1)$, such that $\beta_2(1,b) \geq S_1 + S_2$ for all $b \geq \underline{b}_2$.

We now show that the claim holds for $\overline{T} = 3$. More specifically, $\overline{T} = 3$ implies that

$$N_1^1 = S_1 + S_2 + S_3 - A_2 \frac{\hat{p}(1)b}{1-\hat{p}(1)b} - A_3 \frac{b\left[\hat{p}(1)+\hat{p}(2)-b\hat{p}(1)\hat{p}(2)\right]}{\left[1-\hat{p}(1)b\right]\left[1-\hat{p}(2)b\right]},$$

so that

$$\beta_3(p,b) \equiv pb + A_2 \frac{\hat{p}(1)b}{1-\hat{p}(1)b} + A_3 \frac{b\left[\hat{p}(1)+\hat{p}(2)-b\hat{p}(1)\hat{p}(2)\right]}{\left[1-\hat{p}(1)b\right]\left[1-\hat{p}(2)b\right]} = \sum_{t=1}^{3} S_t. \tag{B.7}$$

As for (B.6), the right-hand side takes values within the unit interval. At the left-hand side of the equation, $\beta_3(p,b)$ is continuous for all $p, b \in (0,1)$. Function $\beta_3(p,b)$ is strictly increasing in p, because pb, $\hat{p}(1)b/\left(1-\hat{p}(1)b\right)$, and

$$b\left[\hat{p}(1)+\hat{p}(2)-b\hat{p}(1)\hat{p}(2)\right] / \left[1-\hat{p}(1)b\right]\left[1-\hat{p}(2)b\right]$$

are all strictly increasing in p. Finally, $\beta_3(p,1) \to p$, $\beta_3(0,b) \to 0$ and $\beta_3(1,b) \to (A_2 b)/(1-b) + b - A_3 + A_3/(1-b)^2 = \beta_2(1,b) - A_3 + A_3/(1-b)^2$. Hence, the same conclusions as for $\overline{T} = 2$ apply (i.e., there exists a lower bound for b, $\underline{b}_3 \in (0,1)$, such that $\beta_3(1,b) \geq S_1 + S_2 + S_3$ for all $b \geq \underline{b}_3$).

Reiterating this analysis, we can invoke the principle of induction to conclude that there exists a lower bound for b, denoted by $\underline{b}_{\overline{T}} \equiv \underline{b} \in (0,1)$, such that $\beta_{\overline{T}}(1,b) \geq S_1 + S_2 + \ldots + S_{\overline{T}}$ for all $b \geq \underline{b}$. Hence, provided $b \geq \underline{b}$, a unique value of $p(b)$ exists, with $p(b) \in (\sum_t S_t, 1)$ and $t = 1, ..., \overline{T}$. Q.E.D.

The main implication of the claims above is that, while we can pin down the per-period costs from the choice probabilities, we can only recover an implicit relationship between b and p, as given by $p(b)$. Therefore, our structural parameters $(p, b, F(t, \sigma_t))$ are not uniquely determined by model and data alone, and we need to impose additional restrictions to separately identify b and p.

B.2 Additional Restriction for Identification

In line with the main text, the assumption we make is that, in $t = 1$, the per-period costs are the same pre- and post-breakthrough:

Assumption B.1. *Pre- and post-breakthrough revision costs are equal in $t = 1$: $F(1,1) = F(1,0) = F(1)$.*

Then, in any $t > 1$, the costs in the pre- and post-breakthrough phases, denoted by $(F(t,0), F(t,1))$, can be recovered by inverting (B.3) and (B.4). Moreover, by Assumption B.1, we can obtain $F(1)$ by inverting (B.3) evaluated in $t = 1$:

$$k_1 \equiv G^{-1}(1 - S_1/N_1^1) = \hat{\pi}(2) - \hat{\pi}(1) - F(1). \tag{B.8}$$

Then, from (B.4) evaluated at $t = 1$, we obtain

$$k_2 \equiv G^{-1}(1 - A_1/N_1^0) = b\hat{p}(1)\hat{\pi}(2) - F(1).$$

Substituting $F(1)$ from (B.8), and rearranging, yields

$$b\hat{p}(1) - (k_2 - k_1) = \frac{\hat{\pi}(2) - \hat{\pi}(1)}{\hat{\pi}(2)}. \tag{B.9}$$

In the next claim, we provide the sufficient conditions such that a value of $b(p)$ solving (B.9) exists and is unique.

Claim B.4. *Let $G(\varepsilon) = \frac{1}{1+\exp(-\varepsilon)}$. If $p \geq \underline{p} \equiv 1 - A_1^2$, then there exists a unique value of $b(p) \in (S_1/p, (1 - A_1)/p)$ such that*

$$b\hat{p}(1) - \frac{k_2 - k_1}{\hat{\pi}(2)} = \frac{\hat{\pi}(2) - \hat{\pi}(1)}{\hat{\pi}(2)}, \tag{B.10}$$

with $0 < S_1/p < (1 - A_1)/p < 1$ for all $p \geq \underline{p}$.

Proof. Under the Logistic distribution,

$$k_1 = \log\left(\frac{N_1^1 - S_1}{S_1}\right) \quad \text{and} \quad k_2 = \log\left(\frac{N_1^0 - A_1}{A_1}\right). \tag{B.11}$$

Thus, both sides of (B.10) are continuous for all $N_1^1 - S_1 = pb - S_1 > 0$, $N_1^0 - A_1 = 1 - pb - A_1 > 0$ and $1 > p, b > 0$. After using $N_1^1 = pb$ and $N_1^0 = 1 - pb$, Equation (B.10) can be rewritten as

$$\alpha(p,b) \equiv b\hat{p}(1) - \left[\log\left(\frac{1 - bp - A_1}{A_1}\right) - \log\left(\frac{bp - S_1}{S_1}\right)\right]\frac{1}{\hat{\pi}(2)} = \frac{\hat{\pi}(2) - \hat{\pi}(1)}{\hat{\pi}(2)}. \tag{B.12}$$

Under our assumption of a strictly increasing and positive function $\hat{\pi}(\cdot)$, the right-hand side is a scalar that (strictly) lies within the unit interval. Moreover,

$$\lim_{b \to S_1/p} \alpha(p,b) = -\infty \quad \text{and} \quad \lim_{b \to (1-A_1)/p} \alpha(p,b) = +\infty.$$

Taken together, the continuity of α and its behavior in the limits imply that a solution exists for all $b \in (S_1/p, (1 - A_1)/p)$. We now determine the sufficient condition guaranteeing that such a solution is unique. Specifically, this boils down to studying the monotonicity of $\alpha(p,b)$ with respect to b.

We then take the derivative of $\alpha(p,b)$ with respect to b and find that

$$\frac{\partial \alpha(p,b)}{\partial b} = \frac{p\left[1-b\left(2-bp\right)\right]}{\left(1-bp\right)^2} + \frac{p\left(1-A_1-S_1\right)}{\left(1-bp-A_1\right)\left(bp-S_1\right)\hat{\pi}(2)}.$$

Because $1-A_1-S_1>0$, $1-pb=N_1^0>A_1$ and $pb=N_1^1>S_1$, the second term in the expression for $\frac{\partial \alpha(p,b)}{\partial b}$ is strictly positive. Using this fact, we provide below the sufficient restrictions under which $\frac{\partial \alpha(p,b)}{\partial b}>0$.

To begin with, we note that if the first term of $\frac{\partial \alpha(p,b)}{\partial b}$ is weakly positive, or $p \geq (2b-1)/b^2$, then $\alpha(p,b)$ is strictly monotone in b, with $(2b-1)/b^2$ increasing in b. Hence, if p is larger than this threshold evaluated at $b \to (1-A_1)/p$ (the upper bound for the values of b such that (B.10) is well-defined), it is larger for all relevant values of b. Doing so, we find that $p \geq 1-A_1^2$ is sufficient to guarantee that a solution to (B.10) is unique.[1] Moreover, $p \geq 1-A_1^2$ implies that $0 < S_1/p < (1-A_1)/p < 1$. Q.E.D.

This claim gives us the sufficient conditions for the existence of a unique $b(p)$, meaning that, thanks to Assumption B.1, point-identification can be achieved by jointly solving for p and b from $p=p(b)$ and $b=b(p)$. To conclude the analysis, we proceed as follows. First, we prove that, given Claim B.4, a unique (constrained) solution exists for p and b when $\overline{T}=2$. That is, we obtain point-identification in $\overline{T}=2$. Second, we give the additional condition under which this result on point-identification extends to the case of a general $\overline{T}$. To close the model, the value of p and b can then be plugged into the expressions for the fixed costs.

Model Solution for $\overline{T}=2$ We first show that a unique (constrained) solution exists in this case.

Claim B.5. *Let $\overline{T}=2$ and $G(\varepsilon)=\frac{1}{1+\exp(-\varepsilon)}$. Then, there exists a unique couple (b^*,p^*) provided $p^*>1-A_1^2$, $b^*\geq \underline{b}$ and $p^*b^*\in(S_1,1-A_1)$.*

Proof. To begin with, we show two properties of $p(b)$ (Claim B.3) that are useful for what follows. Namely, we prove that $p'(b)<0$ and $p''(b)>0$ under $p>\underline{p}$.

After taking the total derivative of Equation (B.6) in the proof of Claim B.3, we find that

$$\frac{dp}{db} = p'(b) \equiv -\frac{\frac{\partial \beta(p,b)}{\partial b}}{\frac{\partial \beta(p,b)}{\partial p}} = -\frac{p}{b}\frac{1+\frac{A_2[1-b(2-bp)]}{[1-b(2-bp)]^2}}{1+\frac{A_2(1-b)}{[1-b(2-bp)]^2}} < 0$$

[1] Were this condition violated (so that $p<(2b-1)/b^2$), monotonicity would hold if and only if the second term in $\frac{\partial \alpha(p,b)}{\partial b}$ were large enough. Since that term is strictly increasing in A_1, this boils down to finding the lower bound on A_1 such that the whole expression is strictly positive. We obtain:

$$\frac{\partial \alpha(p,b)}{\partial b} \geq 0 \iff A_1 \geq \underline{A}_1 \equiv \frac{\left(1-bp\right)^2\left(1-S_1\right)+\left(1-bp\right)\left[1-b\left(2-pb\right)\right]\left(bp-S_1\right)\hat{\pi}(2)}{\left(1-bp\right)^2+\left[1-b\left(2-bp\right)\right]\left(bp-S_1\right)\hat{\pi}(2)},$$

with $\underline{A}_1 < \left(0, N_1^0\right)$ for sufficiently large values of $\hat{\pi}(2)$. Hence, if $p<(2b-1)/b^2$, $\alpha(p,b)$ is strictly increasing in b for all $A_1>\underline{A}_1$.

for all $1-b(2-bp)>0$ (which is implied by $p>1-A_1^2$, see the proof of Claim B.4). Moreover, we find that

$$\frac{dp'(b)}{db}=p''(b)\equiv p\frac{A_2^2\{1-b[2-b(2-p)]\}+B^4+A_2B\{2-b[2+p(4-bC)]\}}{b^2[A_2(1-b)+B^2]^2}>0,$$

where $B\equiv 1-b(2-b)p$, $C\equiv 9+3b^2p-2b(3+p)$, $[A_2(1-b)+B^2]>0$, and the sign of the numerator is strictly positive under $1-b(2-bp)>0$, or $p>1-A_1^2$.

We now plug $p(b)$ into Equation (B.12). We obtain

$$\frac{bp(b)(1-b)}{(1-p(b))+p(b)(1-b)}-\left[\log\left(\frac{1-bp(b)-A_1}{A_1}\right)-\log\left(\frac{bp(b)-S_1}{S_1}\right)\right]\frac{1}{\hat{\pi}(2)}=\frac{\hat{\pi}(2)-\hat{\pi}(1)}{\hat{\pi}(2)},$$

where the left-hand side corresponds $\alpha(p(b),b)$ (see the proof of Claim B.4). Our goal is to establish that this equation has a unique solution in b.

We now study the behavior of $\alpha(p(b),b)$ over the support of b. Recall from Claim B.4 that function α is continuous for all $N_1^1-S_1=p(b)b-S_1>0$ and $N_1^0-A_1=1-p(b)b-A_1>0$. Moreover,

$$\lim_{bp(b)\to S_1}\alpha(p(b),b)=-\infty\quad\text{and}\quad\lim_{bp(b)\to(1-A_1)}\alpha(p(b),b)=+\infty,$$

where, since $p(b)b$ is increasing in b, it can be inverted to compute the lower and upper bounds on b such that $bp(b)\in(S_1,1-A_1)$. This then implies that a solution exists for all b such that $bp(b)\in(S_1,1-A_1)$.

Finally, we take the derivative of $\alpha(p(b),b)$ with respect to b and find that

$$\begin{aligned}\frac{d\alpha(p(b),b)}{db}&=\frac{(1-b)(p(b)+bp'(b))-bp(b)}{1-bp(b)}+\frac{(1-b)(bp(b))(p(b)+bp'(b))}{(1-bp(b))^2}\\&+\frac{(1-A_1-S_1)(p(b)+bp'(b))}{\hat{\pi}(2)(1-bp(b)-A_1)(bp(b)-S_1)}.\end{aligned}$$

After solving for the value of $p(b)$ such that $\beta_2(p,b)=S_1+S_2$, we plug it into $\frac{d\alpha(p(b),b)}{db}$ to establish that $\alpha(p(b),b)$ is increasing for $\hat{\pi}(2)$ sufficiently small. Combined with the analysis above that yields existence of a solution for b, it implies that the solution is unique. We denote this solution by b^*.

To obtain the corresponding value of p^*, we plug b^* into $p(b)$ defined in Claim B.3, and obtain p^* provided $b^*\geq\underline{b}$. Finally, the couple (p^*,b^*) must satisfy $p^*b^*\in(S_1,1-A_1)$. Q.E.D.

The claim proves that, given Claim B.4, the two-period model discussed in Section II.D is point-identified under constraints.

Model Solution for any $\overline{T}>2$ We now provide the sufficient conditions that allow us to extend the result in Claim B.5 to a general $\overline{T}$. This boils down to providing the conditions such that a solution to the system of two equations in two unknowns $p=p(b)$ and $b=b(p)$ exists and is unique. We do this in the following claim.

Claim B.6. *Let $\overline{T} > 2$ and $G(\varepsilon) = \frac{1}{1+\exp(-\varepsilon)}$. First, if a solution (b^*, p^*), with $b^* \geq \underline{b}$ and $p^* b^* \in (S_1, 1 - A_1)$, exists, it is unique if $p^* > h(A_1, \overline{T})$, where $h(\cdot, \overline{T})$ is the lower bound for p in period $\overline{T}$. Let $b^{min} \equiv \arg\min_b g(b)$ and $b^{max} \equiv \max\{\underline{b}, \arg\max_b g(b)\}$, with $g(\cdot) \equiv b^{-1}(\cdot)$. The duple (b^*, p^*) exists if and only if the following conditions are satisfied: if $g(b^{max}) < p(b^{max})$ (resp. $g(b^{max}) > p(b^{max})$) then $g(b^{min}) > p(b^{min})$ (resp. $g(b^{min}) < p(b^{min})$).*

Proof. To establish uniqueness conditional on existence, it is sufficient to prove the monotonicity of $p(b)$ and $b(p)$. We begin with $b(p)$. Taking the total derivative of Equation (B.12), we find that

$$\frac{db}{dp} = b'(p) \equiv -\frac{\frac{\partial \alpha(p,b)}{\partial p}}{\frac{\partial \alpha(p,b)}{\partial b}} = -\frac{b}{p}\frac{\frac{1-b}{(1-bp)^2} + \frac{1-A_1-S_1}{(1-bp-A_1)(bp-S_1)\hat{\pi}(2)}}{\frac{1-b(2-bp)}{(1-bp)^2} + \frac{1-A_1-S_1}{(1-bp-A_1)(bp-S_1)\hat{\pi}(2)}} < 0$$

for all $1 - b(2 - bp) > 0$ (which is implied by $p > 1 - A_1^2$).

As far as $p(b)$ is concerned, we prove in Claim B.4 that, for $\overline{T} = 2$, $p'(b) < 0$ if $p > \frac{2b-1}{b^2}$ (which holds true for $p > 1 - A_1^2$). For $\overline{T} = 3$, similar calculations show that $p'(b) < 0$ for all $p > \frac{(3-b)b-2}{b^2(3-2b)} > \frac{2b-1}{b^2}$. Because $\frac{(3-b)b-2}{b^2(3-2b)}$ increases in b, plugging $b = (1 - A_1)/p$ yields $p > h(A_1, 3) \equiv \frac{1}{4}(1 - A_1)(3 + 3A_1 + \sqrt{1 + A_1(2 + 9A_1)})$, with $\underline{p} = 1 - A_1^2 < h(A_1, 3) < 1$. Proceeding analogously for $\overline{T} = 4$, we find that $p'(b) < 0$ if $p > h(A_1, 4) > h(A_1, 3)$. Thus, we can conclude that, for $\overline{T} > 2$, there exists $h(A_1, \overline{T})$ such that $b(p)$ is strictly monotone for all $p > h(A_1, \overline{T})$.

We now develop the conditions for existence in the second part of the claim. First we note that, since $b(p)$ is monotone, there exists $g(\cdot) \equiv b^{-1}(\cdot)$. Then, we use the following facts:

- Function $p(b)$ is such that $p(\underline{b}) = 1$ and $p(1) = \sum_{t=1}^{\overline{T}} S_t$, with $t = 1, ..., \overline{T}$.
- The support of function $g(b)$ is given by $(S_1/p, (1 - A_1)/p)$. Moreover, $1 > g(b) > \underline{p}$.

Given these properties, existence is guaranteed if $p(b)$ and $b(p)$ cross over their domain, which holds true under the conditions in the claim. Q.E.D.

The claim accomplishes two things:

1. It establishes the condition guaranteeing the monotonicity of functions $p(b)$ and $b(p)$. Specifically, we define a lower bound for p, denoted by $h(A_1, \overline{T})$, with $h(A_1, \overline{T}) > h(A_1, \overline{T} - 1) > ... > h(A_1, 3) > 1 - A_1^2 = \underline{p}$, such that $b(p)$ and $p(b)$ are both strictly decreasing. This implies that, if a solution (b^*, p^*) exists, it is also unique.

2. Monotonicity alone only gives us that $p(b)$ and $b(p)$ cross at most once; that is, it is not sufficient to prove existence. Then, in the second part, the claim gives the conditions under which the functions cross over their domain.

C Likelihood Function with Censored Projects

We define a censored project in T as a project that has not been either abandoned or published by version T. This means, the project has not experienced any high cost shock, which would have implied the termination of the process. Moreover, a censored project might have experienced a breakthrough in any $\tau < T$ (implying the post-breakthrough phase in T). We denote the status of a censored project by $\sigma = \emptyset$, and the probability that a censored project reaches version T by $\Pr(T, \sigma_T = \emptyset)$. This probability is equal to:

$$\Pr(T, \sigma_T = \emptyset) = \frac{\Pr(T, \sigma_T = 1)}{1 - G^1(T)} + \frac{\Pr(T, \sigma_T = 0)}{1 - G^0(T)}$$

where $\Pr(T, \sigma_T = 1) = p \sum_{\tau=0}^{T-1} \rho(\tau, T)$ with $\rho(\tau, T)$ in Equation (13) is the probability that a project is published in T (with status $\sigma_T = 1$), and $\Pr(T, \sigma_T = 0)$ in Equation (15) is the probability that a project is abandoned in T (with status $\sigma_T = 0$). Moreover, we have

$$LL(T, \sigma_T = \emptyset | b, p, \boldsymbol{F}) = \log(\Pr(T, \sigma_T = \emptyset)).$$

We can then rewrite the log-likelihood of the data in Equation (17) as

$$LL(b, p, \boldsymbol{F}) = \sum_{i=1}^{N} LL(T_i, \sigma_{T_i} | b, p, \boldsymbol{F}) \tag{C.1}$$

with index i for a given project $i = 1, \ldots, N$ that reaches version T_i with status $\sigma_{T_i} \in \{0, 1, \emptyset\}$ in T_i. This expression for the log-likelihood accounts for all censored projects.

D Counterfactual Analysis

In this section, we show that the subsidy has a negative impact on the ex-ante values of projects as computed after deducting the subsidy transferred to the team and the cost borne by the team. We consider the two-period example without the post-breakthrough phase so that the project is published upon receiving consensus.[2]

Let the subsidy be denoted by s, then the net expected value of the project (including the paid subsidy) to the team is

$$V(s) = pb\left[\hat{\pi}(1) + (1-b)G(\bar{\varepsilon})\hat{\pi}(2)\right] + (1-pb)\,G(\bar{\varepsilon})\left[-F + s - E[\varepsilon|\varepsilon \leq \bar{\varepsilon}]\right],$$

with $0 < s < F$ and

$$\bar{\varepsilon} \equiv b\hat{p}(1,0)\hat{\pi}(2) - F + s, \quad \hat{p}(1,0) = \frac{p\,(1-b)}{1-pb}.$$

The first term in $V(s)$ corresponds to the expected benefits of a breakthrough, as given by the first period payoff, and the second period payoff conditional on the team deciding to continue in the first period. These terms are multiplied by the total probability that a project is good, and that breakthrough happens in one of the two periods. The expression in the second bracket corresponds to the expected opportunity costs of revising the project, multiplied by the probability that a project is not good, or a good project does not receive a breakthrough in any of the periods.

We further use $\widetilde{V}(s) \equiv V(s) - (1-pb)\,G(\bar{\varepsilon})s$ to denote the net expected value of the project to the organization as a whole (when it internalizes the team's costs and accounts for the cost of the subsidy). It is equal to the team's value net of the direct transfer flowing from the organization paying the subsidy to the team of authors. This expression can be rewritten as

$$\widetilde{V}(s) = pb\left[\hat{\pi}(1) + (1-b)\,G(\bar{\varepsilon})\hat{\pi}(2)\right] + (1-pb)\,G(\bar{\varepsilon})\left[-F - E[\varepsilon|\varepsilon \leq \bar{\varepsilon}]\right].$$

Note that, by construction, $\widetilde{V}(s) = V(0)$, that is, without a subsidy, the team's and organization's value coincide. We are interested in establishing the sign of $\widetilde{V}(s) - V(0)$, that is, the organization's value added from subsidy s.

Given that $V(0)$ is independent of the subsidy, the derivative of $\widetilde{V}(s) - V(0)$ with respect to s is $\frac{\partial\left[\widetilde{V}(s) - V(0)\right]}{\partial s} = \frac{\partial \widetilde{V}(s)}{\partial s}$ with

$$\begin{aligned}\frac{\partial \widetilde{V}(s)}{\partial s} &= \frac{\partial G(\bar{\varepsilon})}{\partial s}\left[pb\,(1-b)\,\hat{\pi}(2) - (1-pb)\left[F + E[\varepsilon|\varepsilon \leq \bar{\varepsilon}]\right]\right] \\ &\quad - (1-pb)\,G(\bar{\varepsilon})\frac{\partial E[\varepsilon|\varepsilon \leq \bar{\varepsilon}]}{\partial s} \qquad \text{(D.1)} \\ &= (1-pb)\left[\frac{\partial G(\bar{\varepsilon})}{\partial s}\left[\hat{p}(1,0)b\hat{\pi}(2) - F - E[\varepsilon|\varepsilon \leq \bar{\varepsilon}]\right] - G(\bar{\varepsilon})\frac{\partial E[\varepsilon|\varepsilon \leq \bar{\varepsilon}]}{\partial s}\right].\end{aligned}$$

[2] As we discuss below, introducing the post-breakthrough phase does not change the results in this appendix. The same holds with the analysis featuring $\overline{T} > 2$. In both cases, the qualitative nature of the conclusions remains, at the cost of making the calculations cumbersome.

Plugging

$$\frac{\partial G(\overline{\varepsilon})}{\partial s} = g(\overline{\varepsilon})\frac{\partial \overline{\varepsilon}}{\partial s} \quad \text{and} \quad \frac{\partial E[\varepsilon|\varepsilon \leq \overline{\varepsilon}]}{\partial s} = \frac{\partial \overline{\varepsilon}}{\partial s}\frac{g(\overline{\varepsilon})}{G(\overline{\varepsilon})}\left[\overline{\varepsilon} - E[\varepsilon|\varepsilon \leq \overline{\varepsilon}]\right]$$

into the expression in Equation (D.1), we obtain

$$\begin{aligned} \frac{\partial \widetilde{V}(s)}{\partial s} &= (1-pb)\, g(\overline{\varepsilon})\frac{\partial \overline{\varepsilon}}{\partial s}\left[\hat{p}(1)b\hat{\pi}(2) - F - \overline{\varepsilon}\right] \\ &= -(1-pb)\, g(\overline{\varepsilon})\frac{\partial \overline{\varepsilon}}{\partial s}s < 0. \end{aligned} \tag{D.2}$$

The second equality follows from the definition of $\overline{\varepsilon}$, and the inequality in Equation (D.2) follows from $\frac{\partial \overline{\varepsilon}}{\partial s} > 0$. Together with $\widetilde{V}(0) - V(0) = 0$, this property of $\widetilde{V}(s)$ implies that the net ex-ante value of the project to the organization falls with any value of the subsidy.

This result reflects the following considerations: in addition to a positive effect on the increase in expected benefits, the subsidy produces a negative effect from the additional costs borne by the team and a negative effect caused by an increase in the average cost shock borne by the team. We find that the first two forces cancel out, but the third remains—as given by the $g(\overline{\varepsilon})\frac{\partial \overline{\varepsilon}}{\partial s}$ term in Equation (D.2).[3]

[3]Note that adding the post-breakthrough phase is likely to reinforce the result. As we discuss in the main text, in the post-breakthrough phase, due to absence of discounting, the team's dynamic incentive to continue is only driven by the marginal increase in the payoffs, but the negative effects remain.

E Data Appendix

E.1 Construction of ID Sample

Our primary data source is the online archives of the IETF, which contain the full text for each version of every Internet Draft (ID) submitted after July 1990, along with various pieces of bibliographic information.

We begin by obtaining a list of all IETF IDs from the Internet Draft Status Summary file at

https://tools.ietf.org/id/all_id.txt.[4]

IDs in this Status Summary file are identified by a string of hyphenated alpha-numeric characters. For example, the ID name for RFC 7368 ("IPv6 Home Networking Architecture Principles") is

`draft-ietf-homenet-arch`.

All IDs begin with `draft`; the string `ietf` indicates that this particular ID is a working group ID (see below); the remaining pieces identify a specific proposal. Different versions of an ID are indexed by a suffix `-00` for the initial version and `-NN` for the NN's revision. The Status Summary file contains a list of the ID-version strings for the last version of an ID. For example, for RFC 7368, this last ID was `draft-ietf-homenet-arch-17` (where `-17` indicates the 17th revision of the initial proposal or the overall 18th version).

The total number of IDs on the list is 28,627. For each, we download the text document of all versions from

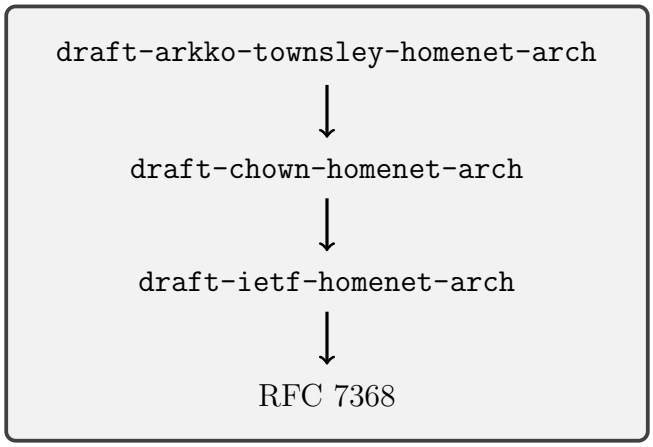
http://tools.ietf.org/id/[ID-version].txt

where `[ID-version]` is the respective ID-version combination. For example, for RFC 7368, we know the first version is `draft-ietf-homenet-arch-00`, and the last version is `draft-ietf-homenet-arch-17`. If available, we download the files from the first to the last version. Moreover, because the Status Summary file contains the date only for the last ID-version, we parse the history information for a given ID at http://datatracker.ietf.org/doc/[ID]/history (with `[ID]` the ID name without the version numbers) to obtain the online publication dates for each version. These are the dates a given version is posted to the online repository. Especially for older IDs, this date is not always the date the version was finished and circulated. We manually clean the date information in two steps. First, assuming that a version $t+1$ is not published before a version t, we find the set of versions for which the time difference between $t+1$ and t is negative (and version $t+1$ has a date before version t). We then manually inspect the draft document to find the correct dates of publication. If a complete publication date is not provided, we use the statutory expiration date (six months after the ID version is posted; if listed in the document) to backtrack the publication date; if the day of the month of the ID version is not given, we use the first of the month (or

[4] Downloaded on September 10, 2015. This date marks our data cut-off. In May 2024, the URL of the ID list is https://www.ietf.org/id/all_id.txt.

the date of previous ID version, whichever comes later); if no date is provided in the draft document, we use the date of the previous ID version.

IDs go through various stages of development and change their ID name along the way. For instance, an ID may be initiated by an individual IETF contributor and is later adopted by a working group (WG). Our example ID (RFC 7368) is such a case. Its original ID name was `draft-arkko-townsley-homenet-arch` (with one version). It was replaced (i.e., superseded) by ID `draft-chown-homenet-arch` (with two versions), before being adopted by the `homenet` WG as ID `draft-ietf-homenet-arch` (with 18 versions).



The Status Summary file contains the status of each ID, including information on the ID name that replaced a given ID. We use this information to link such chains of IDs to create a single *project*. In many cases, a subsequent ID supersedes an "expired" ID. As a statutory rule of the IETF, an unpublished ID expires after six months of inactivity (i.e., no submission of a new version), leading to its removal from "active" status. Authors, however, can reactivate and resume an expired ID. In our data, we see numerous cases where more than six months have passed between two versions of a given ID (see below for how we handle such cases). Likewise, we see numerous cases where more than six months have passed between two subsequent IDs in a project chain. For the construction of our project chains, we implement the following rule (referred to as the "24-months rule"). If less than 24 months have passed between the last version of an ID i and the first version of its successor (the ID $i+1$ that replaces the previous i in the chain), then we link the two IDs to obtain a single project chain. Instead, if more than 24 months have passed between the last version of an ID i and the first version of its successor $i+1$, we delink the two IDs. The first chain ends with the last version of the predecessor ID i, the second chain begins with the first version of the successor ID $i+1$. We thus delink 32 project chains.

After linking IDs, our sample contains 25,532 project chains with 96,770 versions. The longest chain links five IDs; 10 chains link four IDs; 205 chains link three IDs; 2,098 chains link two IDs, and 23,218 chains comprise only one ID.

We have applied our 24-months rule to delink chains if the gap between two presumably related IDs—recall, we obtain the linking information from the IETF's Status Summary file—is too big. Moreover, for many projects, we see considerable gaps between two versions

of the same ID. Gaps longer than six months are in violation of the IETF's statutory rule for expiration, but is fairly common. We therefore take a conservative approach when determining the de-facto status of a project chain. First, if there is a gap of more than 24 months within a project chain (i.e., within an ID of a given project chain), we drop all versions after the gap (a total of 1,780 versions in 451 chains). Second, project chains that have a version that was published less than 24 months before our data cut-off date (September 10, 2015) and that are not published as an RFC are considered "active." We thus get three different outcomes for our project chains:

Published: A project chain is "published" if it is published as an RFC or in the final stages (i.e., in the "Queue" but not yet assigned an RFC number); the full sample contains 5,834 published project chains.

Abandoned: A project chain is "abandoned" if it is not published as an RFC and no new version has been posted after September 10, 2013; the full sample contains 16,038 abandoned project chains.

Active: A project chain is "active" if it is not published as an RFC and a new version has been posted after September 10, 2013; the full sample contains 3,660 active project chains.

For the first version in each project chain, we take the first version for which we have a document. Thus, by construction, we have all initial versions for each project chain. We further miss 432 documents for later version numbers. For version-level information that we obtain from the ID documents (authors, text distance), we interpolate to account for these missing values.

After constructing our raw sample, we take the following steps to construct our estimation sample:

Step 1: Drop 456 project chains from six special technology areas. We provide more details in Section E.2.3 below.

Step 2: Drop 325 non-IETF projects: These projects are 19 IESG projects ("Internet Engineering Steering Committee"), 3 IANA projects ("Internet Assigned Numbers Authority"), 193 IRTF projects ("Internet Research Task Force"), and 110 IAB projects ("Internet Architecture Board").

Step 3: Drop 444 project chains with specialized RFCs. We provide more details in Section E.2.2 below.

Step 4: Drop 529 project chains initiated before 1996.

Step 5: Drop 3,526 active project chains (that means, active projects following the 24-months rule).

Step 6: Drop 4,161 completed (published or abandoned) project chains initiated in 2010 or later. We thus minimize selection on outcomes by dropping completed projects initiated in 2010 or later. Out of the 3,526 projects active at the cut-off date, and 3,479 (98.7%) projects were initiated in 2010 or later.

Table E.1: Sampling

Step	Description	Projects	Versions
	Full Sample (after applying the 24-months rule)	25,532	94,990
Step 1	Drop six areas	25,076	92,249
Step 2	Drop special non-IETF series	24,751	90,845
Step 3	Drop special tracks	24,307	88,187
Step 4	Drop projects initiated before 1996	23,778	86,757
Step 5	Drop active projects	20,252	73,654
Step 6	Drop completed projects initiated in or after 2010	16,091	57,179
	Final Estimation Sample	16,091	57,179

We summarize these steps in Table E.1. Our final sample consists of 16,091 project chains (3,857 published; 12,234 abandoned) with 57,179 versions. This is the sample used for all results but those presented in Column (4) in Table A.2. For this latter set of results, we use the information from right-censored observations. In other words, we do not drop active project chains (**Step 5**). Moreover, because by including active project chains we are not selecting on outcomes when keeping all projects initiated in 2010 or later, we do not drop this set of recent projects (**Step 6**). This extended sample consists of 23,778 project chains (4,781 published; 15,471 abandoned; 3,526 active) with 86,757 versions.

E.2 Further Information on IDs

E.2.1 Working Group

The file-naming convention of the IETF allows us to identify a project chain's relevant working group. IDs with names that begin with `draft-ietf` are IDs from a working group, and the working group name is the third part of the name. For our example project chain (RFC 7368), the third (and last) ID is from the `homenet` working group. We construct a subsample of all project chains that are initiated by a working group and refer to it as the working group sample (or WG sample). A project chain is in the WG sample if its first ID (in the chain) is from a working group. The WG sample consists of 3,932 project chains (2,206 published; 1,726 abandoned) with 22,025 versions. The project chains in this sample are initiated by 285 different working groups (both active and closed) and (at some point in time) part of 315 different working groups.

E.2.2 Standards-Track and Nonstandards-Track

For the distinction of standards-track and nonstandards-track project chains, we use information on the "status" of published RFCs. We first access the "RFC Index" at `http://www.rfc-editor.org/rfc-index.html` to obtain information on the title, the publication date, the status, area, and the working group of published RFCs. An RFC's status holds the information we need to identify a project chain's *track*. An RFC can be of one of a number

of statuses: "Best Current Practice," "Draft Standard," "Experimental," "Historic," "Informational," "Internet Standard," "Proposed Standard," and "Unknown." We obtain three track categories for RFCs:

> **No Track:** An RFC with status equal to "Best Current Practice" (184 projects) "Draft Standard" (104 projects) "Historic" (111 projects), "Internet Standard" (45 projects) or "Unknown." In **Step 3** above (see Table E.1), we drop these projects from our sample.
>
> **Nonstandards Track:** An RFC with status equal to "Experimental" or "Informational."
>
> **Standards Track:** All other RFCs.

To link RFC-level information to our sample of project chains, we access metadata for each RFC through

`http://datatracker.ietf.org/doc/rfc[NNNN]`

where `[NNNN]` is the four-digit RFC number. We parse the website to pull the last ID name that lead to the published RFC. We then match our RFC-level information with our project-chain information. For each published project chain, we thus obtain track information through its RFC status. We further assume that all other project chains (i.e., those that are abandoned in our final estimation sample) are standards-track project chains.

Our final estimation sample consists of 14,444 project chains on the standards track (2,210 published; 12,234 abandoned) and 1,647 project chains on the nonstandards track. Out of these nonstandards-track projects, 1,385 are "Informational" and 262 are "Experimental."

E.2.3 Areas

We obtain area information via two separate avenues. First, working groups are assigned to specific technology areas. We download information for area-working group assignments for active working groups from `https://datatracker.ietf.org/wg` and `https://tools.ietf.org/area` (these lists are overlapping) and for concluded working groups from `https://datatracker.ietf.org/group/concluded` and `https://tools.ietf.org/wg/concluded` (these lists are overlapping). We can thus assign an area to each project chain that has at some point been adopted by a working group. If a project chain has multiple areas, we choose the last area the project chain was assigned to.

For the non-working group sample (i.e., individual projects), we obtain some area information from `http://datatracker.ietf.org/doc/[ID]` (with `[ID]` the ID name): for 1,020 ID names, the information on Document Type contains the line "individual in `[A]` area" where `[A]` is an area identifier. This means, an individual ID (outside a working group) has been assigned to an area. In our estimation sample, we thus obtain area information for 712 additional project chains.

In **Step 1** above, we have dropped projects from six areas from the sample: 246 projects (1,256 versions) from the *gen* area ("General Area"), 160 projects (1,308 versions) from the *art* area ("Applications and Real-Time Area"), 35 projects (134 version) from the *usr* area,

3 projects (11 versions) from the *irtf* area ("Internet Research Task Force"), 1 project (6 versions) from the *adm* area, and 11 projects (26 versions) from the *ora* area. Moreover, we combine 60 projects from the *sub* area with the *rtg* area ("Routing Area") and the 53 projects from the *mgt* area with the *ops* area ("Operations and Management Area"). After further applying **Steps 2** through **6**, we arrive at the following area count in Table E.2:

Table E.2: Areas

		Projects		Versions
	Area Description	Published	Abandoned	
app	Applications Area	492	411	5,173
rai	Real-Time Applications and Infrastructure Area	496	204	5,583
tsv	Transport Area	419	190	4,272
int	Internet Area	599	388	6,121
rtg/sub	Routing Area	498	305	5,713
ops/mgt	Operations and Management Area	423	287	4,447
sec	Security Area	447	324	4,403
	Projects with area information	3,374	2,109	35,667
	Projects without area information	483	10,125	21,512

The IETF technology areas correspond roughly to the various layers in the engineering "protocol stack" as described in Simcoe (2012). From top to bottom, those layers/areas are: Applications (app), Realtime Applications and Infrastructure (rai), Transport (tsv), Internet (int), and Routing (rtg/sub). The IETF also recognizes two areas that cut across the various layers: Operations (ops/mgt) and Security (sec).

For the full sample, we have area information for 7,948 project chains. For our estimation sample, we have area information for 5,483 project chains.

E.2.4 Author Information

We use Jari Arkko's `authorstats` tool (https://www.arkko.com/tools/authorstats.html) to parse ID documents and extract author names, affiliations, and email addresses. The goal is to identify authors and their project history at the IETF.[5]

After running the `authorstats` tool on all documents of our full sample, we conduct some simple cleaning procedures. Using Python's name-entity-recognition libraries, we drop entries (for authors) in the output file that do not represent names. We further extract, where possible, author names from email addresses to confirm author names.

We then use two auxiliary data sets to further clean and complete the author information. First, we use the author names and organizations of authors on published RFCs. The information is manually collected (Simcoe, 2012) and serves as a reliable source. We merge the RFC author data with our project chain data to correct and complete the name information for project chains associated with an RFC. Second, we use the attendance lists of

[5]For a brief analysis of authors' affiliations, see https://www.arkko.com/tools/docstats.html.

IETF meetings (meeting 29 in April 1994 through meeting 93 in July 2015) to construct a list of IETF contributors/members that serve as the list of potential authors. For most of the meetings, we obtain author names, affiliations, and email addresses. Email address information allows us to fill gaps in the author information for project chains. In the last step, we conduct basic author-name disambiguation to obtain unique author identifiers.

Using unique author-name identifiers for project chains, we obtain an author's project history at the IETF. For our measure of team size for a project chain, we count the number of authors on the first version of the project. We further construct two measures of author experience. First, for a given project chain, we count an author's number of successfully completed project chains (i.e., published as RFCs) before that project chain is initiated. At the project-chain level, we then use the experience of the most *successful* author of the author team (if multiple authors are listed) of the initial draft. Second, for a given project chain, we count an author's number of completed project chains (i.e., published or abandoned) before that project chain is initiated. At the project-chain level, we then use the experience of the most *prolific* author of the author team (if multiple authors are listed) of the initial draft. Because we have, by construction, all first-version documents, author information is missing for project chains only if the `authorstats` tool is not able to parse the document.

E.2.5 Patent Citations

We use the number of citations from U.S. patents as a proxy for commercial impact. In U.S. patents, IETF RFCs or IDs are cited in the non-patent literature (NPL) references. We obtain the list of non-patent literature references for all U.S. utility patents granted between 1976 and February 2016 from http://www.patentsview.org/download/ (file: otherreference.zip). We count the number of citations of a given project chain from patents in three steps:

Step 1: Using regular expressions that help us take typos and formatting errors into account, we search the NPL references for entries in the form of ID names (e.g., `draft-ietf-homenet-arch`). We pre-sample and use for the search only NPL references that include the terms "IETF," "internet," or "draft" (we use regular expressions to account for differences in capitalization). We use the full sample of ID names (from all project chains) as our search sample. When possible, we also extract the version number that is cited.

Step 2: Using regular expressions that help us take typos and formatting errors into account, we search the NPL references for entries in the form of RFC numbers (e.g., "RFC 7368"). We pre-sample and use for the search only NPL references that include the terms "RFC," "IETF," "internet," or "draft" (we use regular expressions to account for differences in capitalization).

Step 3: As part of the initial download of ID and RFC metadata, we also obtain the titles of all IDs and RFCs. We search the NPL references for these titles, using only titles that have at least four words and 10 characters. We pre-sample and use for the search only NPL references that include the terms "IETF," "internet," "RFC,"

"Request For Comment(s)," or "draft" (we use regular expressions to account for differences in capitalization).

These three approaches result in overlapping lists of citing patents for a given project chain. We therefore tally the number of unique patents that cite the given project chain in its NPL references. Our approach (counting not only citations to RFC numbers but also including ID identifiers) allows us to capture citations from U.S. patents that were made while the project chain was still active (before publication).

The project chain-level data we obtain from the above procedure is the basis for our patent citation estimates $\hat{\pi}(t)$. For our main results in Table 2, we use the patent citations for successful project chains (published RFCs) in the estimation sample to estimate $\hat{\pi}(t)$.

E.2.6 RFC Citations

In addition to patent citations, we also count the number of RFC citations each project chain receives. Each RFC contains a list of references that typically contains several citations to previously published RFCs. We download the RFC documents via `http://www.rfc-editor.org/rfc/rfc[NNNN].txt` (where `[NNNN]` is the RFC number)[6] and scrape the documents to obtain a list of each citation. We construct a count of RFC citations that is equal to the number of times a given RFC is cited by some other RFC published at a later date. We then link the citations for a given RFC with our project-chain data. Unlike the data for patent citations, our numbers for RFC citations do not include citations to the unpublished RFC (while the project chain was still active).

As with patent citations, we think of RFC citations as an indication that the citing references build on the ideas contained in the cited references, so that citations serve as a measure of the cited RFC's technological importance in a cumulative innovation context. Simcoe (2012) shows that RFC and U.S. patent citations are highly correlated. For our updated data (for both RFC citations and U.S. patent citations), the correlation coefficient is 0.61.

E.2.7 Text Distances

The goal is to calculate the textual distance (or, dissimilarity) of an ID version t from the initial version. For project-level information, we use the distance of the last version, $t = T$, from the initial version, $t = 1$. We use R text-analysis libraries to load and preprocess the documents. Preprocessing steps are: drop punctuation, drop stop words, drop numbers, convert all text to lower case, and eliminate excessive white space. We then create document-term matrices. Each row i of such a matrix represents a document, and each column j represents a term in the corpus (the collection of all documents). The value in cell (i, j) is the number of times a term j is used in document i. A document-term matrix gives us vector-space representation of a document i, where a document is a vector $x_i = (x_{i,1}, \ldots, x_{i,n})$ of term frequencies, with n unique terms in the corpus.

As our measure of textual distance (or dissimilarity) $\text{dist}(T, 1)$ between document $t = T$ and $t = 1$ we use the *cosine dissimilarity* (we refer to it as *cosine distance*). More specifically,

[6] Our data sample is as of June 2017.

we calculate 1 − cosine similarity, where the latter is the cosine of the angle between two term-frequency vectors x_T and x_1, so that the cosine distance is

$$\text{dist}(T,1) \equiv 1 - \frac{x_T \cdot x_1}{||x_T||\ ||x_1||} \quad \text{with} \quad ||x_i|| = \sqrt{\sum_{l=1}^{n} x_{i,l}^2}\,. \tag{E.1}$$

E.3 Emails

We use emails exchanged among IETF contributors (authors and members of the community who provide feedback) to construct two variables:

Email Count: The number of emails sent in response to a given version t of a project chain. At the version level it is, for version t, the number of emails sent in response to the previous version $t-1$. At the project chain-level level it is, for project chain i, the average number of emails sent in response to any version of i.

Commerciality: The suit-to-beard ratio as the share of corporate email addresses (from which emails are sent) over all email addresses.

We construct our email data in a number of steps.

Step 1: We first download the "IETF Mail Archive" and the "Concluded WG IETF Mail Archive" via `ftp://ftp.ietf.org/`.[7] These email archives contain all emails sent to various IETF email lists. We obtain the emails in monthly batch files. We use a Python script to first split the monthly batch files into individual emails.

Step 2: We extract the email address from which an email was sent and the date and time it was sent.

Step 3: In order to assign a given email to a given version of a project chain, we search for each ID name (from the complete list of all ID names) in the text of that email. Any email that mentions an ID name is considered to be sent in response to that ID. We also extract the version number of that listed ID name (if possible). The email archive provides us with emails from all email lists, including organizational lists and the `i-d-announce` lists (in its various forms: before 1998, between 1998 and 2004, and after 2004). A new version of an ID is announced to the community at large via this email list. It provides a very clean account of the date the version is posted (with an occasional delay of a day or two relative to the posted date listed on the draft). We do not use ID announcement (or other organizational emails) for our email counts, but use the date information (for each ID version) to fill missing version numbers in the next step.

Step 4: If a version number (for the email assignment) is not available, we take two approaches to fill the gap. First, if an email j without an assigned ID version number

[7] Our email data is as of October 17, 2015.

was sent between two emails that do have the same version number, we use the ID version number of the two embracing emails. Second, if an email j without an assigned ID version number was sent between two consecutive ID announcement emails, we use the earlier email's ID version number as the assigned ID version number for email j.

Step 5: We delete all ID announcement emails and other organizational emails so as to not include them in email counts. These emails do not contain any information pertinent to the content of the ID.

Step 6: In a last step, we delete all duplicate entries (where the same email is assigned to the same version multiple times). This does not mean that we delete duplicate emails, because an email j can be assigned to two or more project chains. Such an email enters our email database multiple times. We also delete all emails sent before the first ID version was posted and all emails sent more than 24 months after the last ID version was posted.

We have at least one email for 20,532 project chains out of 25,532 project chains in the full sample (a total of 645,064 emails); and at least one email for 12,396 project chains out of 16,091 project chains in the estimation sample. This number amounts to 331,094 emails in the estimation sample.

The emails in the full sample are sent from 20,968 unique addresses (from 7,533 domains). The earliest email in the full sample is from July 6, 1990. The last email is from October 17, 2015. The emails in the estimation sample are sent from 15,404 email addresses (from 5,775 domains). The earliest email in the estimation sample is from January 4, 1996, the last email from October 4, 2015 (assigned to a project chain that was published as RFC).

To construct the suit-to-beard ratio, we categorize email addresses as corporate if their top-level domain (TLD) is `com`, `net`, or `biz`. If the TLD is a country domain, we use `.co.` as an identifier for corporate email addresses. In order to avoid counting email addresses from email provider services, we compile a list of more than 3,000 free (and for-cost) email providers. Many of these use, for instance, `com` as their TLD. We do not categorize such email addresses as corporate.

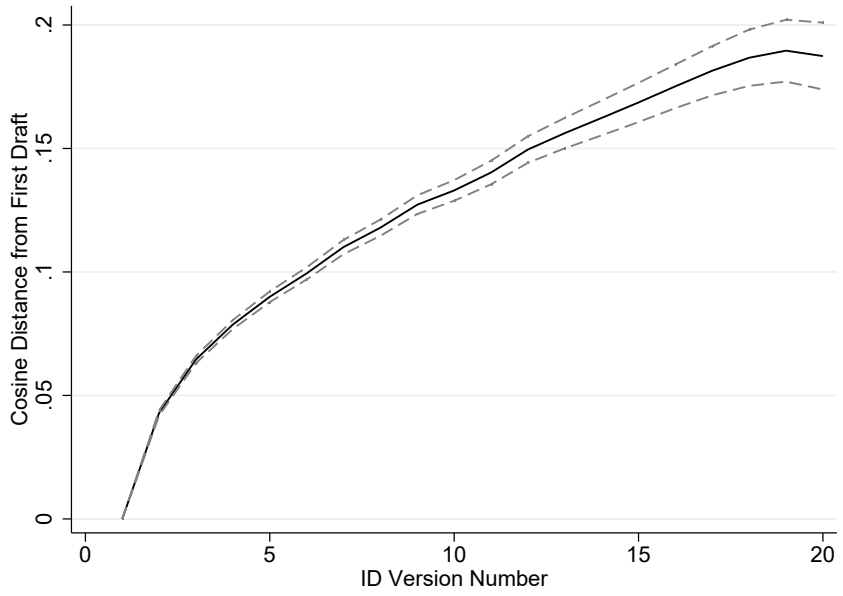
The suit-to-beard ratio for a given project chain is the number of corporate email addresses from which the project chain's assigned emails are sent, divided by all email addresses of that project chain. We have at least one email address (and are able to construct the suit-to-beard ratio) for 10,710 standards-track projects (out of 14,444 standards-track projects in the estimation sample).

F Descriptive Analysis

To motivate our theoretical model of learning, we now present several stylized facts about the internet standard setting process. This part of the analysis has three objectives: (1) establish that proposals are changing over time, (2) characterize the hazard rates for success and abandonment, and (3) link revisions with commercial significance.

Fact 1: Revision Distance. Figure F.1 shows how projects change during the revision process.[8] To create the figure, we represent each version by a vector containing the frequencies of the terms used in that version (that is, we use a "bag-of-words" approach for document representation). We use the cosine distance as our textual distance metrics and calculate that distance between a given version and the original submission, and then plot the mean distance (conditional on the number of revisions).

Figure F.1: Cosine Distance to First Draft



Notes: The figure presents the textual distance (i.e., cosine distance as one minus the cosine similarity) of an ID version number t from the initial version. Representing a document by a vector x_i containing the frequencies of terms used in version i, the cosine distance of a version t from the initial version 1 is defined as $1 - \frac{x_t \cdot x_1}{\|x_t\|\|x_1\|}$. See Appendix E for more details, including text preprocessing steps.

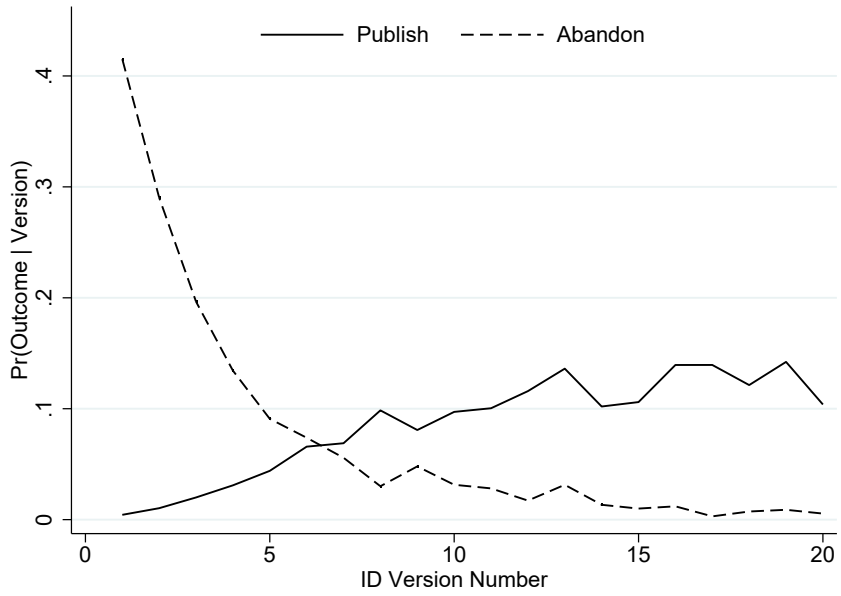
Although it is hard to give a natural interpretation to this distance metric, the figure suggests that each revision takes the proposal further from the initial submission, with the rate of change rapid at first, and gradually diminishing.

Fact 2: Hazard Rates. Figure F.2 plots the hazard rate (i.e., probability of exit, conditional on survival to version t) of the two possible outcomes for each Internet Draft: aban-

[8] We use the terms "revision" and "version" interchangeably. The number of revisions, therefore, includes the very first version of the draft (and is thus equal to the total number of versions).

donment or publication as an RFC. The most striking aspect of the figure is the rapidly decreasing hazard of abandonment. Almost 40 percent of proposals are never revised, but conditional on reaching a fifth version, the probability of abandonment falls below 10 percent. At the same time, the hazard of becoming a Proposed Standard climbs from zero on initial submission to about 10 percent by the eighth version. Together, these two hazard rates imply that the probability of publication increases sharply with the number of revisions. The increasing odds of publication are consistent with the arrival of "good news" over time, which will be a feature of our learning model.

Figure F.2: Hazard Rates



Notes: The figure presents the hazard rates for all standards-track proposals. The probability of publication in t (conditional on survival to that version t) is depicted by the solid line, the probability that a proposal is abandoned in t (given survival to t) is depicted by the dashed line.

Another feature of Figure F.2 is that the publication hazard never climbs above 16 percent, even though the hazard of abandonment falls to nearly zero. This suggests that author-teams continue to revise proposals even after it becomes clear that they will be published, and it leads us to propose a two-phase model rather than a theory where consensus leads to immediate publication.

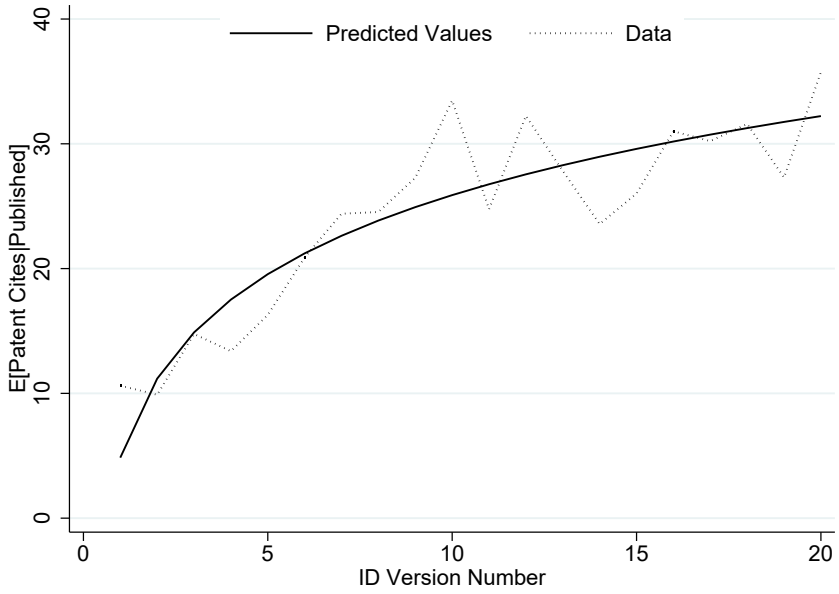
Fact 3: Revision and Citation. The final piece of our descriptive analysis examines the relationship between revisions and U.S. patent citations to published RFCs (our proxy for commercial impact).[9] Specifically, we estimate linear regressions of the form

$$Cites_i = \alpha_y + \beta_1 \times \log(Versions_i) + \varepsilon_i \tag{F.1}$$

[9] Roach and Cohen (2013) provide a discussion of measurement issues associated with counting non-patent prior art references in U.S. patents, and suggests that these citations are a superior measure of knowledge reuse than more commonly used metrics based on citations to other patents.

where $Cites_i$ is a count of citations to published RFC i, and α_y are publication-year effects that control for right-censoring in the citation process. Figure F.3 plots fitted values for standards-track RFCs, assuming a publication year of 2000. The dotted lines correspond to estimates from a non-parametric model where we replace $\log(Versions_i)$ with dummies for each value of the variable $Versions_i$.

Figure F.3: Expected U.S. Patent Citations



Notes: The figure presents the predicted number of patent citations of published RFCs. The solid line depicts the fitted values for the model in Equation (F.1) that log-linear in versions; the dotted line depicts fitted values from a non-parametric model with dummies for each version number.

The main take-away from this analysis is that citations increase (at a declining rate) with the number of revisions. We use fitted values from these regressions as our measure of expected payoffs in the structural model described below. Thus, in the absence of monetary measures of cost and benefit, citations will serve as the numeraire, providing a unit of measure for teams' opportunity costs and the evaluation of policies in our simulations.

References

Roach, Michael, and Wesley Cohen. 2013. "Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research." *Management Science*, 59(2): 504–525.

Simcoe, Timothy. 2012. "Standard Setting Committees: Consensus Governance for Shared Technology Platforms." *American Economic Review*, 102(1): 305–336.