Modelling interactions in large longitudinal social networks: Mixed Additive REM

Prof. Ernst C. Wit

Università della Svizzera italiana, Switzerland

7 May, 2024

Remember: modelling hazard = modelling network

Let

$$\lambda_{sr}(t) = \underbrace{Y_{sr}(t)}_{\text{Is } (s,r) \text{ at risk?}} \times \underbrace{\lambda_0(t)}_{\text{baseline hazard}} \times \underbrace{e^{f_{sr}(x)}}_{\text{edge specific risk factors}}$$

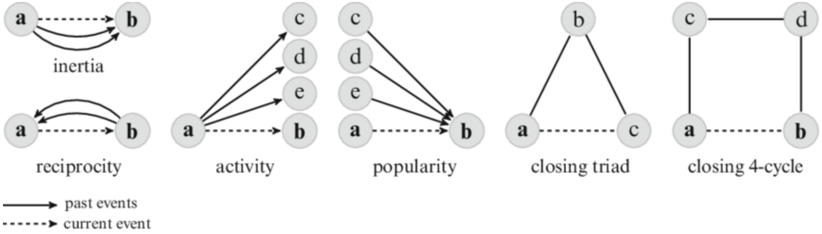
where

- $Y_{sr}(t)$: edge specific risk indicator (known)
- $\lambda_0(t)$: global risk determinants (unknown)
- $f_{sr}(x)$: edge-specific risk determinants (unknown)

QUESTIONS:

- Should we extend f_{sr} beyond $f_{sr}(x) = x_{sr}\beta$?
- If so, can we?

Endogenous edge-specific determinants of interactions



Endogenous effects: features depending on past interactions

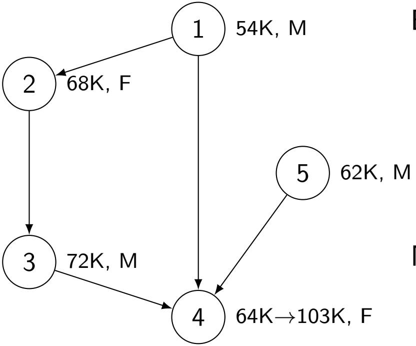
1. **monadic:** activity, popularity
2. **dyadic:** inertia, reciprocity
3. **triadic:** transitivity, cyclic closure, sender/receiver balance
4. **higher-order:** 4-cycle, k-star, ...

Endogenous effects relate to **emergence** and **virality**!

Exogenous edge-specific determinants of interactions

Exogenous effects: *inherent* features of sender and receiver

1. **monadic:** Depending on either sender or receiver only:
   - **known:** time-varying income, gender (measured covariates)
   - **unknown:** popularity, sociability (random effects)
2. **dyadic:** depending on sender and receiver (e.g. same gender)



Events:

1. (54K, M) $\longrightarrow$ (68K, F)
2. (68K, F) $\longrightarrow$ (72K, M)
3. Salary increase 4: 64K $\rightarrow$ 103K.
4. (54K, M) $\longrightarrow$ (103K, F)
5. (72K, M) $\longrightarrow$ (103K, F)
6. (62K, M) $\longrightarrow$ (103K, F)

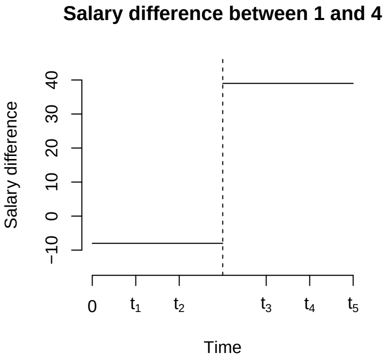
Note:

- Opposite sex attracts interaction
- Higher salary attracts interaction

Time-varying covariates

The covariate "salary difference" varies over time:

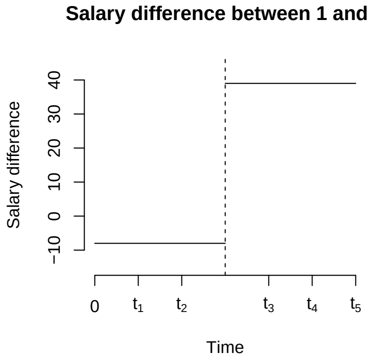
$$x_{sr}(t) = \text{Salary}_r(t) - \text{Salary}_s(t)$$

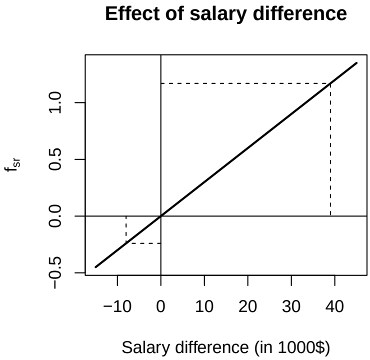


Time-varying covariates

The covariate "salary difference" varies over time:

$$x_{sr}(t) = \text{Salary}_r(t) - \text{Salary}_s(t)$$





... but its effect could be linear: $f_{sr}(x(t)) = x_{sr}(t)\beta$ (here: $\beta = 0.03$).

Drivers of species invasions: time-varying covariates

Most drivers change in time:

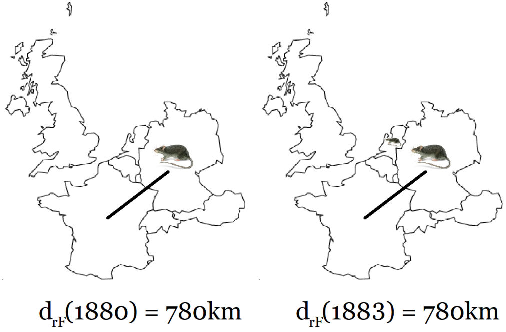
- $l_r(t)$: landuse in region r at time t.
- $d_{sr}(t)$: distance to region nearest to r invaded by s by time t.
- $tr_{sr}(t)$: annual trade between r and regions invaded by s by time t.
- $dt_{sr}(t)$: min temp diff between r and regions invaded by s by time t.
- $k_{sr}(t)$: presence of s at time t in colonial power to which r belongs.

$d_{rF}(1880) = 780\text{km}$

Drivers of species invasions: time-varying covariates

Most drivers change in time:

- $l_r(t)$: landuse in region r at time t.
- $d_{sr}(t)$: distance to region nearest to r invaded by s by time t.
- $tr_{sr}(t)$: annual trade between r and regions invaded by s by time t.
- $dt_{sr}(t)$: min temp diff between r and regions invaded by s by time t.
- $k_{sr}(t)$: presence of s at time t in colonial power to which r belongs.



Drivers of species invasions: time-varying covariates

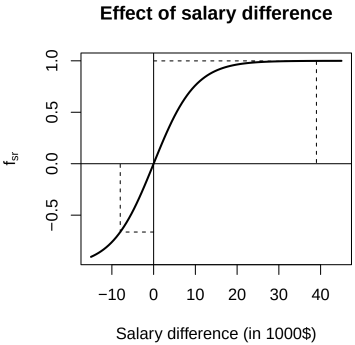
Most drivers change in time:

- $l_r(t)$: landuse in region r at time t.
- $d_{sr}(t)$: distance to region nearest to r invaded by s by time t.
- $tr_{sr}(t)$: annual trade between r and regions invaded by s by time t.
- $dt_{sr}(t)$: min temp diff between r and regions invaded by s by time t.
- $k_{sr}(t)$: presence of s at time t in colonial power to which r belongs.

$d_{rF}(1880) = 780\text{km}$ $d_{rF}(1883) = 780\text{km}$ $d_{rF}(1889) = 570\text{km}$

Non-linear effect of salary difference

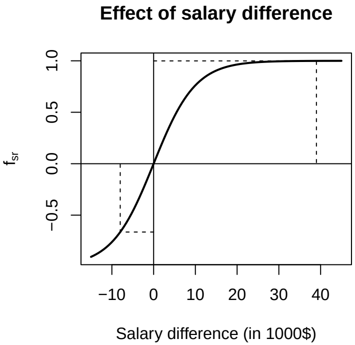
... but perhaps effect of salary difference is non-linear:

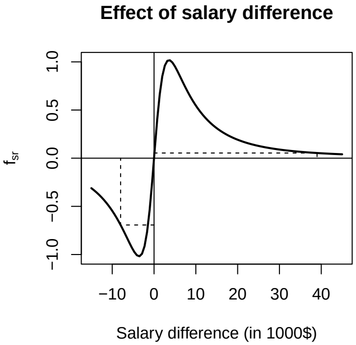


Maybe "salary difference" effect saturates

Non-linear effect of salary difference

... but perhaps effect of salary difference is non-linear:





Maybe "salary difference" effect saturates

Maybe it is more pronounced for small differences, but less for large ones

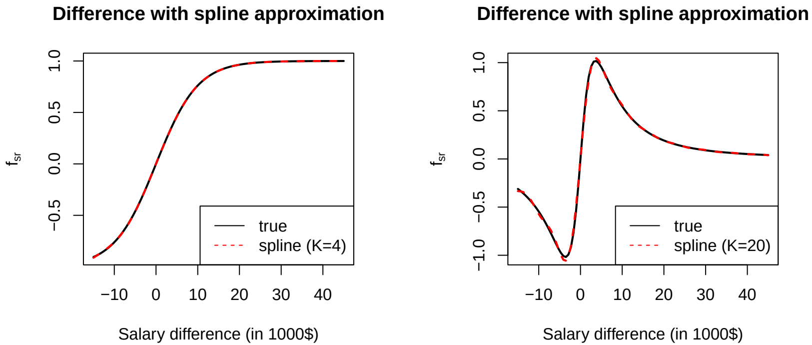
How to account for such forms without strong assumptions?

Splines: data-driven non-linear effects

Rather than using a particular form, we allow for a flexible function:

$$f_{sr}(x) = \sum_{k=1}^{K} \theta_k b_k(x),$$

where $\{b_1, \ldots, b_K\}$ is some convenient spline basis.

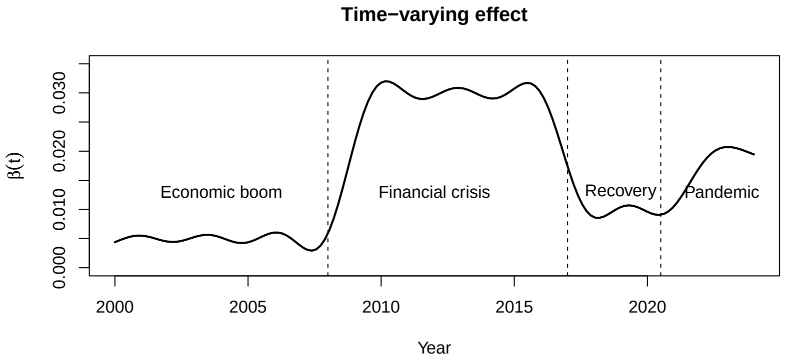


Time-varying effects

Alternatively, effect of salary difference might **change over time**:

$$f_{sr}(x(t)) = x_{sr}(t)\beta(t),$$

For example,



This can also be fitted with splines: $f_{sr}(x(t)) = \sum_k \theta_k b_k(t) x_{sr}(t).$

Why include random effects? Hierarchy Principle

Definition (Hierarchy principle)

Model with higher-order interactions **should also include all** lower-order interactions

Sociological models often include higher-order effects, e.g.,

- 2nd order interactions: repetition, reciprocity,
- 3rd order interactions: triadic closure,

BUT: No 1st order/node effects **violates** hierarchy principle.

Two types of 1st order effects:

- Endogenous:
  - number of interactions received
  - number of interactions initiated
- **Exogenous: unmeasured heterogeneity = random effects**

Mathematical form of f_{sr}: beyond linear!

Up till now, most social scientists considered *linear effects*.

Instead, we propose:

$$f_{sr}(x(t), z(t)) = \underbrace{\beta' x_{sr}^{(1)}(t)}_{\text{linear}} + \underbrace{\beta'(t) x_{sr}^{(2)}(t)}_{\text{time-varying}} + \underbrace{f(x_{sr}^{(3)}(t))}_{\text{non-linear}} + \underbrace{\gamma' z_{sr}(t)}_{\text{random}}$$

This is crucial:

- **Time-varying:** effects may change in long-term
- **Non-linear:**
  - **Optimum:** effects might have an optimum.
  - **Saturation:** effects might saturate
  - **Temporal:** effects may have temporal structure
- **Heterogeneity:** people do not react in same way

Event history model for time-to-invasion

Hazard for all species $s \in \mathcal{S}$ and regions $r \in \mathcal{C}$:

$$\lambda_{sr}(t) = \text{hazard of species } s \text{ invading region } r \text{ in year } t.$$

by means of

$$\lambda_{sr}(t) = Y_{sr}(t)\lambda_0(t)e^{x'_{sr}(t)\beta(t)+z'_{sr}(t)\gamma}$$

where

- $Y_{sr}(t)$: at risk indicator of invasion of region r by species s
- $\lambda_0(t)$: baseline hazard
- $x_{sr}(t)$: time-varying covariates
- $z_{sr}(t)$: random effect covariates
- $\gamma \sim N(0, \Sigma_\gamma)$: random effects

Estimation of Relational Event Model

Case-control Partial Likelihood:

Randomly sample 1 non-event (t_i, s_i^*, r_i^*) for each event (t_i, s_i, r_i).

$$(\hat{\beta}, \hat{\theta}, \widehat{\Sigma}_\gamma) = \arg\max \prod_{i=1}^{n} \frac{e^{\Delta x_i \beta + \Delta b_i \theta + \Delta z_i \gamma}}{1 + e^{\Delta x_i \beta + \Delta b_i \theta + \Delta z_i \gamma}}$$

subject to smoothness constraints $\theta^t S \theta \leq c$, where

- $\Delta b_i = (b_1(x_{s_i r_i}) - b_1(x_{s_i^* r_i^*}), \ldots, b_K(x_{s_i r_i}) - b_K(x_{s_i^* r_i^*}))$
- $\Delta x_i = x_{s_i r_i} - x_{s_i^* r_i^*}$ and $\Delta z_i = z_{s_i r_i} - z_{s_i^* r_i^*}$.
- S is a penalty matrix involving second derivatives.

This is equivalent with **additive mixed effect logistic regression**

Use function `gam` from R-package `mgcv`

How to fit non-linear effect using gam

Fit interactions as non-linear function of salary difference (`x`).

- Let `x.ev` and `x.nv` be $n \times 1$ vector of events & non-events.
- Let `ones` be $n \times 1$ vector of ones.
- Define

```
X = cbind(x.ev,x.nv)
I = cbind(ones,-ones)
```

- Fit the non-linear model via:

```
gam(ones~-1 + s(X, by=I), family = binomial)
```

This fits hazard function:

$$\lambda_{sr}(t) = \lambda_0(t) e^{f_{sr}(x(t))}$$

How to fit time-varying effect using gam

Fit interactions as linear function of `x` with time-varying $\beta(t)$:

- Let `x.ev` and `x.nv` be $n \times 1$ vector of events & non-events.
- Let `ones` be $n \times 1$ vector of ones.
- Let `tms` be $n \times 1$ vector of event times.
- Define
  ```
  T = cbind(tms,tms)
  X = cbind(x.ev,-x.nv)
  ```
- Fit the time-varying effect model via:
  ```
  gam(ones~-1 + s(T, by=X), family = binomial)
  ```

This fits hazard function:

$$\lambda_{sr}(t) = \lambda_0(t) e^{x_{sr}(t)\beta(t)}$$

How to fit random effects using gam

Fit interactions with **random sender effect**.

- Let `s.ev` be $n \times 1$ factor of event senders.
- Let `s.nv` be $n \times 1$ factor of non-events senders.
- Let ones be $n \times 1$ vector of ones.
- Define
```
S = cbind(s.ev,s.nv)
I = cbind(ones,-ones)
```
- Fit the random effect model via:
```
gam(ones∼-1+s(X,by=I,bs="re"),family=binomial)
```

This fits hazard function:

$$\lambda_{sr}(t) = \lambda_0(t)e^{\gamma_s}$$

Species invasions

Event history model for time-to-invasion

Hazard for all species $s \in \mathcal{S}$ and regions $r \in \mathcal{C}$:

$$\lambda_{sr}(t) = \text{hazard of species } s \text{ invading region } r \text{ in year } t.$$

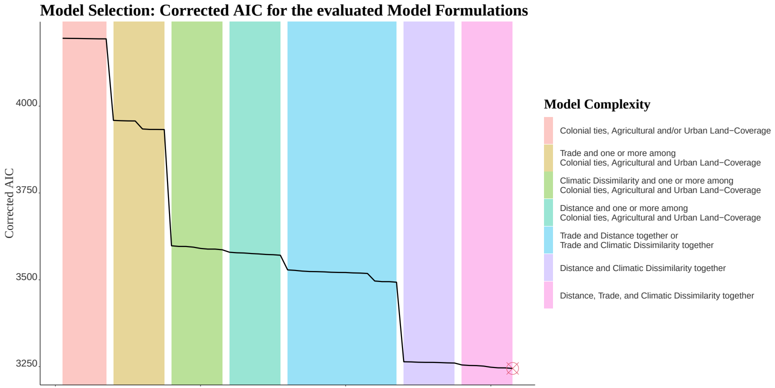
by means of

$$\lambda_{sr}(t) = Y_{sr}(t)\lambda_0(t)e^{x'_{sr}(t)\beta(t)+z'_{sr}(t)\gamma}$$

where

- $Y_{sr}(t)$: 0 if s is already present in r at time t
- $\lambda_0(t)$: baseline hazard
- $x_{sr}(t)$: climate, distance, trade, colonial ties, land-use
- $z_{sr}(t)$: species, region, species-interaction

Species invasions: model selection



Trade, climate & distance: most important factors in species dispersions.

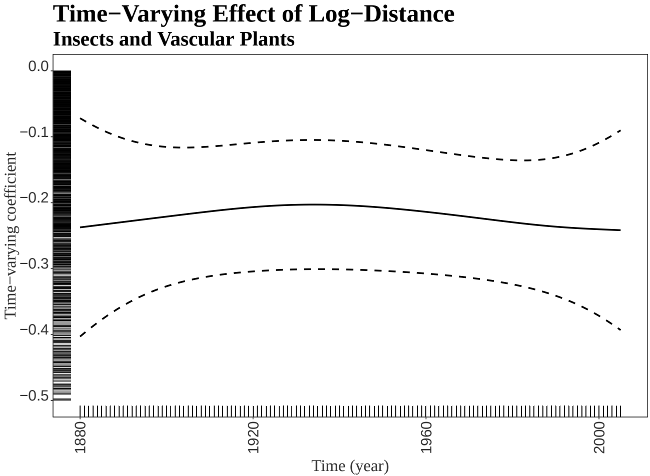
Results: fixed effects

	Birds	Plants	Insects	Mammals
Colonial ties	0.16	-0.09	0.31	0.13
Difference in temperature	-0.08	-0.04	-0.11	-0.07

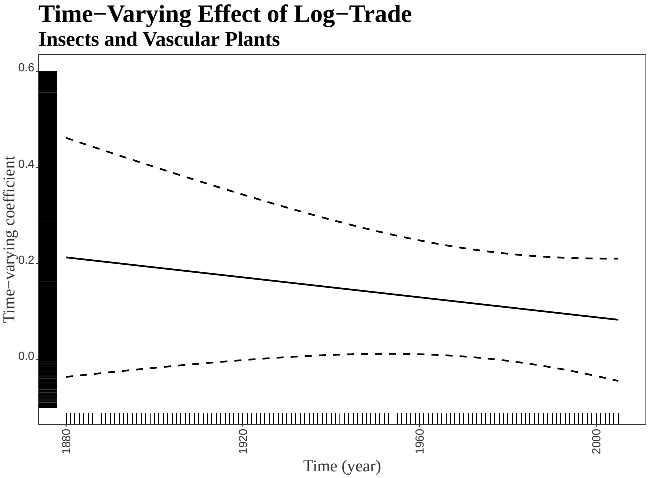
From this we can conclude:

- Colonial ties only has an impact in dispersion of plants.
- Species tend to invade countries with same climatic conditions.

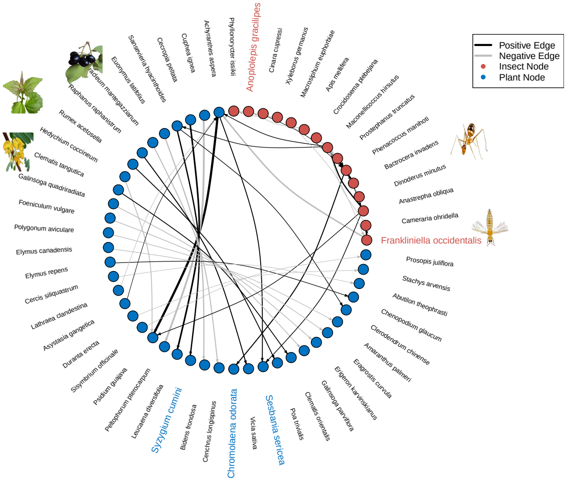
Results: distance reduces invasions



Results: trade is becoming less important

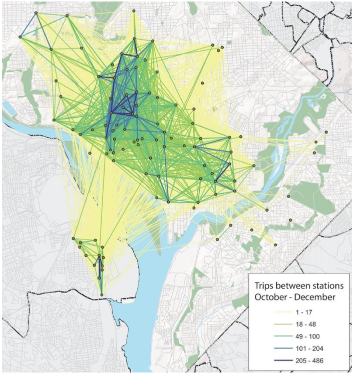


Results: Species have a tendency to coinvade



Bike sharing in Washington DC

Reminder: bike sharing network (100K-1M)



- **nodes:** 1300+ bike stations in Washington DC
- **edges:** 350K rides
- **time:** between 9-31 July, 2023.

We define **relational events**:

$$(d_k, a_k, t_k)$$

- d_k: departure station
- a_k: arrival station
- t_k: departure time
- k: $1, \ldots, 350,000$

Bike-sharing model with global covariates

We consider following **hazard** model:

$$\lambda_{sr}(t) = \lambda_0(t)\exp\{g_{\text{temp}}(x^{(\text{temp})}(t)) + g_{\text{prec}}(x^{(\text{prec})}(t)) + g_{\text{tod}}(x^{(\text{ToD})}(t))$$
$$+ x_s^{(\text{comp})}\beta + x_r^{(\text{comp})}\gamma$$
$$+ f_{\text{dist}}(x_{sr}^{(\text{dist})}) + f_{\text{rep}}(x_{sr}^{(\text{rep})}(t)) + f_{\text{rec}}(x_{sr}^{(\text{rec})}(t))\}.$$

where

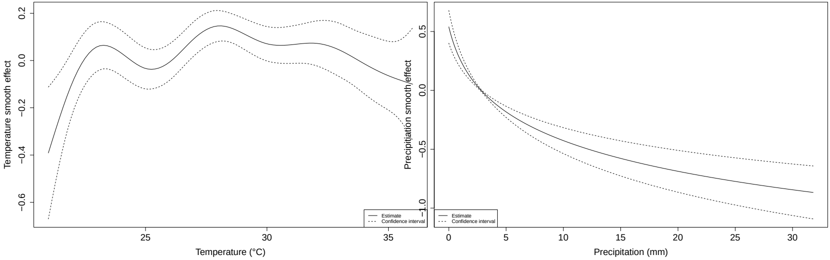
- **Global covariates:** temperature, precipitation, time-of-day
- **Linear effects:** sender/receiver competition
- **Edge-specific:** distance, repetition, reciprocity

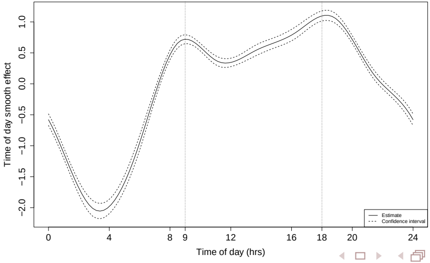
No sender and receiver competition (yet)

	Coef.	S.E.	p-value
β	-0.2145	0.0103	< 0.00001
γ	-0.1885	0.0101	< 0.00001

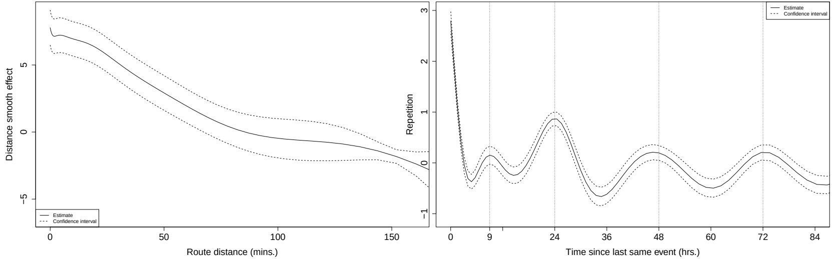
Negative competition: volume of bike shares is still too low compared to geographical concentration of bike stations.

Global effects: temperature, precipitation, time-of-day



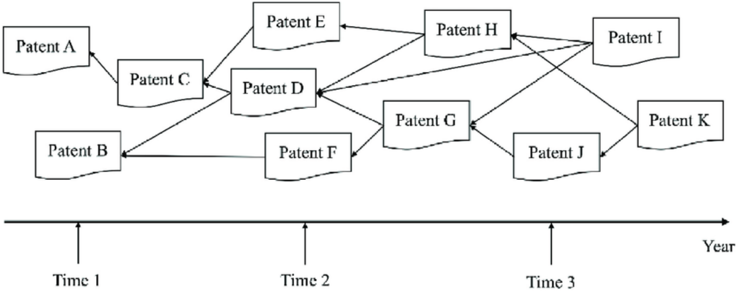


Edge-effects: distance and repetition



Dynamics of Innovation: patent citations

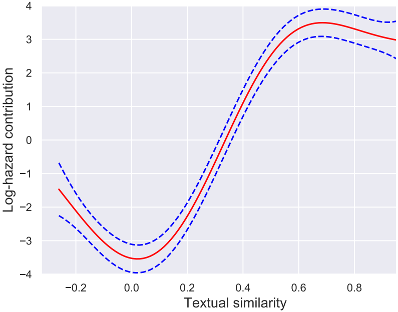
Reminder: patent citations (1M-100M)

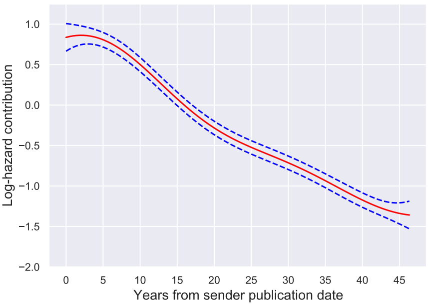


We study

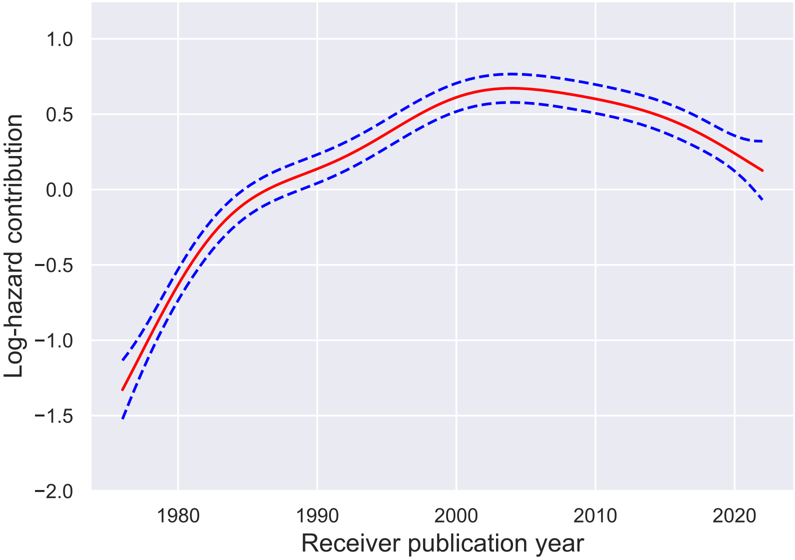
- 123M **patent citations**
- between 10M **patents**
- deposited some time between 1976 and 2023
- at US patent office

Drivers: similarity & time-lag

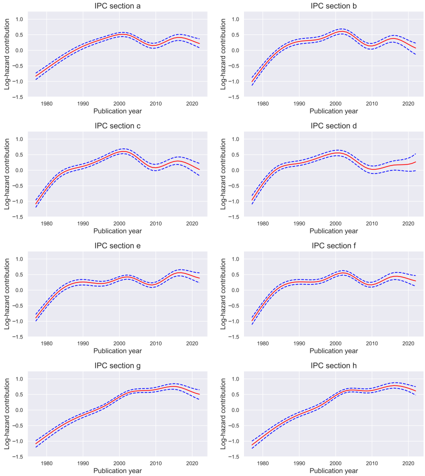




Innovation is declining since 2000



…but it depends on scientific field



A	Human necessities
B	Performing operations; transporting
C	Chemistry; metallurgy
D	Textiles; paper
E	Fixed constructions
F	Mechanical engineering; lighting; heating; weapons; blasting
G	Physics
H	Electricity

Take-home messages

This tutorial considered **effects in dynamic networks**:

1. **Covariates** are either:
   - Endogenous: depend on past of network
     often depend on time (reciprocity, triadic closure,...), or
   - Exogenous: depend on features of nodes
     can depend on time (e.g. income changes over time)
2. **Effects** of covariates can be:
   - Linear: $f_{sr}(x(t)) = x_{sr}(t)\beta$
   - Time-varying: $f_{sr}(x(t)) = x_{sr}(t)\beta(t)$
   - Non-linear: $f_{sr}(x(t))$ arbitrary function
3. Random effects account for unmodelled **heterogeneity**.
4. **Estimation** of parameters via sampled partial likelihood:
   - for each event (s_i, r_i, t_i) sample one non-event (s_i^*, r_i^*, t_i)
   - Fit **mixed additive logistic regression** with `mgcv::gam`