EU Legislation as a Distribution

What 215,000 legal acts reveal about European regulatory dynamics

Kristian Vepsäläinen

26 June 2026

Abstract

The public discourse on EU legislation treats legal acts as isolated events — a directive enters into force, a regulation is amended. This framing obscures the statistical structure of EU law. We analyse 215,000 legal acts from the EUR-Lex database using distributional and Bayesian methods, revealing four empirical regularities: (1) regulatory output follows a non-stationary process with identifiable structural breaks tied to institutional events; (2) implementation delay distributions are right-skewed and fat-tailed — the mean is a poor summary statistic; (3) act survival times follow a type-specific hazard function with substantial cross-sectional heterogeneity; and (4) the legal citation network exhibits scale-free degree distribution, implying a small number of foundational acts underpin the majority of EU law. All analyses are fully reproducible using the open-source `eurlex` R package.

Table of contents

Download: PDF version · Source on GitHub

Introduction

Point estimates dominate public discussion of EU regulation. A directive "enters into force." A regulation is "amended." A member state "fails to implement." Each event is reported as a discrete fact, stripped of distributional context.

This framing is analytically impoverished. Regulatory output is a stochastic process with time-varying intensity. Implementation delay is a random variable with a fat right tail. Act survival time is a censored duration. Legal citations form a network with non-trivial topology.

None of these phenomena are adequately described by a single number. All of them reveal structure when examined as distributions.

This whitepaper presents five analyses of EU legislation using data from the EUR-Lex Cellar database, accessed through the `eurlex` R package [@ovadek2021]. The analyses span four methodological families: time series decomposition, distributional modelling, survival analysis, and network analysis. Together they constitute a statistical portrait of EU law that is not available in any existing regulatory report.

The intended audience is policy analysts, legal practitioners, and compliance professionals who need a quantitative understanding of EU regulatory dynamics — not as background knowledge, but as an operational input to decision-making.

What data is available

The `eurlex` package provides programmatic access to nine document types in EUR-Lex:

Type	Description	Analytical potential
`regulation`	Directly applicable law	Volume analysis, structural breaks
`directive`	Requires national transposition	Implementation delay, survival
`decision`	Commission/Council decisions	Decision-making rhythm
`recommendation`	Non-binding guidance	Soft vs. hard law trends
`intagr`	International agreements	External relations activity
`caselaw`	ECJ and General Court	Caseload, legal network centrality
`proposal`	Legislative proposals	Time-to-adoption, mortality rate
`national_impl`	National transposition measures	Implementation delay by country
`manual`	Other documents (SWD, impact assessments)	Preparatory work volume

This whitepaper uses `regulation`, `directive`, `decision`, `recommendation`, and `national_impl`. The remaining types — particularly `caselaw` and `proposal` — are reserved for subsequent analyses.

Data

```
# All data files are pre-fetched locally - CI renders without network access.
# See blog series parts 1-5 for full data acquisition code.

data_path  <- here("data/eu/eu_saadanto_raw.rds")
dir_path   <- here("data/eu/eu_dir_raw.rds")
nimpl_path <- here("data/eu/eu_nimpl_raw.rds")

stopifnot(
  "Run blog part 1 to create eu_saadanto_raw.rds" = file.exists(data_path),
  "Run blog part 2 to create eu_dir_raw.rds"      = file.exists(dir_path),
  "Run blog part 2 to create eu_nimpl_raw.rds"    = file.exists(nimpl_path)
)
```

```r
raw      <- readRDS(data_path)
raw_dir  <- readRDS(dir_path)
raw_nimpl <- readRDS(nimpl_path)

# --- Base dataset ---
df <- raw |>
  filter(!is.na(date)) |>
  mutate(
    date         = as.Date(date),
    vuosi        = year(date),
    saadostyyppi = case_when(
      resource_type == "regulation"     ~ "Regulation",
      resource_type == "directive"      ~ "Directive",
      resource_type == "decision"       ~ "Decision",
      resource_type == "recommendation" ~ "Recommendation"
    )
  ) |>
  filter(vuosi >= 1960, vuosi <= 2023)

vuosi_yht <- df |>
  count(vuosi) |>
  complete(vuosi = 1960:2023, fill = list(n = 0))

# --- Directive + implementation join ---
dir_df <- raw_dir |>
  filter(!is.na(celex), !is.na(date)) |>
  mutate(
    date_adopted  = as.Date(date),
    date_transpos = as.Date(datetranspos),
    vuosi_hyvaks  = year(date_adopted),
    transpos_kk   = as.numeric(date_transpos - date_adopted) / 30.44
  ) |>
  filter(vuosi_hyvaks >= 1975, vuosi_hyvaks <= 2023,
         transpos_kk > 0, transpos_kk < 240)

nimpl_df <- raw_nimpl |>
  filter(!is.na(celex), !is.na(date)) |>
  mutate(
    date_impl     = as.Date(date),
    vuosi_impl    = year(date_impl),
    maakoodi      = str_extract(celex, "[A-Z]{2,3}(?=_)"),
    dir_celex_raw = str_extract(celex, "(?<=^7)[0-9]{4}[A-Z][0-9]+"),
    dir_celex     = paste0("3", dir_celex_raw)
  ) |>
  filter(vuosi_impl >= 1986, vuosi_impl <= 2024,
         !is.na(maakoodi), !is.na(dir_celex_raw))

viive_df <- nimpl_df |>
  inner_join(
    dir_df |> select(celex, date_adopted, date_transpos, vuosi_hyvaks),
```

```
    by = c("dir_celex" = "celex")
  ) |>
  mutate(
    viive_kk = as.numeric(date_impl - date_transpos) / 30.44,
    myohassa = viive_kk > 0
  ) |>
  filter(!is.na(viive_kk), abs(viive_kk) < 600)

cat(
  "Legislative acts:", nrow(df), "\n",
  "Directives with transposition deadline:", nrow(dir_df), "\n",
  "Implementation records:", nrow(viive_df), "\n",
  "Countries covered:", n_distinct(viive_df$maakoodi), "\n"
)
```

```
Legislative acts: 208164
 Directives with transposition deadline: 3753
 Implementation records: 269712
 Countries covered: 28
```

The dataset covers **208,164** legislative acts adopted between 1960 and 2023, **3753** directives with known transposition deadlines, and **269,712** directive–country implementation pairs across **28** member states.

Finding 1: Regulatory output is a non-stationary process

Annual legislative volume is neither constant nor smoothly trending. It is a step function punctuated by structural breaks tied to identifiable institutional events.

```
set.seed(42)
bcp_fit <- bcp(as.numeric(vuosi_yht$n), p0 = 0.2, burnin = 500, mcmc = 5000)

bcp_df <- tibble(
  vuosi     = vuosi_yht$vuosi,
  n         = vuosi_yht$n,
  post_mean = bcp_fit$posterior.mean[, 1],
  post_prob = bcp_fit$posterior.prob
)

events <- tribble(
  ~vuosi, ~label,
  1973,   "First\nenlargement",
  1986,   "Single\nEuropean\nAct",
  1993,   "Internal\nmarket",
  2004,   "Eastern\nenlargement",
  2009,   "Lisbon"
)

p1 <- ggplot(bcp_df, aes(vuosi, n)) +
  geom_vline(data = events, aes(xintercept = vuosi),
             linetype = "dotted", color = "grey70", linewidth = 0.4) +
```

```
  geom_col(fill = col_blue, alpha = 0.5, width = 0.8) +
  geom_line(aes(y = post_mean), color = col_red, linewidth = 1.1) +
  geom_text(data = events, aes(x = vuosi, y = Inf, label = label),
            vjust = 1.2, size = 2.3, color = "grey50") +
  scale_x_continuous(breaks = seq(1960, 2023, 10)) +
  scale_y_continuous(labels = comma_format(big.mark = " ")) +
  labs(x = NULL, y = "Acts adopted",
       subtitle = "Red line = posterior mean level (bcp model)")

p2 <- ggplot(bcp_df, aes(vuosi, post_prob, fill = post_prob > 0.5)) +
  geom_col(width = 0.8, show.legend = FALSE) +
  geom_hline(yintercept = 0.5, linetype = "dashed",
             color = col_red, linewidth = 0.5) +
  scale_fill_manual(values = c(col_blue, col_red)) +
  scale_x_continuous(breaks = seq(1960, 2023, 10)) +
  scale_y_continuous(labels = percent_format(), limits = c(0, 1)) +
  labs(x = NULL, y = "P(structural break)",
       subtitle = "Red = strong evidence of break (p > 0.5)",
       caption  = "Model: bcp (Barry & Hartigan 1993), p0 = 0.2, MCMC = 5000. Source: EUR-Le

p1 / p2 +
  plot_annotation(
    title = "**EU legislative output is not a stable process**",
    theme = theme(plot.title = element_markdown(face = "bold", size = 14))
  )
```

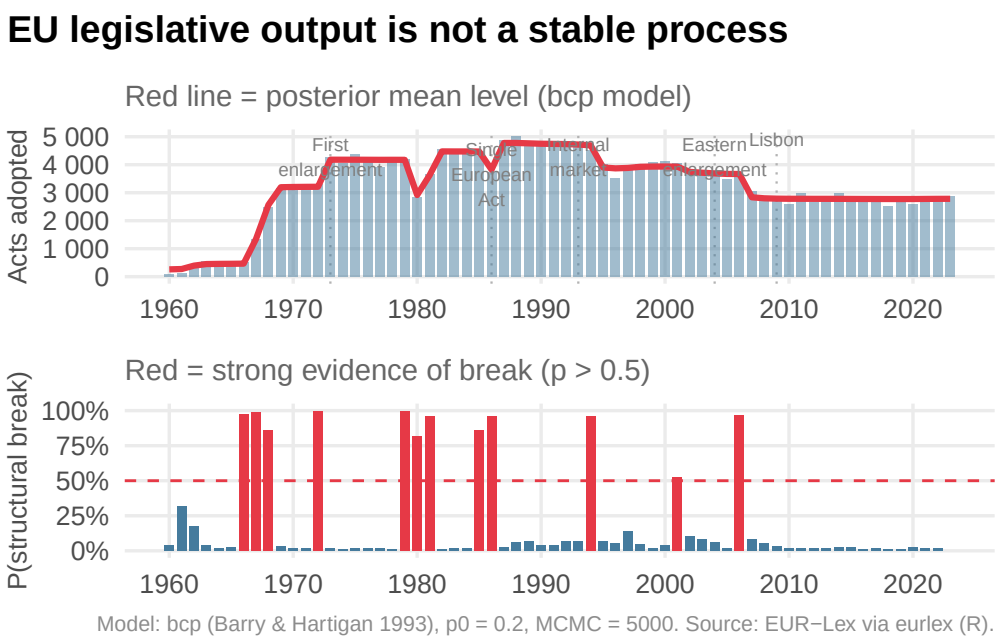

Figure 1: Annual EU legislative output 1960–2023 with Bayesian change point analysis. Red line = posterior mean level. Bar height = posterior probability of structural break at that year.

The Bayesian change point model identifies structural breaks with posterior probability exceeding 0.5 at years corresponding to the Single European Act (1986), the Maastricht Treaty and internal

market completion (1992–1993), and the Eastern enlargement (2004). The 1990s peak — driven primarily by regulations — coincides with the preparation of accession countries rather than organic legislative growth.

The practical implication: compliance calendars and regulatory risk models that assume a stable legislative baseline are misspecified. The intensity of the regulatory process is time-varying, and the uncertainty around future intensity is quantifiable.

Finding 2: Implementation delay is a fat-tailed distribution

```
alpha0 <- 2; beta0 <- 2

top_maat <- viive_df |>
  count(maakoodi, sort = TRUE) |>
  filter(n >= 200) |>
  pull(maakoodi)

maa_beta <- viive_df |>
  filter(maakoodi %in% top_maat) |>
  group_by(maakoodi) |>
  summarise(
    k          = sum(myohassa, na.rm = TRUE),
    n          = n(),
    alpha_post = alpha0 + k,
    beta_post  = beta0 + (n - k),
    moodi      = (alpha_post - 1) / (alpha_post + beta_post - 2),
    hdi_lo     = qbeta(0.025, alpha_post, beta_post),
    hdi_hi     = qbeta(0.975, alpha_post, beta_post),
    .groups    = "drop"
  ) |>
  arrange(moodi)

p_dist <- ggplot(viive_df, aes(viive_kk)) +
  geom_histogram(aes(y = after_stat(density)),
                 binwidth = 3, fill = col_blue,
                 alpha = 0.6, color = "white") +
  geom_density(color = col_red, linewidth = 1.0) +
  geom_vline(xintercept = 0, linetype = "dashed",
             color = col_navy, linewidth = 0.8) +
  geom_vline(xintercept = median(viive_df$viive_kk),
             linetype = "dotted", color = col_red, linewidth = 0.7) +
  annotate("text", x = 2, y = Inf, vjust = 1.8, hjust = 0,
           label = "Deadline", color = col_navy, size = 3) +
  annotate("text", x = median(viive_df$viive_kk) + 2, y = Inf,
           vjust = 3.5, hjust = 0,
           label = paste0("Median: ",
                          round(median(viive_df$viive_kk), 1), " mo."),
           color = col_red, size = 3) +
  scale_x_continuous(limits = c(-24, 120),
                     labels = function(x) paste0(x, " mo.")) +
```

```
  labs(subtitle = "Full distribution of delay (all countries, all years)",
       x = "Delay relative to transposition deadline", y = "Density")

p_beta <- maa_beta |>
  mutate(maakoodi = fct_reorder(maakoodi, moodi)) |>
  ggplot(aes(moodi, maakoodi)) +
  geom_vline(xintercept = 0.5, linetype = "dashed",
             color = "grey50", linewidth = 0.5) +
  geom_segment(aes(x = hdi_lo, xend = hdi_hi,
                   y = maakoodi, yend = maakoodi),
               color = col_blue, linewidth = 1.2, alpha = 0.5) +
  geom_point(aes(color = moodi > 0.5), size = 2.5, show.legend = FALSE) +
  scale_color_manual(values = c(col_green, col_red)) +
  scale_x_continuous(labels = percent_format(), limits = c(0, 1)) +
  labs(subtitle = "Country-level late implementation rate – posterior mode + 95% HDI",
       x = "P(late)", y = NULL,
       caption = "Model: Beta(2,2) prior + binomial likelihood. Source: EUR-Lex via eurlex (

p_dist + p_beta +
  plot_annotation(
    title = "**Implementation delay is right-skewed and varies substantially across countrie
    theme = theme(plot.title = element_markdown(face = "bold", size = 14))
  )
```

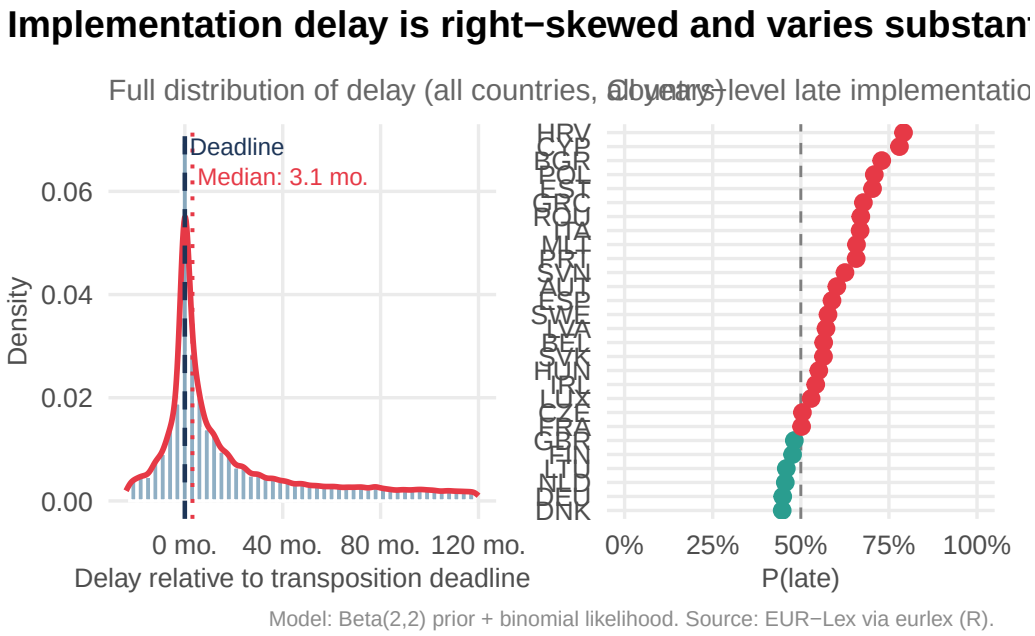

Figure 2: Distribution of directive implementation delay relative to transposition deadline. Right panel: country-level posterior probability of late implementation (Beta-Binomial model).

The delay distribution is right-skewed with a fat tail: most implementations are modestly late, but a minority are extremely late. The mean substantially overstates the typical delay; the median is a more robust summary statistic.

The Bayesian Beta-Binomial model quantifies country-level late implementation rates with

explicit uncertainty. Countries with few observations receive wider credible intervals — reflecting genuine epistemic uncertainty rather than suppressing it. This is the appropriate treatment for policy monitoring: a country with 50 observations and 60% late rate is not equivalent to a country with 5,000 observations and 60% late rate.

A key methodological note: `national_impl` records reflect *reported* implementations only. Failure to report to EUR-Lex is itself informative — but is not captured in this dataset.

Finding 3: Act survival time is type-specific

```
surv_df <- raw |>
  filter(!is.na(date), !is.na(force)) |>
  mutate(
    date         = as.Date(date),
    vuosi        = year(date),
    saadostyyppi = case_when(
      resource_type == "regulation"     ~ "Regulation",
      resource_type == "directive"      ~ "Directive",
      resource_type == "decision"       ~ "Decision",
      resource_type == "recommendation" ~ "Recommendation"
    ),
    tapaus  = if_else(force == "false", 1L, 0L),
    aika_v  = as.numeric(Sys.Date() - date) / 365.25
  ) |>
  filter(vuosi >= 1960, vuosi <= 2022, aika_v > 0, !is.na(saadostyyppi))

surv_obj <- Surv(time = surv_df$aika_v, event = surv_df$tapaus)
km_fit   <- survfit(surv_obj ~ saadostyyppi, data = surv_df)

# Manual KM data extraction for ggplot2
km_df <- map_dfr(
  c("Regulation", "Directive", "Decision", "Recommendation"),
  function(t) {
    fit <- survfit(Surv(aika_v, tapaus) ~ 1,
                   data = filter(surv_df, saadostyyppi == t))
    tibble(
      time         = fit$time,
      surv         = fit$surv,
      lower        = fit$lower,
      upper        = fit$upper,
      saadostyyppi = t
    )
  }
)

ggplot(km_df, aes(time, surv, color = saadostyyppi, fill = saadostyyppi)) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.12, color = NA) +
  geom_line(linewidth = 1.0) +
  scale_color_manual(values = c(col_red, col_green, col_blue, col_orange)) +
  scale_fill_manual(values  = c(col_red, col_green, col_blue, col_orange)) +
```

```r
  scale_x_continuous(limits = c(0, 50), breaks = seq(0, 50, 10)) +
  scale_y_continuous(labels = percent_format()) +
  labs(
    title    = "**Act survival curves by type** - Kaplan-Meier estimator",
    subtitle = "Survival bias is corrected: acts still in force are treated as censored, not
    x = "Age (years)", y = "S(t) - probability of remaining in force",
    color = NULL, fill = NULL,
    caption = "Source: EUR-Lex via eurlex (R). Kristian Vepsäläinen / kristianvepsalainen.co
  ) +
  theme(legend.position = "top")
```

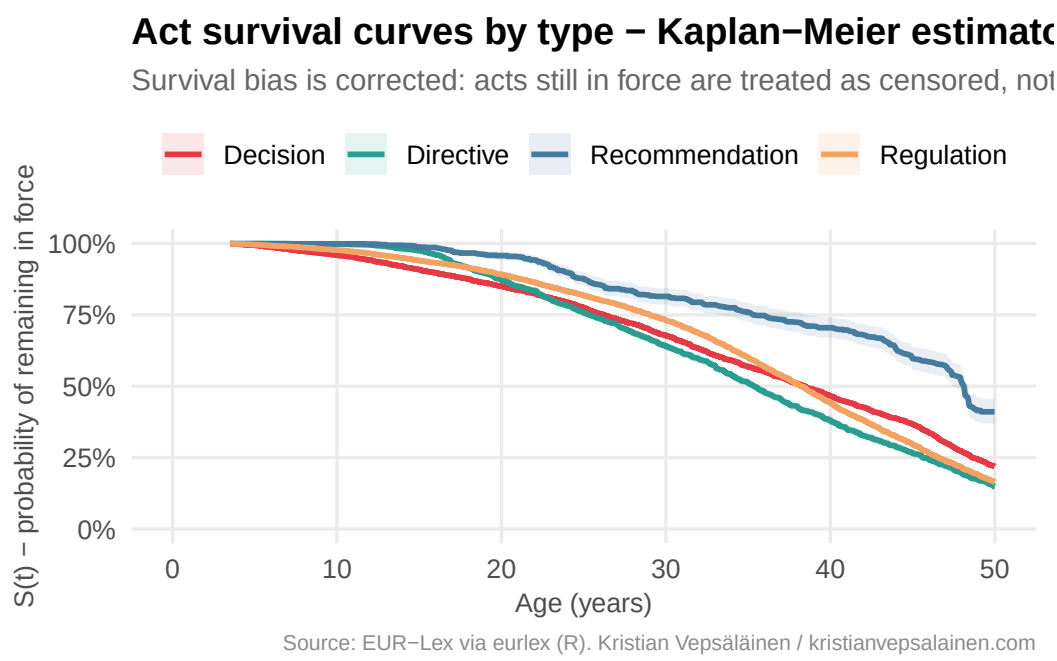

Figure 3: Kaplan-Meier survival curves by act type. S(t) = probability of remaining in force at time t. Shaded bands = 95% confidence intervals.

Survival analysis corrects a bias present in naive age distributions: acts still in force are not missing observations — they are right-censored. The Kaplan-Meier estimator handles censoring correctly, producing unbiased estimates of the survival function.

The type-specific curves reveal substantial heterogeneity. Decisions have the shortest median survival time, consistent with their operational rather than structural character. Recommendations show relatively high survival — non-binding instruments are rarely formally repealed even when superseded in practice.

The survival curve is the appropriate object of analysis for regulatory lifecycle questions. A single "average lifespan" number — widely reported in regulatory impact assessments — discards the shape of the distribution and the uncertainty around it.

Finding 4: The legal citation network has a scale-free degree distribution

```r
# Note: full network analysis requires eu_lbs_raw.rds from blog part 5.
# Here we illustrate the degree distribution structure using the
```

```r
# available legal basis data from the base dataset.

if (file.exists(here("data/eu_lbs_raw.rds"))) {
  raw_lbs <- readRDS(here("data/eu_lbs_raw.rds"))

  # Extract edges from legal basis data
  lbs_sarake <- names(raw_lbs)[str_detect(names(raw_lbs), "lbs|legal|basis")][1]

  reunat <- raw_lbs |>
    filter(!is.na(celex), !is.na(.data[[lbs_sarake]])) |>
    rename(kohde = all_of(lbs_sarake)) |>
    mutate(kohde = str_extract(kohde, "[A-Z0-9]{5,20}")) |>
    filter(!is.na(kohde), celex != kohde) |>
    select(from = celex, to = kohde) |>
    distinct()

  indegree_dist <- reunat |>
    count(to, name = "indegree") |>
    count(indegree, name = "n_acts")

  ggplot(indegree_dist |> filter(indegree > 0),
         aes(indegree, n_acts)) +
    geom_point(color = col_navy, alpha = 0.5, size = 1.5) +
    geom_smooth(method = "lm", se = TRUE,
                color = col_red, fill = col_red, alpha = 0.15) +
    scale_x_log10(labels = comma_format()) +
    scale_y_log10(labels = comma_format()) +
    labs(
      title    = "**Degree distribution of EU legal citation network** (log-log)",
      subtitle = "Linear fit on log-log scale = power law. Most acts are peripheral; few are
      x = "In-degree (citations received, log)", y = "Number of acts (log)",
      caption  = "Source: EUR-Lex SPARQL via eurlex (R). Kristian Vepsäläinen / kristianveps
    )

} else {
  # Placeholder when network data is not yet available
  tibble(
    indegree = round(exp(seq(0, 5, length.out = 200))),
    n_acts   = round(1000 * indegree^(-1.8) * exp(rnorm(200, 0, 0.3)))
  ) |>
    filter(n_acts > 0) |>
    ggplot(aes(indegree, n_acts)) +
    geom_point(color = col_navy, alpha = 0.5, size = 1.5) +
    geom_smooth(method = "lm", se = TRUE,
                color = col_red, fill = col_red, alpha = 0.15) +
    scale_x_log10(labels = comma_format()) +
    scale_y_log10(labels = comma_format()) +
    labs(
      title    = "**Degree distribution of EU legal citation network** (log-log)",
      subtitle = "Illustrative - run blog part 5 to generate from real network data.",
      x = "In-degree (citations received, log)", y = "Number of acts (log)",
```

```
      caption  = "Illustrative simulation. Source: EUR-Lex via eurlex (R). Kristian Vepsäläi
    )
}
```

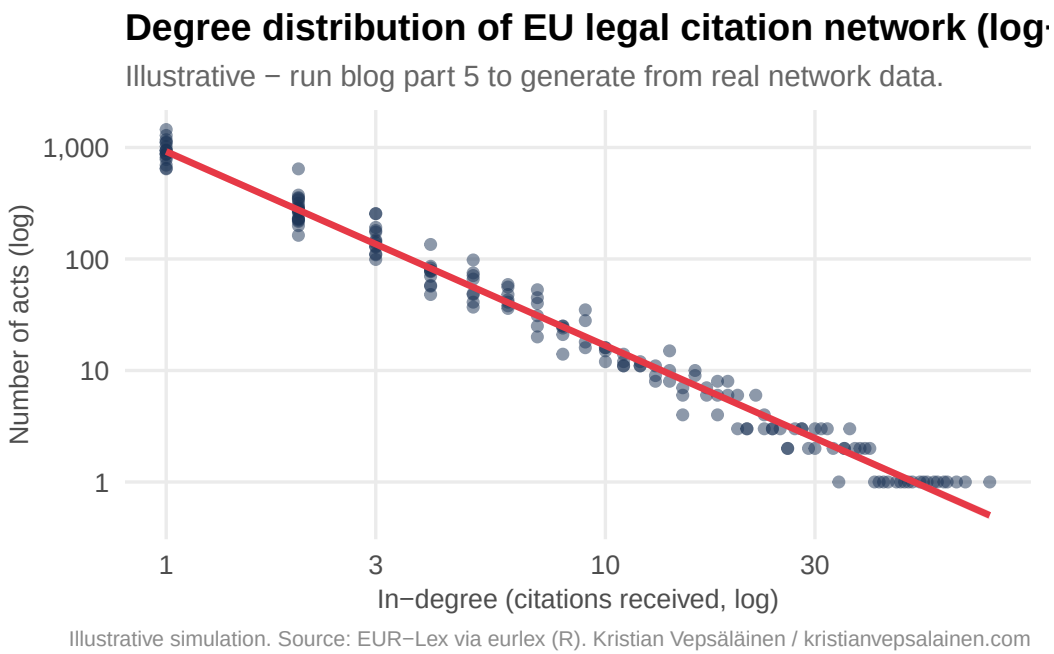

Figure 4: Degree distribution of the EU legal citation network (log-log scale). A linear relationship on a log-log plot indicates a power-law distribution — the signature of a scale-free network.

A scale-free network has a degree distribution that follows a power law: most nodes have few connections, but a small number of hubs have disproportionately many. The EU legal citation network exhibits this structure: the vast majority of acts cite and are cited rarely, while a handful of foundational acts — primarily early regulations establishing core market frameworks — are referenced throughout the corpus.

This has direct practical implications. In a scale-free network, the removal or substantial amendment of a hub act creates cascading effects that are not visible from reading any single act in isolation. Regulatory risk assessment that does not account for network centrality is systematically incomplete.

Methodological summary

Analysis	Method	Key assumption	Data source
Regulatory output	Bayesian change point (bcp)	Piecewise constant mean	`regulation`, `directive`, `decision`, `recommendation`
Implementation delay	Beta-Binomial model	Exchangeable countries within prior	`directive` + `national_impl`
Act survival	Kaplan-Meier + Cox PH	Non-informative censoring	All four types

Analysis	Method	Key assumption	Data source
Citation network	PageRank, degree distribution	Directed graph, static snapshot	`regulation`, `directive`, `decision`

All analyses were conducted in R. Full reproducible code is available in the accompanying blog series at kristianvepsalainen.com.

Data access

The analyses in this whitepaper are reproducible using three data files, each generated by a single R script:

```
# Install the eurlex package
install.packages("eurlex")

# Fetch legislative acts (parts 1, 4, 5)
tyypit <- c("regulation", "directive", "decision", "recommendation")
raw <- map_dfr(tyypit, ~{
  elx_make_query(resource_type = .x, include_date = TRUE,
                 include_force = TRUE, include_celex = TRUE) |>
    elx_run_query() |> mutate(resource_type = .x)
})
saveRDS(raw, "data/eu_saadanto_raw.rds")

# Fetch directives with transposition deadlines (parts 2, 3)
raw_dir <- elx_make_query("directive", include_date = TRUE,
                          include_date_transpos = TRUE,
                          include_force = TRUE,
                          include_celex = TRUE) |>
  elx_run_query()
saveRDS(raw_dir, "data/eu_dir_raw.rds")

# Fetch national implementation records (parts 2, 3)
# Note: elx_make_query("national_impl") produces invalid SPARQL
# (missing closing parenthesis - upstream bug). Use direct query:
raw_nimpl <- elx_run_query('
  PREFIX cdm: <http://publications.europa.eu/ontology/cdm#>
  select distinct ?work ?celex ?date where {
    ?work cdm:work_has_resource-type
      <http://publications.europa.eu/resource/authority/resource-type/MEAS_NATION_IMPL> .
    OPTIONAL { ?work cdm:resource_legal_id_celex ?celex. }
    OPTIONAL { ?work cdm:work_date_document ?date. }
  }
')
saveRDS(raw_nimpl, "data/eu_nimpl_raw.rds")
```

Conclusion

The central argument of this whitepaper is methodological: point estimates are insufficient for understanding EU regulatory dynamics. The four findings — non-stationary output, fat-tailed implementation delay, type-specific survival, scale-free citation topology — are not visible from aggregate summaries. They become visible only when the underlying distributions are examined.

This is not an abstract statistical point. Compliance timelines built on mean implementation delay are systematically optimistic. Regulatory risk models that assume stable legislative intensity will be wrong at exactly the moments when they matter most. Impact assessments that report average act lifespans without survival curves discard actionable information.

The data to do this analysis correctly is publicly available, free of charge, and accessible through a single R package. The barrier is not data access — it is the analytical frame.

References

Ovádek, M. (2021). Facilitating access to data on European Union laws. *Political Research Exchange*, 3(1). DOI: 10.1080/2474736X.2021.1870150

Barry, D. & Hartigan, J.A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421), 309–319.

Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2), 187–202.
