

Standard Setting Committees:
Consensus Governance for Shared Technology Platforms

Web Appendix

Timothy Simcoe

April 4, 2011

A1 Data Set Construction

Dependent Variables

Time-to-Consensus

Data on every Internet Draft published between 1993 and 2003 was obtained from two sources: 1) the “ietf-announce” mailing list, and 2) www.watersprings.org. Publication dates come from the ietf-announce list in over 90 percent of the cases, and the watersprings site was used to fill in dates when one version in a particular series was missing.

Citations

I collected citation data from three sources. RFC citations were gathered by using a Perl program to examine the reference section of an ASCII text copy of every published RFC. The complete RFC archives are available at www.rfc-editor.org. Patent citations were collected from the USPTO website, where I searched for all non-patent prior art citations containing the string “RFC” or “Request for Comments” followed by a four digit number. Finally, academic journal citations were collected by performing a similar cited-reference search using all journals in the ISI Web of Science database. These searches were conducted in July 2008.

Independent Variables

Internet Drafts

The IETF’s file naming conventions were used to match each Internet Draft to a particular Working Group. In a few cases where individual ID’s were later adopted by a Working Group, the two series were matched by hand. Author attributes were collected by using a Perl script to parse the header and acknowledgements section of each ID. Similarly, key word counts were obtained by using Perl to scan an ASCII text copy of the proposal.

E-mail Addresses and Working Group Discussion Lists

Data on committee demographics were obtained from the archived e-mail discussion lists of 278 IETF Working Groups. Many of these can be found at ftp.ietf.org/ietf-mail-archive, and the remainder were located by searching the Internet. Collectively, these data go back as far as the late 1980’s, and comprise more than 690,000 messages. Most Working Group communication takes place on these e-mail discussion lists.

I used Perl scripts to examine each message and construct measures of WG demographics, participation, and experience. All of these variables are based on information contained in message headers—specifically the date, sender’s address, and subject line fields. (Mark Overmeer’s Mail::Box perl modules were invaluable for processing the email data.) I used a number of well known domain-naming conventions to classify the institutional affiliation of users from less common top-level domains (e.g. bt.co.uk, rotman.utoronto.ca, or alvestrand.no).

I used several approaches to address the problems of spam and hosted mail. First, I exclude all messages with subject lines related to pornography, home mortgages, hot stocks, or exciting new business opportunities\$!\$!\$! Second, I removed messages originating from the most popular hosted mail sites (e.g. Yahoo! and Hotmail). While this criteria may drop some legitimate messages, I found that most IETF participants have several e-mail accounts—one of which was generally within the domain of their employer. Finally, the results in the paper are based on a sample of messages whose sender (originating address) appeared four or more times on the same list on different dates with different subject lines. I also constructed a sample based on messages that were part of a discussion thread, i.e. either generated a reply or replied to an earlier message. All of the results are robust to a variety of changes in these rules and the criteria used to screen messages.

The e-mail data can be aggregated at three levels: message, sender (unique address), or organization (unique top-level domain). For all variables, the correlation across these different levels is extremely high, e.g. Suit-share measures based on messages, participants, and firms are all correlated above 0.98. Consequently, the results do not change if I change the aggregation level for a particular variable, but they do become unstable when I try to include all three levels.

The NBER Patent Data Merge

The organizational affiliation of each Internet Draft author is identified using an email address from inside the ID. This approach identifies 1,460 organizations with one or more contributions to an IETF Working Group. I focus on 498 organizations that appear on two or more Internet Drafts, and attempted to match each organization to a standardized patent assignee number by hand.¹

I successfully matched 193 organizations to an USPTO assignee code. All of the top 100 IETF contributors were either matched or determined to be non-US patent holders. Many of the unmatched organizations were non-profits (e.g. the World Wide Web Consortium), network operators (e.g. MERIT) or non-US academic institutions. Because of the skew in contribution rates, one or more of the matched organizations appeared as an author on over 90 percent of the proposals in the estimation sample. However, some small patenting firms may not show up in the data because of lags in the US Patent and Trademark Office’s patent review and reporting process.

For each firm in the matched group, I construct a five year unadjusted patent stock, and a five year stock depreciated at 15 percent. I also calculated the uncentered correlation coefficient

¹See <http://elsa.berkeley.edu/~bhhall/pat/namematch/namematch.html> for a discussion of assignee names and numbers.

over 3-digit USPTO technology classes between the firm’s patent portfolio and the total stock of patents owned by IETF participants. The $\log(\text{Patents})$ variable used in the analysis is the sum (over all firms with one or more proposals before a Working Group) of the five-year depreciated patent stock weighted by the firm-specific correlation coefficient.

A2 Propensity Score Matching

I use fitted values $\widehat{P}(x)$ from a probit model of standards-track assignment to estimate the propensity score. Estimates from this probit model are presented in Table A1. Figure A1 shows the empirical distribution of the estimated propensity-score for standards and nonstandards in the estimation sample. The solid vertical lines correspond to the 5th percentile of the estimated propensity score distribution for Proposed Standards (P_5^s) and the 95th percentile of the propensity score distribution for nonstandards-track RFCs (P_{95}^n). (The dotted lines correspond to the 1st and 99th percentiles.) The region where $P_5^s \leq \widehat{P}(X_i) \leq P_{95}^n$ is the common support of the p-score distribution. Discarding observations that fall outside this range (i.e. “trimming” or “blocking” on the propensity score) leaves 952 out of the original 1,166 RFCs, or 82 percent of the initial sample.

Figure A1: Propensity Score Distribution

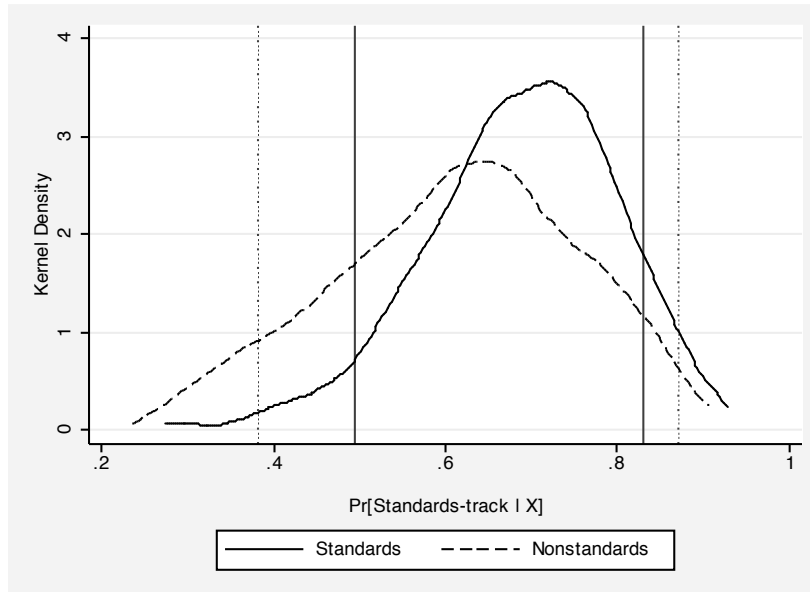


Table A1: Propensity Score Probit

	Probit Regression DV = S-Track	
	Coefficients	Marginal Effects
Suit-share	0.01 (0.00)**	0.00 (0.00)**
Lag Suit-share	-0.01 (0.00)**	-0.00 (0.00)**
log(Patents)	-0.04 (0.02)*	-0.02 (0.01)*
log(Drafts)	-0.90 (0.45)*	-0.32 (0.16)*
log(Cum Drafts)	0.65 (0.29)*	0.23 (0.10)*
log(Members)	0.11 (0.08)	0.04 (0.03)
log>Email)	-0.01 (0.05)	-0.00 (0.02)
log(Cum Email)	0.06 (0.12)	0.02 (0.04)
log(Cum Drafts) ²	-0.10 (0.05)*	-0.04 (0.02)*
log(Cum Email) ²	-0.00 (0.01)	-0.00 (0.00)
log(Drafts) ²	0.23 (0.11)*	0.08 (0.04)*
Sponsors	-0.06 (0.03)+	-0.02 (0.01)+
WG Chair	0.19 (0.09)*	0.07 (0.03)*
log(Filesize)	0.14 (0.05)**	0.05 (0.02)**
Dot-org	0.17 (0.16)	0.06 (0.05)
Dot-edu	-0.17 (0.10)+	-0.06 (0.04)
Dot-gov	-0.09 (0.17)	-0.03 (0.06)
Tech Class Effects	5 [0.01]	
ID Cohort Effects	2 [0.34]	
WG Cohort Effects	2 [0.22]	
Pr(S-Track=1)	0.67	
Observations	1166	
Model dof	26	

Model includes technology class effects and quadratic in draft-year and WG cohort.

A3 Additional Descriptive Statistics

Table A2: Top IETF Contributors[†] (1992-2004)

<u>1992-1994</u>		<u>1992-2004</u>	
1. Cisco	94	1. Cisco	1,787
2. Carnegie Mellon	51	2. Nortel	694
3. mtview.ca.us	48	3. Microsoft	581
4. IBM	44	4. Nokia	539
5. SNMP Research	38	5. Sun Microsystems	513
<u>1995-1997</u>			
1. Cisco	214	6. AT&T	513
2. IBM	140	7. IBM	490
3. Microsoft	140	8. Ericsson	398
4. Sun Microsystems	84	9. Lucent	343
5. USC (ISI)	79	10. Bell Labs	301
<u>1998-2000</u>			
1. Cisco	517	11. Alcatel	299
2. Nortel	321	12. Juniper Networks	260
3. AT&T	223	13. Intel	225
4. Microsoft	221	14. Columbia U.	220
5. Sun Microsystems	180	15. Siemens	200
<u>2001-2004</u>			
1. Cisco	962	16. Dynamicssoft	196
2. Nokia	404	17. USC (ISI)	195
3. Nortel	354	18. ACM	185
4. Ericsson	279	19. MIT	152
5. Sun Microsystems	234	20. NTT	149

[†]Rankings are based on the number of Internet Drafts submitted during the relevant period.

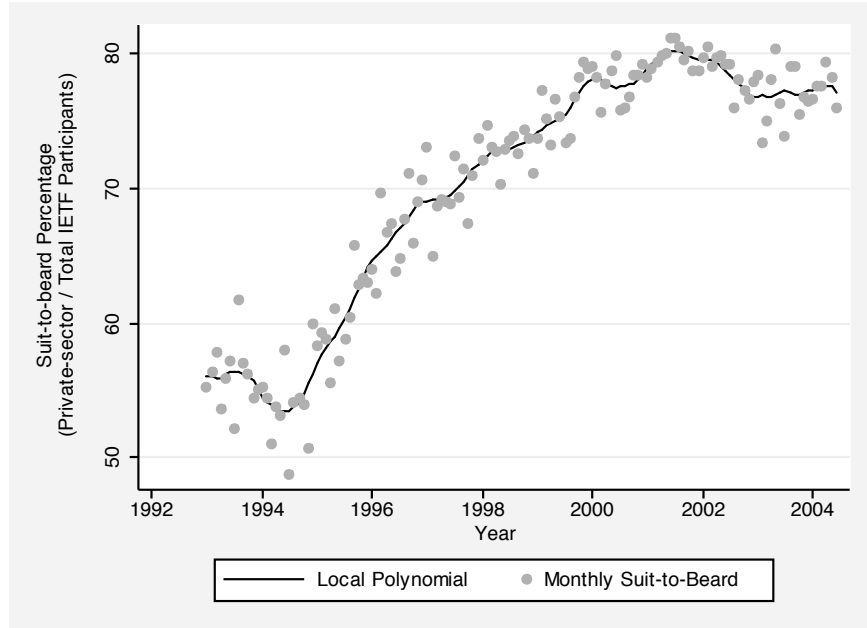


Figure A2: IETF Commercialization (Monthly Suit-share)

Table A3: Sample Construction

	Individual Drafts	Working Group Drafts [†]	One or more Revisions	Estimation Sample [‡]
	Total Observations			
Total Internet Drafts	6,481	3,521	2,662	2,601
Working Groups	0	283	272	249
	Internet Draft Outcomes			
Censored	47	62	62	62
Expired	5,729	1,730	950	933
Nonstandards-track RFC	534	641	586	553
Proposed Standard	171	1,088	1,064	1,053

[†]Excludes IDs from General and User Service Areas, IDs originating before 1993 or after 2003, BCP's, Historic RFC's, Draft Standards and Internet Standards.

Table A4: WG Publication (RFC) Counts[†]

Proposed Standards	Nonstandards-track RFCs						Total
	0	1	2	3	4	≥ 5	
0	25	27	9	4	2	6	73
1	21	12	5	3	1	4	46
2	9	11	1	2	0	4	27
3	8	6	1	2	1	3	21
4	7	4	2	2	0	0	15
≥ 5	9	11	15	4	5	23	67
Total	79	71	33	17	9	40	249

[†] Each cell contains a count of WGs that published the number of standards (nonstandards) indicated by the row (column) headings.

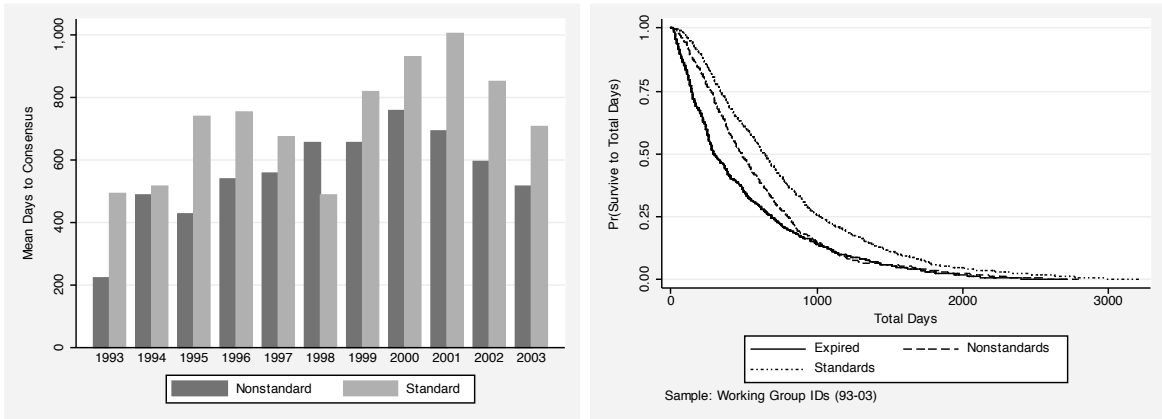


Figure A3: Mean Duration by Cohort (left) & Survival Curves (right)

A4 Switching Model

I estimate a switching model with partially observed regimes. The model has three equations:

$$\begin{aligned} T_i^m &= X_i\beta^m + \varepsilon_i^m \\ S_i &= 1[Z_i + \nu_i] \end{aligned}$$

where $m \in \{s, n\}$ indexes standards and nonstandards respectively, and the unobservables have a trivariate normal distribution $(\varepsilon^s, \varepsilon^n, \nu) \sim N(0, \Sigma)$. The correlation between ε^s and ε^n is undefined, since we never observe both T^s and T^n . Following the approach described in Lokshin and Sajaia (2004), I define

$$\eta_{im} = \frac{\gamma Z_i + \rho_{mv}\varepsilon_i^m/\sigma_m}{\sqrt{1 - \rho_{mv}^2}}$$

where $\rho_{mv} = \sigma_{mv}/\sigma_m$ is the coefficient of correlation between ε^m and ν . Since the unobservables have a trivariate normal distribution, there is a simple closed form expression for uncensored observations (i.e. Proposed Standards and nonstandards-track RFCs).

$$\begin{aligned} Pr(S_i = 1, T_i^s = T) &= \frac{\Phi(\eta_{is})}{\sigma_s} \phi\left(\frac{\varepsilon_i^s}{\sigma_s}\right) \\ Pr(S_i = 0, T_i^n = T) &= \frac{(1 - \Phi(\eta_{in}))}{\sigma_n} \phi\left(\frac{\varepsilon_i^n}{\sigma_n}\right) \end{aligned}$$

I treat expired proposals as censored observations whose intended publication status is unknown. Since no proposal is on both the standards- and nonstandards-track, the probability of observing a censored or expired proposal must be

$$Pr(S_i = 1, T_i^s > T_i) + Pr(S_i = 0, T_i^n > T_i) = \Phi\left(\gamma Z_i, \frac{-\varepsilon_i^s}{\sigma_s}, \rho_{sv}\right) + \Phi\left(-\gamma Z_i, \frac{-\varepsilon_i^n}{\sigma_n}, -\rho_{nv}\right)$$

where Φ is the cumulative bivariate normal distribution with correlation parameter ρ .

The log-likelihood is $\sum_i \ln(Pr(S_i, T_i))$. Code for estimating this model in Stata was adapted from the movestay routine developed by Lokshin and Sajaia (2004) and is available on the author's web site: <http://people.bu.edu/tsimcoe>.

A5 Additional Results

Table A5: Instrumenting for Suit-share

	Obs = Internet Draft DV = Total Days			Obs = WG-Year-Track DV = Avg. Total Days			
Model	GMM IV (1)	GMM IV (2)	GMM IV (3)	OLS (4)	OLS (5)	Arellano Bond (6)	Arellano Bond (7)
Suit-share * S-track	5.7 (2.2)**	9.1 (2.9)**	15.9 (11.2)	4.8 (1.7)**	3.7 (4.3)	10.5 (5.7)+	11.5 (8.8)
Suit-share		-3.3 (2.4)		-1.8 (2.8)	-0.1 (4.0)		0.5 (7.3)
	First-Stage Statistics						
<i>Instruments</i>							
Lag-Suit-share	Y	Y	N				
Lag-Share * S-track	N	Y	N				
Non-WG Tech Area Growth	N	N	Y				
<i>F-Test of Excluded IVs</i>							
Suit-share * S-track	203.5 [0.00]	286.5 [0.00]	2.75 [0.07]				
Suit-share		147.5 [0.00]					
	Additional Controls & Statistics						
Tech Class FEs	Y	Y					
PubCohort FEs	Y	Y	Y	Y	Y	Y	Y
Working Group FEs	N	N	Y	Y			
WG-Track FEs	N	N	N	N	Y	Y	Y
Additional Controls	Y	Y	Y	Y	Y	N	N
Obs (S-track)	773	667	737	385	385	118	118
Obs (N-track)	0	289	0	281	281	0	51

+10% *5% **1% ***0.1% significance levels (clustered on Working Group).

Notes: Models (1) through (3) take the RFC as a unit of observation and instrument for Suit-share using lag Suit-share or the growth rate of other WGs in the same Technology Area. Models (4) through (7) collapse the data to WG-Year-Track level. All models omit observations with Total Days > 2007 (5.5 years) to correct for right-truncation of the DV. Models(1) and (2) use the matched sample for comparison to Table 4. Additional controls in (1) through (3) are main effects and S-track interactions for main effects and S-Track interactions for: log(Filesize), log(Drafts), log(WG Mail), log(Cum Mail), Dot-gov, 1[ID Sponsors=2], and 1[ID Sponsors>2]. Additional controls in (4) and (5) are main effects and S-Track interactions for: log(Filesize), log(Drafts), log(WG Mail), log(Cum Mail), Dot-gov, 1[ID Sponsors=2], and 1[ID Sponsors>2].

Table A6: Diff-in-diffs: Alternative Measures and Specifications

Model	Big-Firm Effects		ID Suit-share		Poisson QML		Cox Hazard		75th Quantile Regression	
Suit-share * S-track	7.34 (2.00)***	6.09 (2.18)**			0.66 (0.19)***	0.51 (0.22)*	-1.79 (0.47)***	-2.10 (0.45)***	6.58 (2.27)**	
log(Patents) * S-track	4.42 (10.50)	-1.59 (11.65)	2.03 (10.56)	3.88 (11.13)	0.00 (0.02)	0.00 (0.02)	0.03 (0.02)	-0.00 (0.03)	9.16 (11.18)	
Org-edu * S-Track	171.43 (75.87)*	76.16 (84.10)	117.54 (73.08)	85.94 (74.13)	0.25 (0.13)+	0.16 (0.14)	-0.33 (0.16)*	-0.13 (0.17)	87.70 (70.01)	
Suit-share	-1.46 (1.69)	0.13 (2.77)			-0.16 (0.16)	0.06 (0.25)	0.98 (0.38)*	0.17 (0.54)		-2.21 (3.33)
log(Patents)	3.55 (8.46)	11.71 (12.53)	0.72 (8.05)	13.22 (12.35)	0.01 (0.01)	0.01 (0.02)	-0.04 (0.02)*	-0.03 (0.03)		-13.75 (19.41)
Org-edu	-45.58 (80.12)	-48.14 (95.19)	-18.96 (66.93)	-4.64 (76.06)	-0.11 (0.13)	-0.09 (0.15)	0.19 (0.14)	0.06 (0.15)		-57.64 (99.65)
ID Suit-share * S-track			3.18 (1.41)*	3.09 (1.34)*						
ID Suit-share			1.84 (1.15)	3.48 (1.42)*						
Tech Class Effects	Y	na	Y	na	Y	na	Y	na	Y	Y
WG Effects	N	Y	N	Y	N	Y	N	Y	N	N
PubCohort Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Large Firm Effects	Y	Y	N	N	N	N	N	N	N	N
Additional Controls [†]	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Obs. (Standards)	662	662	662	662	662	662	935	935	773	0
Obs. (Nonstds)	290	290	290	290	290	290	474	474	0	393
Obs. (Censored)							878	878		

+10% *5% **1% ***0.1% significance levels (clustered on Working Group).

Notes: Big-Firm, ID Suit-share and Poisson models are estimated on the Matched Sample. Cox Hazard model assumes all expired proposals are censored standards-track observations. [†]Additional controls in all models are S-track are main effects and S-Track interactions for: log(Filesize), log(Drafts), log(Cum Mail), 1[ID Sponsors=2], and 1[ID Sponsors>2].

Table A7: Publication Delay and Forward Citations

	Poisson Regressions DV = Total Cites					
	(1)	(2)	(3)	(4)	(5)	(6)
log(Suit-share)	-0.11 (0.44)		-0.11 (0.50)	1.48 (0.53)**		1.50 (0.54)**
log(Suit-share) * S-track	0.07 (0.45)		0.07 (0.50)	-0.88 (0.49)		-0.96 (0.51)
log(Total Days)		-0.10 (0.10)	-0.10 (0.09)		-0.06 (0.11)	-0.05 (0.11)
log(Total Days) * S-track		0.36 (0.12)**	0.36 (0.12)**		0.36 (0.12)**	0.35 (0.12)**
log(ID Mail)	0.35 (0.04)***	0.34 (0.05)***	0.34 (0.04)***	0.36 (0.05)***	0.35 (0.05)***	0.34 (0.05)***
log(Cites/Month) _{-i}	0.64 (0.07)***	0.64 (0.07)***	0.64 (0.07)***			
	Control Variables & Regression Statistics					
S-track x Cohort FE's	20 [0.00]	20 [0.00]	20 [0.00]	20 [0.00]	20 [0.00]	20 [0.00]
RFC Month Polynomial	4 [0.00]	4 [0.00]	4 [0.00]	4 [0.00]	4 [0.00]	4 [0.00]
Additional Controls	6 [0.00]	6 [0.00]	6 [0.00]	6 [0.00]	6 [0.00]	6 [0.00]
Tech-Class FEs	5 [0.00]	5 [0.00]	5 [0.00]			
WG FEs				Y	Y	Y
Obs. (Standards)	743	743	743	734	734	734
Obs. (Nonstandards)	403	403	403	393	393	393
Model d.o.f.	39	39	41	33	33	35
Pseudo-LogL x 10 ³	-8.30	-8.21	-8.21	-5.83	-5.82	-5.75

*5% significance; **1% significance; ***0.1% significance (all SEs clustered on Tech Class by S-Track).
Notes: The sample in models (1) through (6) is all RFCs from Draft Cohorts 1993 through 2003 published before 2007. For controls, the table indicates number of parameters and a p-value for joint significance. Additional controls are main effects and standards-track interactions for: log(Drafts), log(WG Mail), Sponsors, log(Filesize), Dot-edu, Dot-gov and Dot-org. See the text for additional details.