Modelling interactions in large longitudinal social networks: Simulation

Prof. Ernst C. Wit

Università della Svizzera italiana, Switzerland

7 May, 2024

Social networks

Social networks are **often** represented as a (un)directed graph

An edge (s, r) expresses a symmetric/asymmetric relation:

- s is friend of r
- s provides assistance to r
- s loans money to r

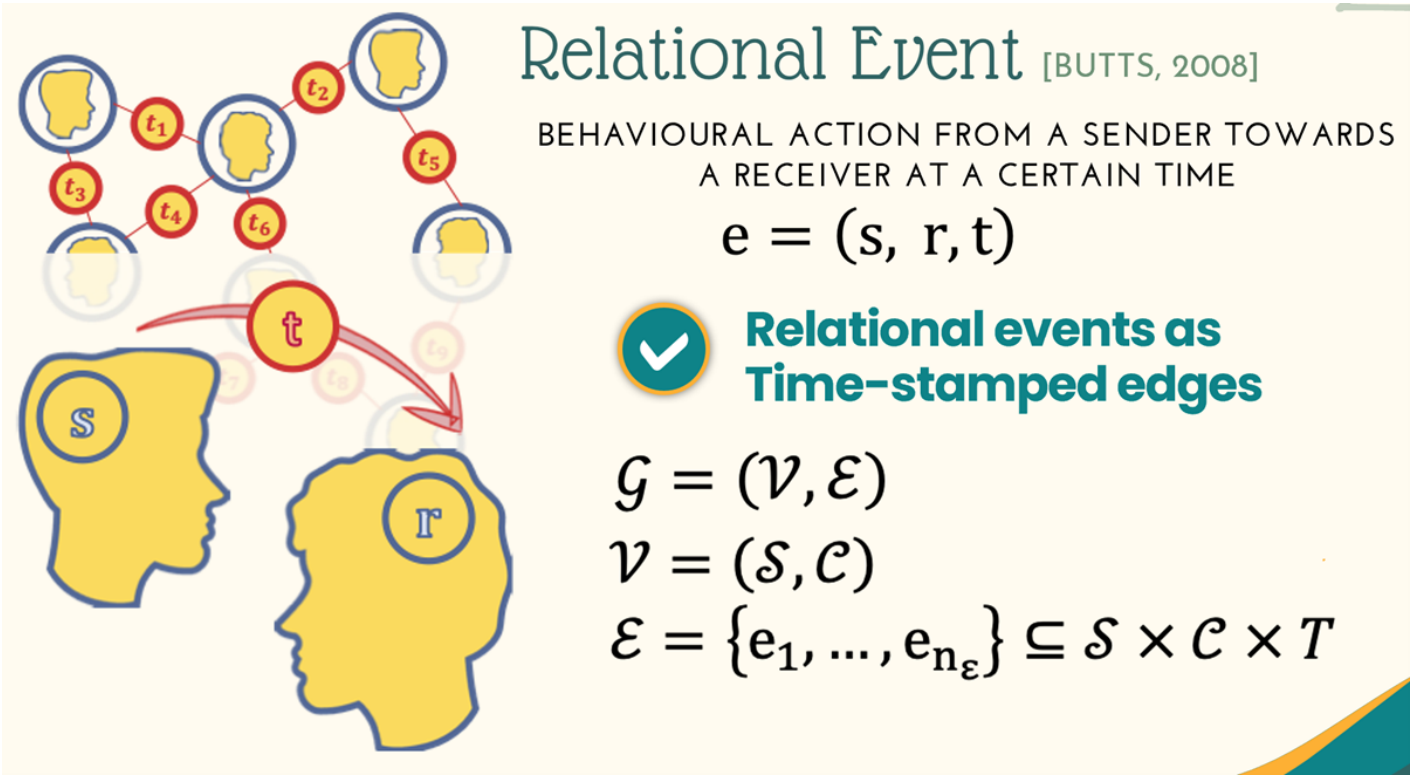
BUT: this ignores **temporal structure** of relationships.

What is a "longitudinal social network"?

Longitudinal social network

A social network where edges have a temporal component:

- **relational state:** edges appear and persist in time (friendship)
- **relational event:** edges are "instantaneous" (communication)



Example: alien species invasions (10K-100K)

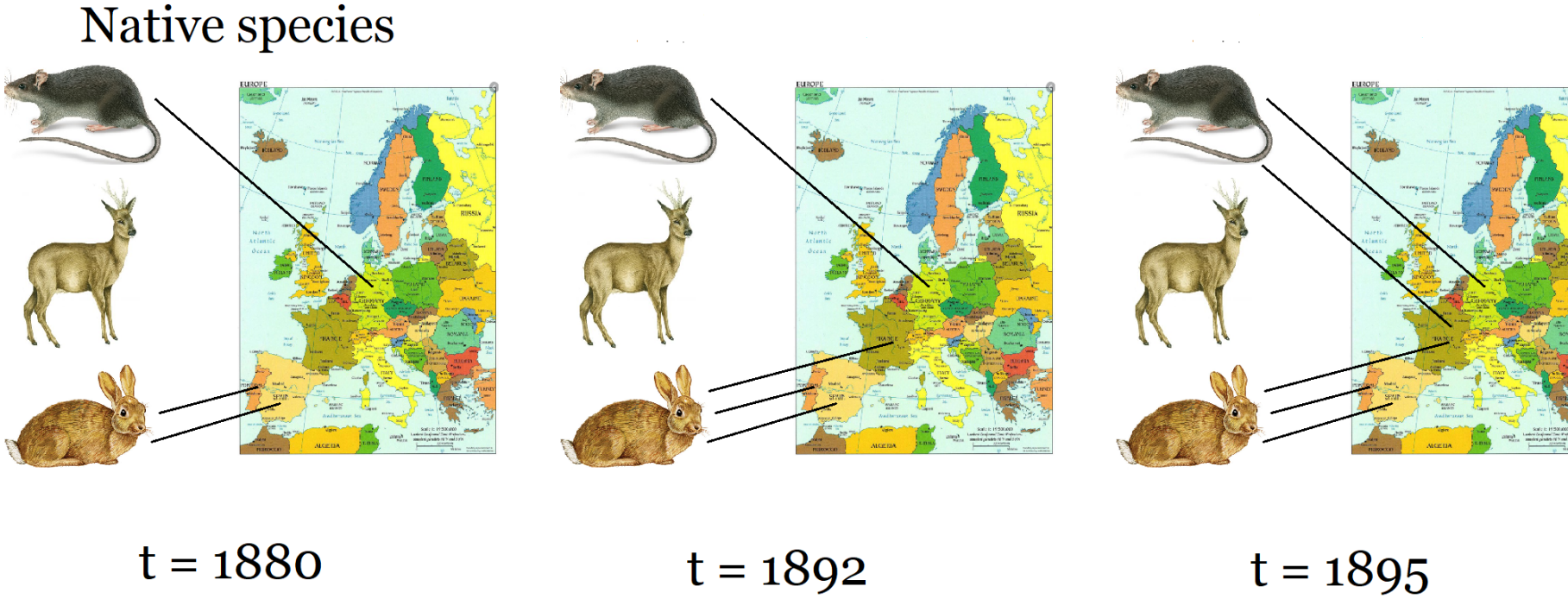
Data source: Alien Species First Records (Seebens et al., 2018)

H	I	J	K	L	M	N	O	P
OrigName	LifeForm	Region	PresentStatus	FirstRecor	FirstRecor	DataQualit	Source	
Acanthophora muscoides Linnaeus, 175	Algae	Turkey		1986	1986		Cinar et al. (2005)	
Acanthophora nayadiformis	Algae	Cyprus	alien	1997	1997	NEW_Befo	DAISIE	
Acanthophora nayadiformis (Delile) Pap	Algae	Turkey		1970	1970		Cinar et al. (2005)	
Acanthophora spicifera	Algae	Hawaiian Islands		1952	1952		Carlton & Eldrege (2009)	
Acetabularia calyculus	Algae	Israel	established	1943	1943	NEW_Befo	DAISIE	
Acetabularia calyculus	Algae	Spain	established	1957	1957	NEW_Befo	DAISIE	
Achnanthes pseudogroenlandica	Algae	Bulgaria		1984	1984		aquaNIS	
Achnanthes pseudogroenlandica	Algae	Romania		1984	1984		aquaNIS	
Achnanthes pseudogroenlandica	Algae	Ukraine		1984	1984	NEW_Befo	DAISIE	
Acrochaetium catenulatum	Algae	Netherlands		1967	1967		aquaNIS	
Acrochaetium kylinii	Algae	Turkey		2007	2000 - 2009	NEW_rand	aquaNIS	
Acrochaetium leptonema	Algae	Bulgaria		2006	2000 - 2009	NEW_rand	aquaNIS	
Acrochaetium leptonema	Algae	Turkey		2007	2000 - 2009	NEW_rand	aquaNIS	

Effectively giving information for

1. for each **species** (inside 1 of 16 life forms)
2. for each **"region"** (of 275 regions)
3. the **first moment** that the species is recorded there.

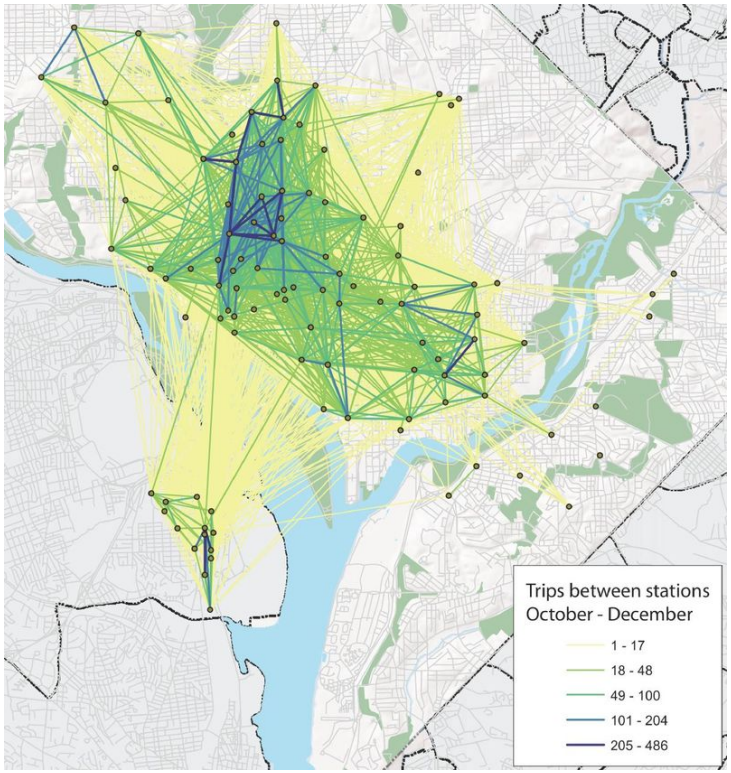
Dynamic two-mode species-region network



In a dynamic two-mode species-region network:

- **species** and **regions** are node-sets;
- The edges in network are **time-stamped invasions**;
- At time 0 ($t = 1880$) there are native species.

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



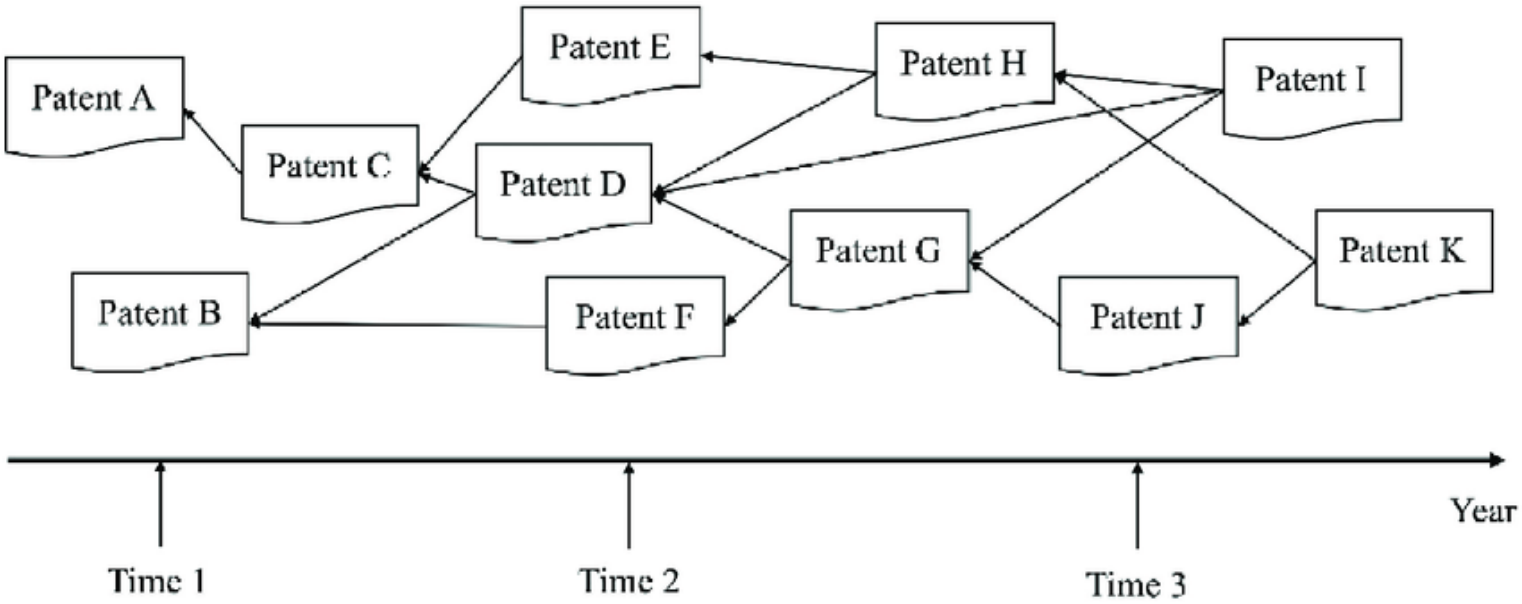
- **nodes:** 1300+ bike stations in Washington DC
- **edges:** 350K rides
- **time:** between 9-31 July, 2023.

We define **relational events**:

$$(d_k, a_k, t_k)$$

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

Example: patent citations (1M-100M)



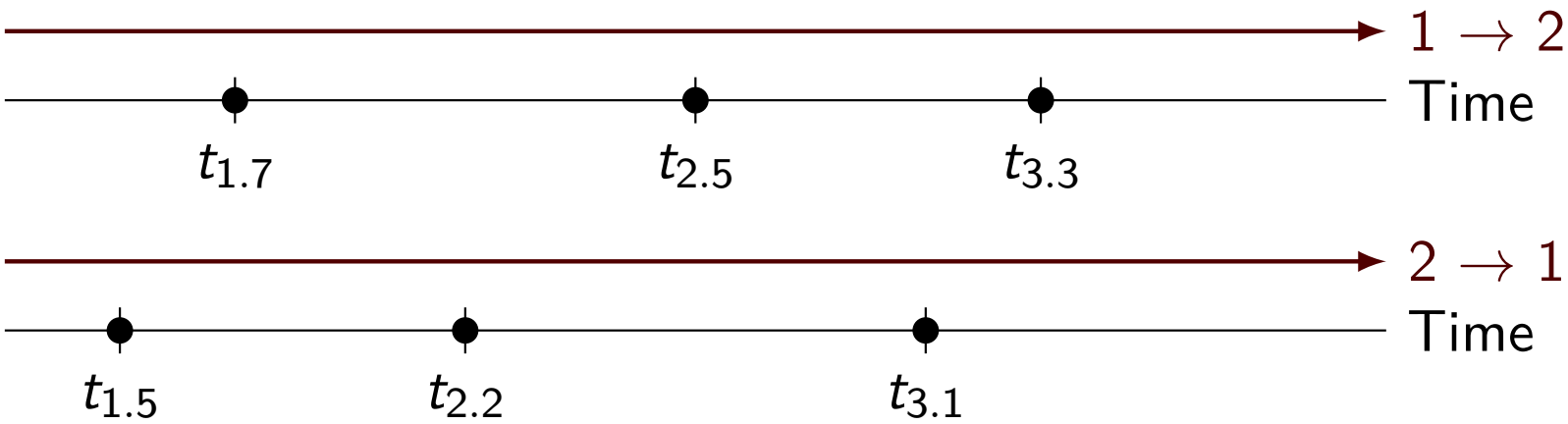
We study

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

Relational event data

A **relational event process** involves following components:

- Set of **actors** $V = \{1, 2, ..., p\}$.
- A relational event can be defined by a triple: $e = (s, r, t)$
  - $s \in V$: Sender of interaction event e
  - $r \in V$: Receiver of interaction event e
  - $t \in \mathbb{R}^+$: Time of interaction event

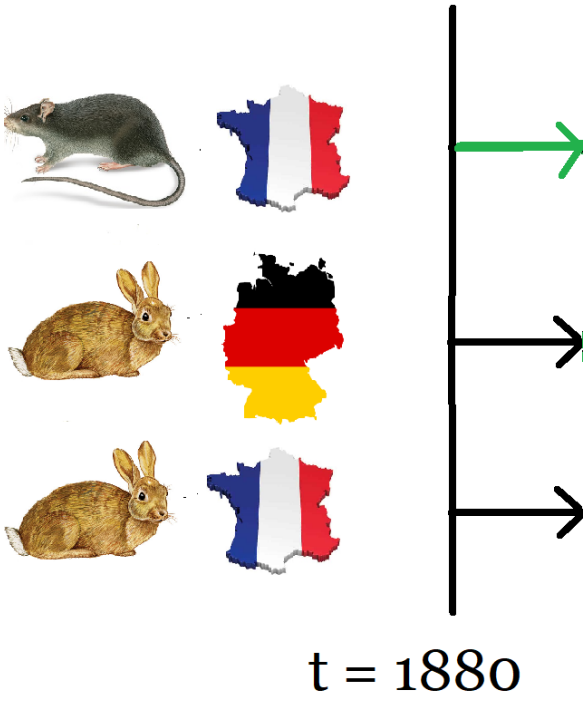


- Collection of time-stamped events $E = \{e_1, e_2, \ldots e_N\}$.

Dynamic two-mode network for species invasions

- Let $\mathcal{S}_L = \{\text{rattus, cuniculus}\}$ be all mammals.
- Let $\mathcal{C} = \{\text{Germany, France}\}$ be all regions.
- Let *rattus* be native to *Germany*.

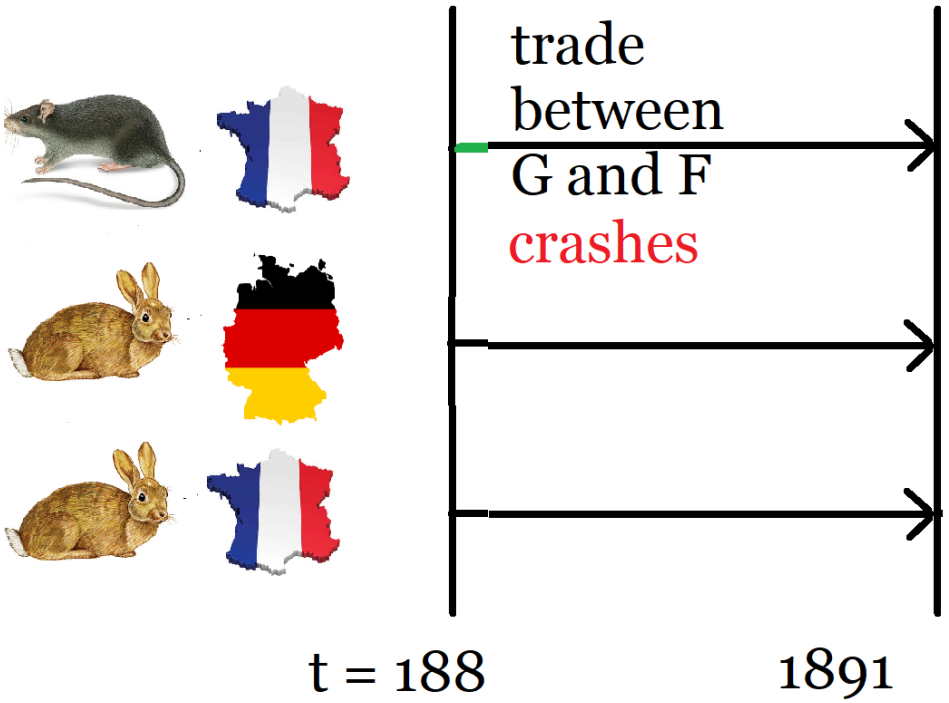
A "race" between T_{rF}, T_{cG} and T_{cF} (T_{rG} already arrived!):



Dynamic two-mode network for species invasions

- Let $\mathcal{S}_L = \{\text{rattus, cuniculus}\}$ be all mammals.
- Let $\mathcal{C} = \{\text{Germany, France}\}$ be all regions.
- Let *rattus* be native to *Germany*.

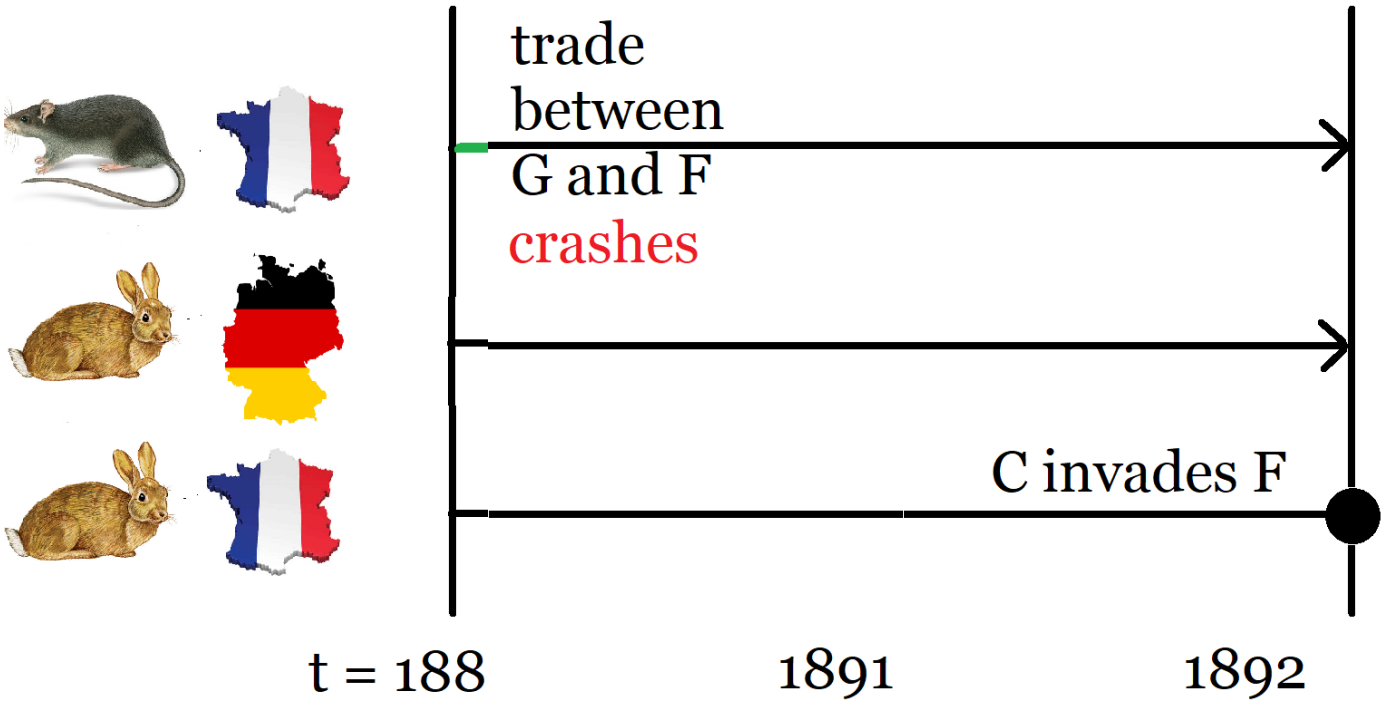
A "race" between T_{rF}, T_{cG} and T_{cF} (T_{rG} already arrived!):



Dynamic two-mode network for species invasions

- Let $\mathcal{S}_L = \{\text{rattus, cuniculus}\}$ be all mammals.
- Let $\mathcal{C} = \{\text{Germany, France}\}$ be all regions.
- Let *rattus* be native to *Germany*.

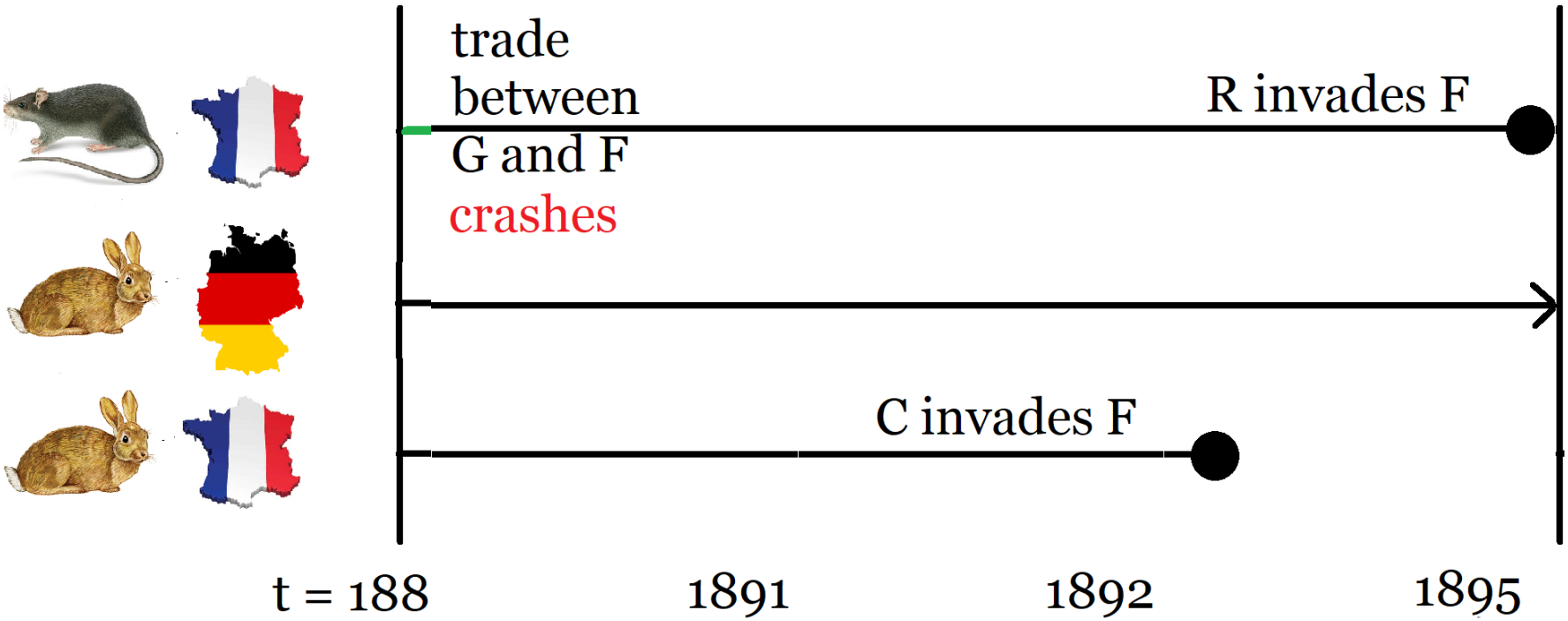
A "race" between T_{rF}, T_{cG} and T_{cF} (T_{rG} already arrived!):



Dynamic two-mode network for species invasions

- Let $\mathcal{S}_L = \{\text{rattus, cuniculus}\}$ be all mammals.
- Let $\mathcal{C} = \{\text{Germany, France}\}$ be all regions.
- Let *rattus* be native to *Germany*.

A "race" between T_{rF}, T_{cG} and T_{cF} (T_{rG} already arrived!):



Modelling longitudinal networks

Let $\{N_{sr}(t)\}_{srt}$ be **longitudinal network**:

- $N_{sr}(t) =$ number of interactions from s to r until time t.
- $N_{sr}(0) = 0$

Mathematically, it is always possible to decompose:

$$N_{sr}(t) = \underbrace{\Lambda_{sr}(t)}_{\text{structural part}} + \underbrace{M_{sr}(t)}_{\text{noise part}}$$

If it exists, then the **hazard**

$$\lambda_{sr}(t) = \frac{d\Lambda_{sr}}{dt}(t)$$

is *instantaneous "probability"* that (s, r) occurs at time t.

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
  e.g. a past patent *cannot* cite a future patent.
- $\lambda_0(t)$: global risk determinants
  e.g. very few bike rides when it rains at time t.
- $f_{sr}(x)$: edge-specific risk determinants
  e.g. if r is a popular person, then (s, r) more likely.

Simulating a relational event network

Let's consider 5 individuals applying for a job:

- Each has "work experience"-score WE given by

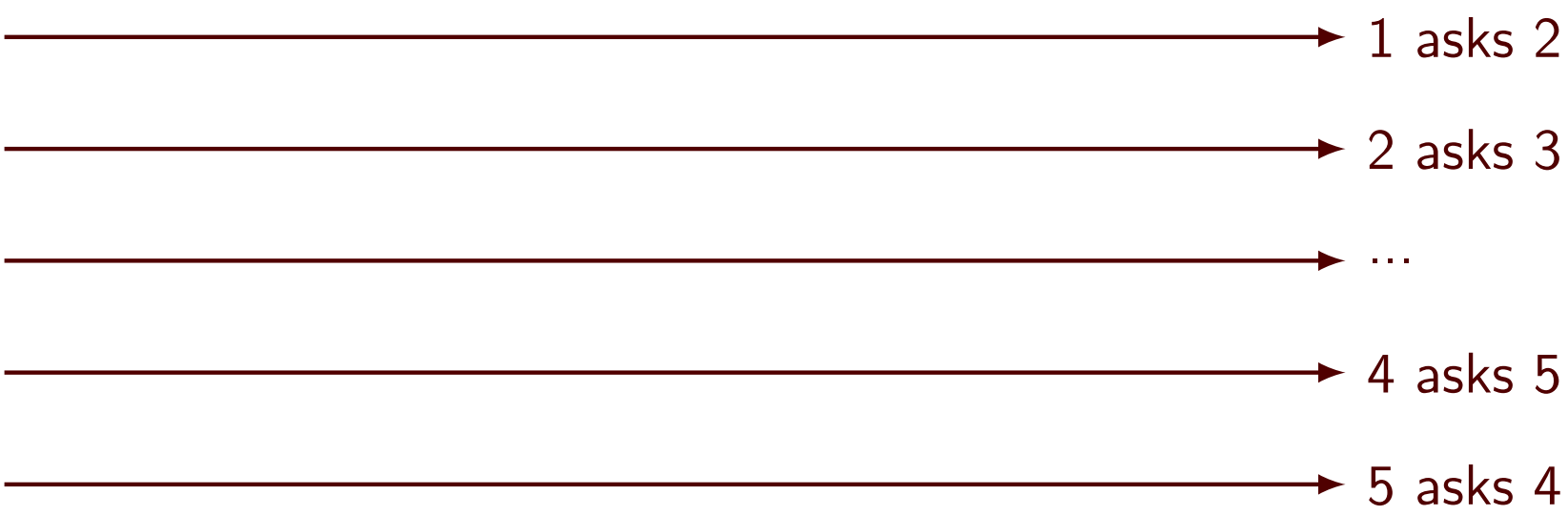
$$0,\ 3,\ 5,\ 9,\ 11 \quad \text{years}$$

- They need to get **two references letters** from each other.
- Person s contacts person r for a reference letter with rate,

$$\lambda_{sr}(t) = Y_{sr}(t) e^{-1.5|WE_s - WE_r|}.$$

- Value -1.5 identifies
  - **relevance of WE:** it is not zero
  - **direction of WE:** larger WE difference $\Rightarrow$ fewer requests
- Y_{sr} has value
  - 1 until person s has reached out to 2 people,
  - after which it returns to 0.

Distribution of event times



There are $5 \times 4 = 20$ requests "waiting" to be asked. Let

$$\Delta T_{sr} = \text{waiting time until } s \text{ asks } r \text{ for a letter}$$

The distribution of ΔT_{sr} is given as

$$\Delta_{sr} \sim \text{Exp}(e^{-1.5|WE_s - WE_r|})$$

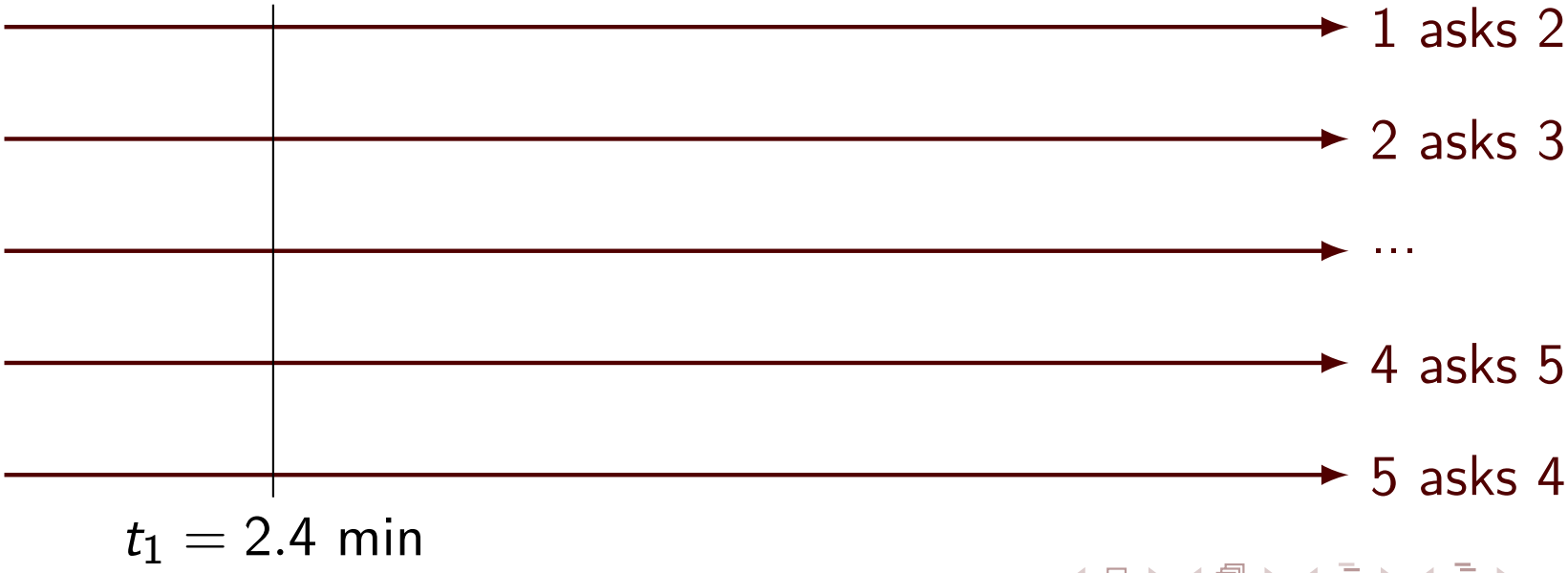
When will first request be sent out?

The first request will happen at time T_1, where

$$T_1 = \min_{sr} \Delta T_{sr}$$

Using property that minimum of exponentials is exponential,

$$T_1 \sim \text{Exp}\left(\sum_{s,r} e^{-1.5|WE_s - WE_r|} \approx 0.23 \right)$$



Which pair of individuals make up first request?

We now know that first event happened at t_1. But which one?

$$E_1 = \text{event that happened at time } t_1.$$

Using property that which exponential is minimum is multinomial,

$$E_1 \sim \text{Multinomial}\left(\frac{e^{-1.5|WE_1 - WE_2|}}{\sum_{s,r} e^{-1.5|WE_s - WE_r|}}, \ldots, \frac{e^{-1.5|WE_5 - WE_4|}}{\sum_{s,r} e^{-1.5|WE_s - WE_r|}}\right)$$

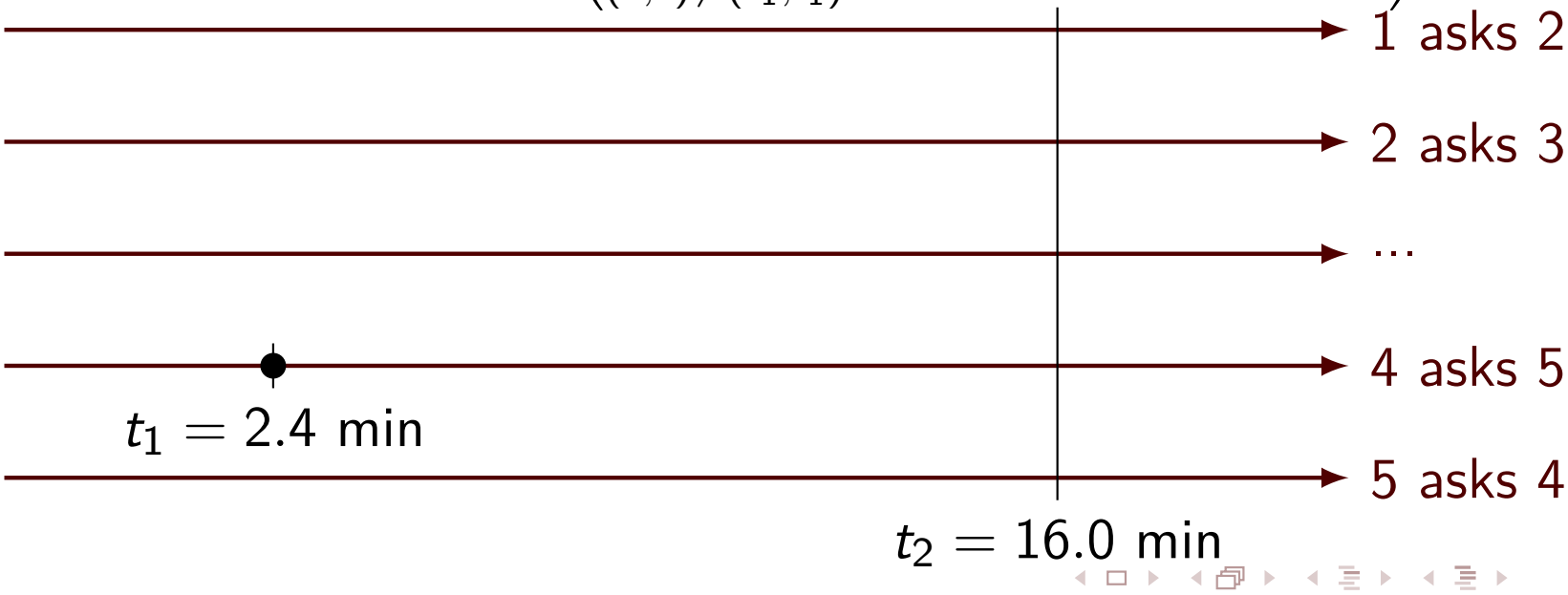
When will second request be sent out?

The second request will happen at time T_2, where

$$T_2 = \min_{sr \neq s_1 r_1} \Delta T_{sr}$$

Using memoryless property of an exponential,

$$T_2 \sim T_1 + \text{Exp}\left(\sum_{(s,r)\neq(s_1,r_1)} e^{-1.5|WE_s - WE_r|} \approx 0.18 \right)$$



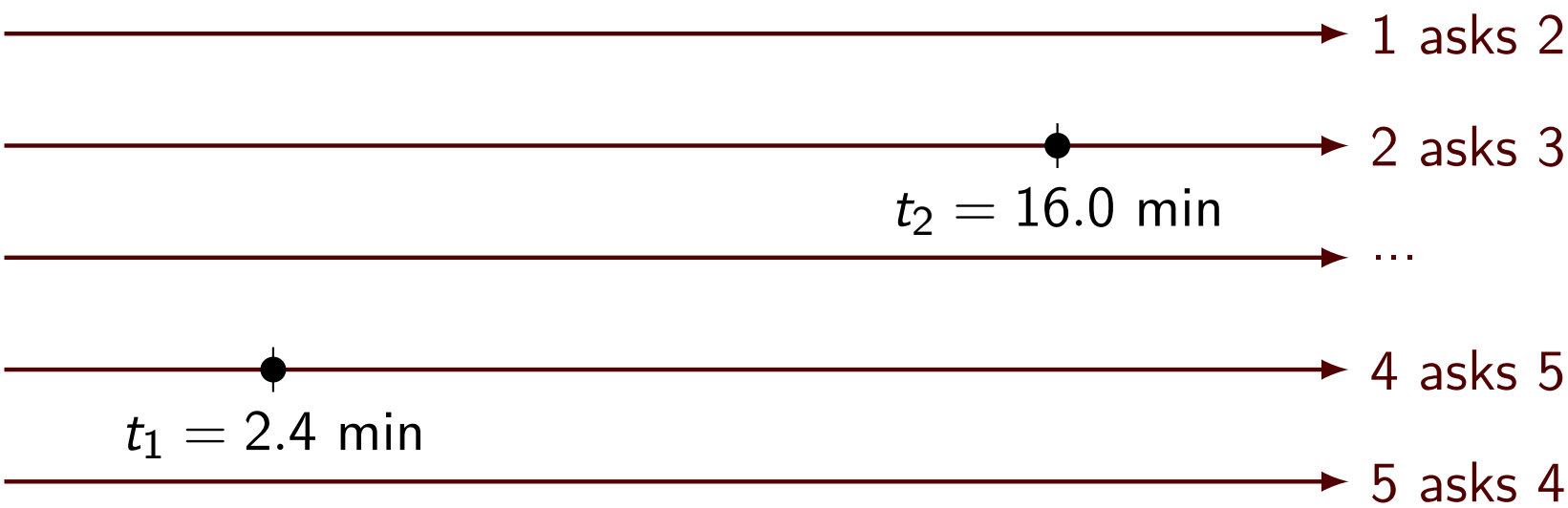
Which pair of individuals make up second request?

We know second event happened at $t_2 = 16.0$ min. Which one?

$$E_2 = \text{event that happened at time } t_2.$$

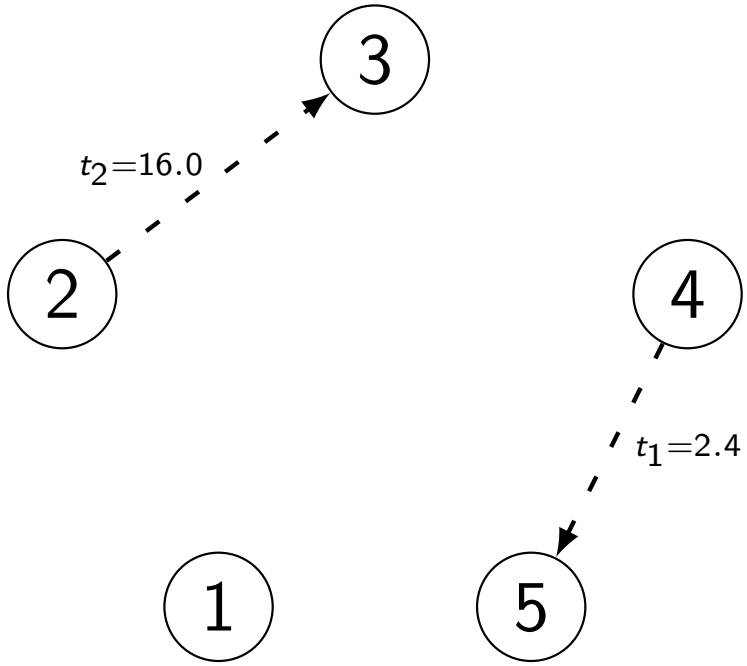
Using property that which exponential is minimum is multinomial,

$$E_2 \sim \text{Multinomial}\left(\frac{e^{-1.5|WE_1 - WE_2|}}{0.18}, \ldots, \frac{e^{-1.5|WE_5 - WE_4|}}{0.18}\right)$$



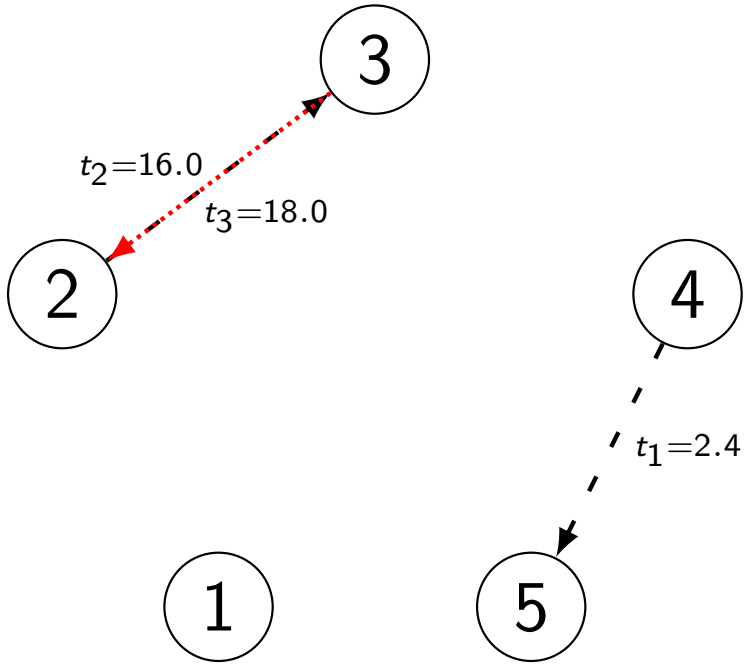
Repeat... to obtain a dynamic network

Repeating above procedure another 8 times, results in:



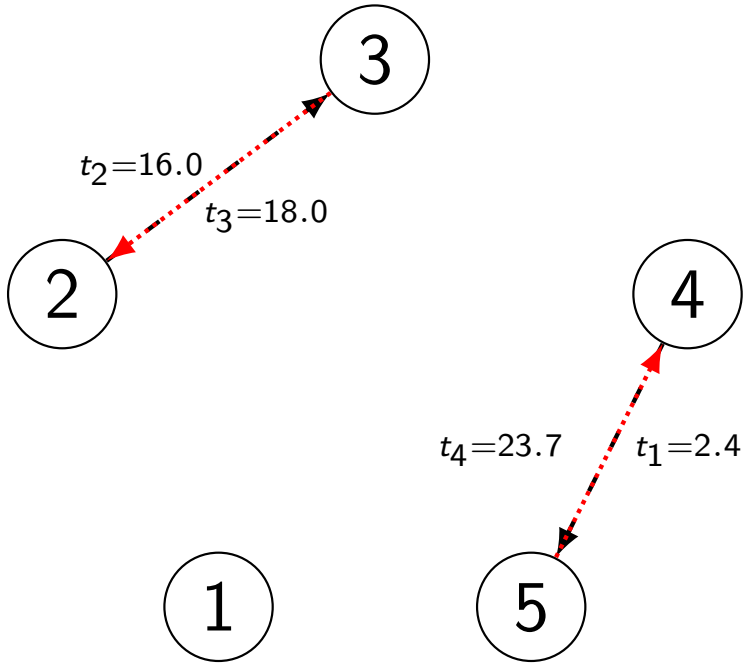
Repeat... to obtain a dynamic network

Repeating above procedure another 8 times, results in:



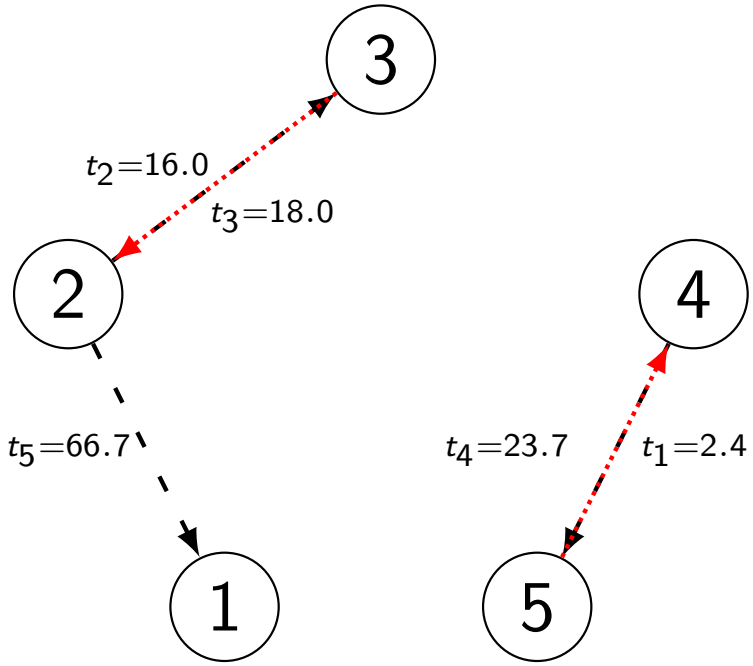
Repeat... to obtain a dynamic network

Repeating above procedure another 8 times, results in:



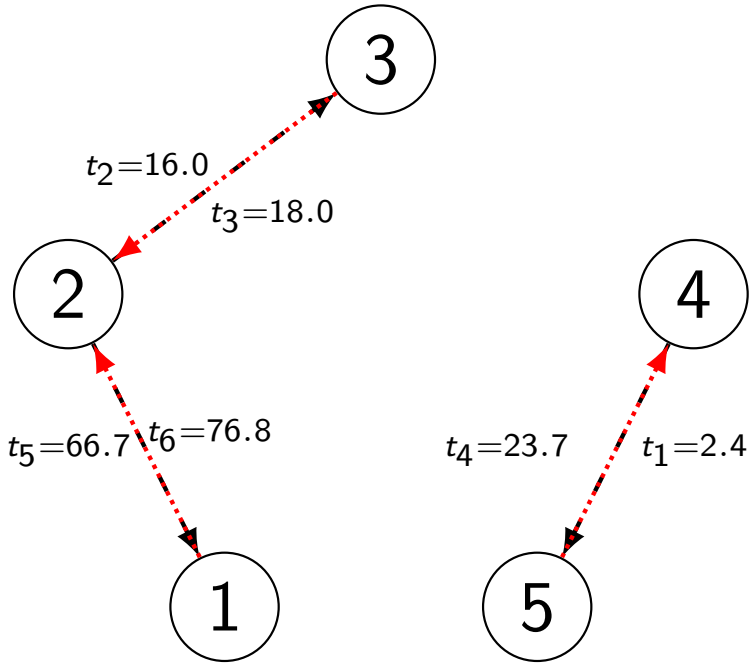
Repeat... to obtain a dynamic network

Repeating above procedure another 8 times, results in:



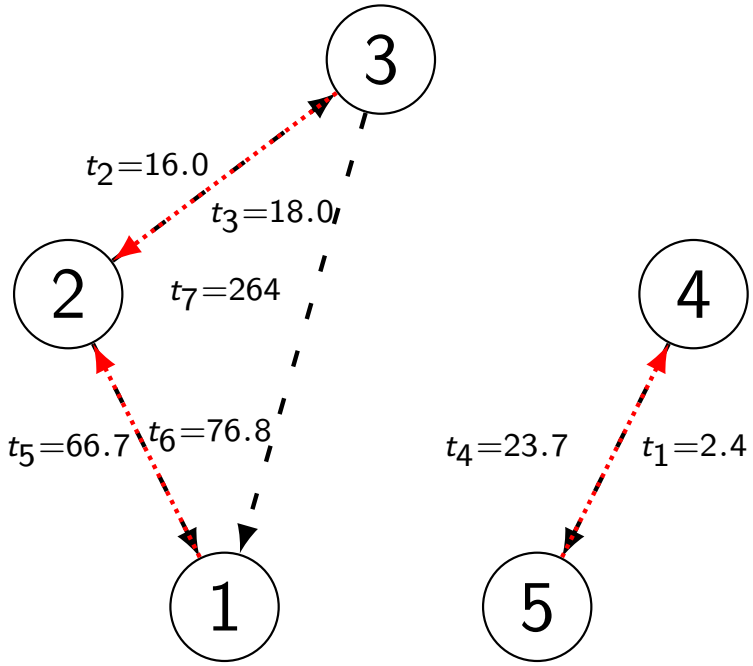
Repeat... to obtain a dynamic network

Repeating above procedure another 8 times, results in:



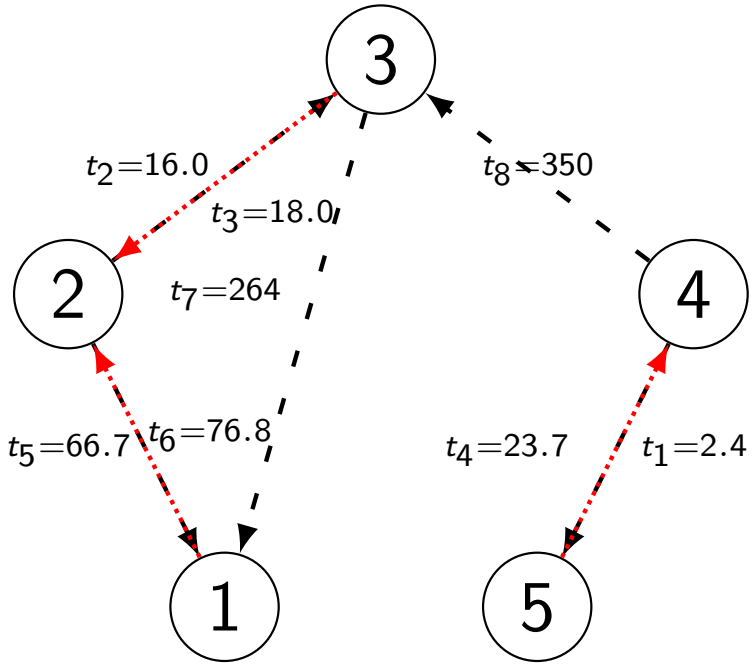
Repeat... to obtain a dynamic network

Repeating above procedure another 8 times, results in:



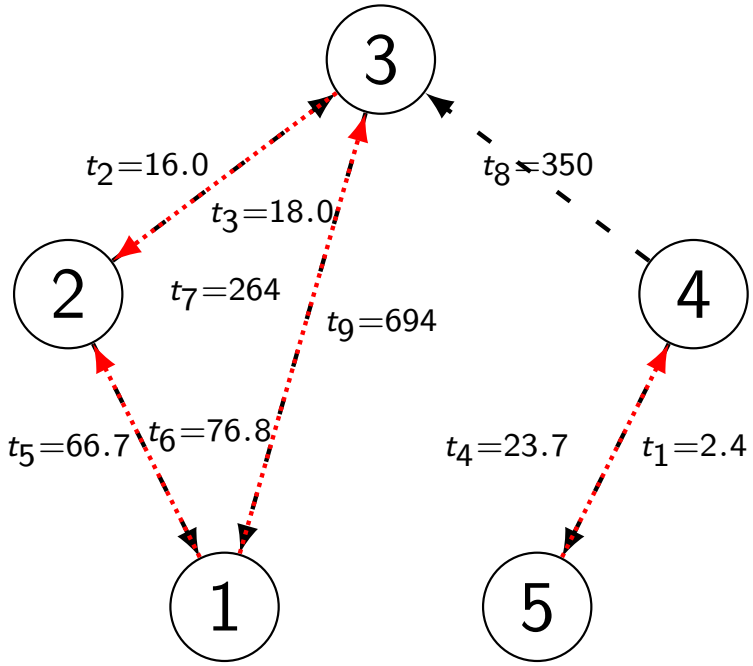
Repeat... to obtain a dynamic network

Repeating above procedure another 8 times, results in:



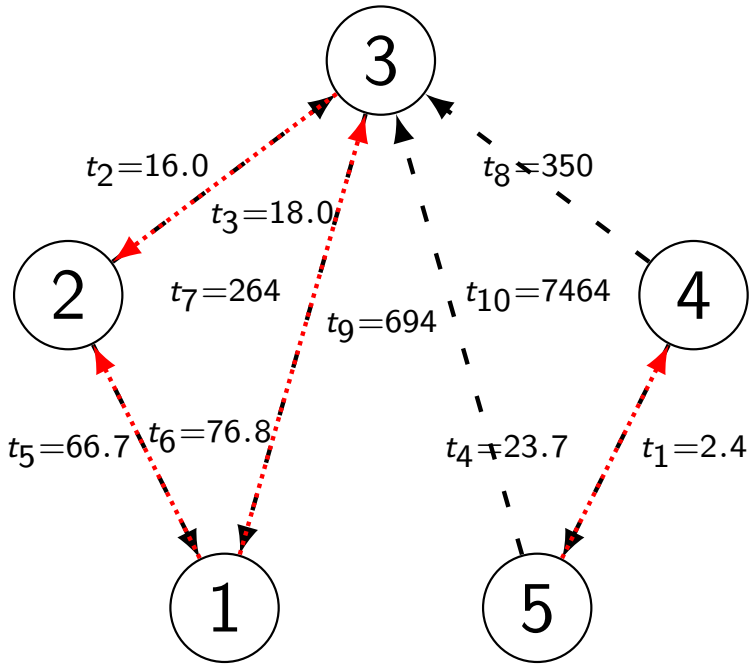
Repeat... to obtain a dynamic network

Repeating above procedure another 8 times, results in:



Repeat... to obtain a dynamic network

Repeating above procedure another 8 times, results in:



From Data to Answers

Let's assume that some gave us this dynamic network:

	time	solicitor	receiver
1	2.4	4	5
2	16.0	2	3
3	18.0	3	2
4	23.7	5	4
5	66.7	2	1
6	76.8	1	2
7	264.4	3	1
8	349.8	4	3
9	694.1	1	3
10	7463.6	5	3

Could we determine influence of WE on this dynamic?

Maximum sampled partial likelihood

There are a number of estimation paradigms:

- **MLE/Bayesian:** intractable even for small networks.
- **Partial likelihood:** denominator of PL is still too large.

Maximum sampled partial likelihood

There are a number of estimation paradigms:

- **MLE/Bayesian:** intractable even for small networks.
- **Partial likelihood:** denominator of PL is still too large.

Case-control Partial Likelihood:

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

$$\hat{\beta} = \arg\max \prod_{i=1}^{10} \frac{e^{\beta |WE_{s_i} - WE_{r_i}|}}{e^{\beta |WE_{s_i} - WE_{r_i}|} + e^{\beta |WE_{s_i^*} - WE_{r_i^*}|}}$$

Maximum sampled partial likelihood

There are a number of estimation paradigms:

- **MLE/Bayesian:** intractable even for small networks.
- **Partial likelihood:** denominator of PL is still too large.

Case-control Partial Likelihood:

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

$$\begin{aligned}\hat{\beta} &= \arg\max \prod_{i=1}^{10} \frac{e^{\beta|WE_{s_i} - WE_{r_i}|}}{e^{\beta|WE_{s_i} - WE_{r_i}|} + e^{\beta|WE_{s_i^*} - WE_{r_i^*}|}} \\ &= \arg\max \prod_{i=1}^{10} \frac{e^{\beta\left(|WE_{s_i} - WE_{r_i}| - |WE_{s_i^*} - WE_{r_i^*}|\right)}}{1 + e^{\beta\left(|WE_{s_i} - WE_{r_i}| - |WE_{s_i^*} - WE_{r_i^*}|\right)}}\end{aligned}$$

This is equivalent with **logistic regression**, where

- responses are ones, $y = (1, 1, \ldots, 1)$
- covariates are differences, $|WE_{s_i} - WE_{r_i}| - |WE_{s_i^*} - WE_{r_i^*}|$.

Sampling non-events

Original data:

	time	s	r
1	2.4	4	5
2	16.0	2	3
3	18.0	3	2
4	23.7	5	4
5	66.7	2	1
6	76.8	1	2
7	264.4	3	1
8	349.8	4	3
9	694.1	1	3
10	7463.6	5	3

Sampling non-events

Original data with sampled non-event:

	tm	s	r	non.s	non.r
1	2.4	4	5	5	3
2	16.0	2	3	2	1
3	18.0	3	2	3	5
4	23.7	5	4	1	3
5	66.7	2	1	4	3
6	76.8	1	2	3	5
7	264.4	3	1	3	4
8	349.8	4	3	4	2
9	694.1	1	3	5	3
10	7463.6	5	3	5	2

Sampling non-events

Original data with sampled non-event and covariates:

	tm	s	r	non.s	non.r	x.WE	x.non.WE
1	2.4	4	5	5	3	2	6
2	16.0	2	3	2	1	2	3
3	18.0	3	2	3	5	2	6
4	23.7	5	4	1	3	2	5
5	66.7	2	1	4	3	3	4
6	76.8	1	2	3	5	3	6
7	264.4	3	1	3	4	5	4
8	349.8	4	3	4	2	4	6
9	694.1	1	3	5	3	5	6
10	7463.6	5	3	5	2	6	8

Sampling non-events

Original data with sampled non-event and covariate differences:

	tm	s	r	non.s	non.r	x.WE	x.non.WE	WE.diff
1	2.4	4	5	5	3	2	6	-4
2	16.0	2	3	2	1	2	3	-1
3	18.0	3	2	3	5	2	6	-4
4	23.7	5	4	1	3	2	5	-3
5	66.7	2	1	4	3	3	4	-1
6	76.8	1	2	3	5	3	6	-3
7	264.4	3	1	3	4	5	4	1
8	349.8	4	3	4	2	4	6	-2
9	694.1	1	3	5	3	5	6	-1
10	7463.6	5	3	5	2	6	8	-2

Sampling non-events

Original data with sampled non-event with WE difference:

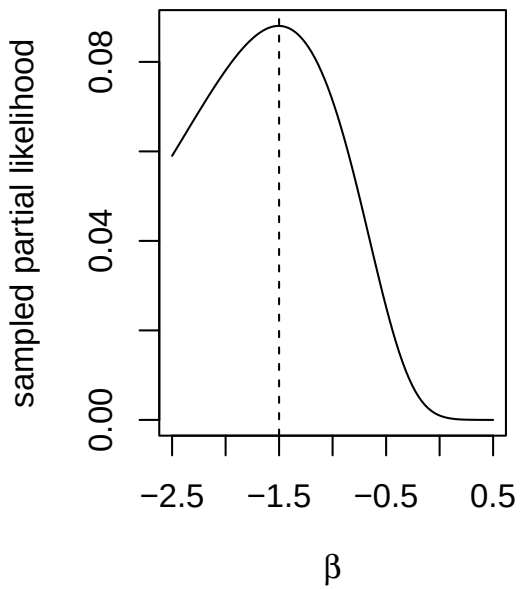
	tm	s	r	non.s	non.r	WE.diff
1	2.4	4	5	5	3	-4
2	16.0	2	3	2	1	-1
3	18.0	3	2	3	5	-4
4	23.7	5	4	1	3	-3
5	66.7	2	1	4	3	-1
6	76.8	1	2	3	5	-3
7	264.4	3	1	3	4	1
8	349.8	4	3	4	2	-2
9	694.1	1	3	5	3	-1
10	7463.6	5	3	5	2	-2

$$L(\beta) = \frac{e^{-4\beta}}{1+e^{-4\beta}} \times \frac{e^{-1\beta}}{1+e^{-1\beta}} \times \cdots \times \frac{e^{-2\beta}}{1+e^{-2\beta}}$$

Sampling non-events

Original data with sampled non-event with WE difference:

	tm	s	r	non.s	non.r	WE.diff
1	2.4	4	5	5	3	-4
2	16.0	2	3	2	1	-1
3	18.0	3	2	3	5	-4
4	23.7	5	4	1	3	-3
5	66.7	2	1	4	3	-1
6	76.8	1	2	3	5	-3
7	264.4	3	1	3	4	1
8	349.8	4	3	4	2	-2
9	694.1	1	3	5	3	-1
10	7463.6	5	3	5	2	-2

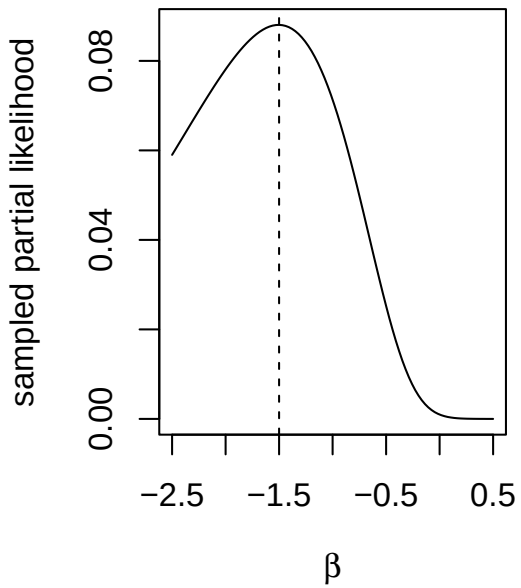


$$L(\beta) = \frac{e^{-4\beta}}{1+e^{-4\beta}} \times \frac{e^{-1\beta}}{1+e^{-1\beta}} \times \ldots \times \frac{e^{-2\beta}}{1+e^{-2\beta}}$$

Sampling non-events

Original data with sampled non-event with WE difference:

	tm	s	r	non.s	non.r	WE.diff
1	2.4	4	5	5	3	-4
2	16.0	2	3	2	1	-1
3	18.0	3	2	3	5	-4
4	23.7	5	4	1	3	-3
5	66.7	2	1	4	3	-1
6	76.8	1	2	3	5	-3
7	264.4	3	1	3	4	1
8	349.8	4	3	4	2	-2
9	694.1	1	3	5	3	-1
10	7463.6	5	3	5	2	-2



$$L(\beta) = \frac{e^{-4\beta}}{1+e^{-4\beta}} \times \frac{e^{-1\beta}}{1+e^{-1\beta}} \times \ldots \times \frac{e^{-2\beta}}{1+e^{-2\beta}}$$

Sampled partial likelihood is maximized at ... $\beta = -1.5$!

Fitting dynamic networks with logistic regression!

We can obtain the same results directly via logistic regression:

```
> WE.diff<-abs(WE[dat$solicitor]-WE[dat$receiver]) -
+              abs(WE[dat$non.solicitor]-WE[dat$non.receiver])
> y<-rep(1,length(WE.diff))
> summary(glm(y ~ -1 + WE.diff, family = binomial))

Call:
glm(formula = y ~ -1 + WE.diff, family = binomial)

Coefficients:
        Estimate Std. Error z value Pr(>|z|)
WE.diff  -1.5042     0.9037  -1.665    0.096 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13.8629  on 10  degrees of freedom
Residual deviance:  4.8597  on  9  degrees of freedom
AIC: 6.8597

Number of Fisher Scoring iterations: 6
```

Fitting dynamic networks with logistic regression!

We can obtain the same results directly via logistic regression:

```
> WE.diff<-abs(WE[dat$solicitor]-WE[dat$receiver]) -
+              abs(WE[dat$non.solicitor]-WE[dat$non.receiver])
> y<-rep(1,length(WE.diff))
> summary(glm(y ~ -1 + WE.diff, family = binomial))

Call:
glm(formula = y ~ -1 + WE.diff, family = binomial)

Coefficients:
        Estimate Std. Error z value Pr(>|z|)
WE.diff  -1.5042     0.9037  -1.665    0.096 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13.8629  on 10  degrees of freedom
Residual deviance:  4.8597  on  9  degrees of freedom
AIC: 6.8597

Number of Fisher Scoring iterations: 6
```

Fitting dynamic networks with logistic regression!

We can obtain the same results directly via logistic regression:

```
> WE.diff<-abs(WE[dat$solicitor]-WE[dat$receiver]) -
+              abs(WE[dat$non.solicitor]-WE[dat$non.receiver])
> y<-rep(1,length(WE.diff))
> summary(glm(y ~ -1 + WE.diff, family = binomial))

Call:
glm(formula = y ~ -1 + WE.diff, family = binomial)

Coefficients:
        Estimate Std. Error z value Pr(>|z|)
WE.diff  -1.5042     0.9037  -1.665    0.096 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13.8629  on 10  degrees of freedom
Residual deviance:  4.8597  on  9  degrees of freedom
AIC: 6.8597

Number of Fisher Scoring iterations: 6
```

Fitting dynamic networks with logistic regression!

We can obtain the same results directly via logistic regression:

```
> WE.diff<-abs(WE[dat$solicitor]-WE[dat$receiver]) -
+              abs(WE[dat$non.solicitor]-WE[dat$non.receiver])
> y<-rep(1,length(WE.diff))
> summary(glm(y ~ -1 + WE.diff, family = binomial))

Call:
glm(formula = y ~ -1 + WE.diff, family = binomial)

Coefficients:
        Estimate Std. Error z value Pr(>|z|)
WE.diff  -1.5042     0.9037  -1.665    0.096 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13.8629  on 10  degrees of freedom
Residual deviance:  4.8597  on  9  degrees of freedom
AIC: 6.8597

Number of Fisher Scoring iterations: 6
```

Fitting dynamic networks with logistic regression!

We can obtain the same results directly via logistic regression:

```
> WE.diff<-abs(WE[dat$solicitor]-WE[dat$receiver]) -
+              abs(WE[dat$non.solicitor]-WE[dat$non.receiver])
> y<-rep(1,length(WE.diff))
> summary(glm(y ~ -1 + WE.diff, family = binomial))

Call:
glm(formula = y ~ -1 + WE.diff, family = binomial)

Coefficients:
        Estimate Std. Error z value Pr(>|z|)
WE.diff  -1.5042     0.9037  -1.665    0.096 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13.8629  on 10  degrees of freedom
Residual deviance:  4.8597  on  9  degrees of freedom
AIC: 6.8597

Number of Fisher Scoring iterations: 6
```

Take-home messages

This tutorial considers **dynamic networks**

1. As series of *relational events* in time,
2. between set of Senders and Receivers:
   - One-mode networks: Senders $=$ Receivers
   - Bipartite networks: Senders $\neq$ Receivers
3. Model dynamic network as **general counting process**:
   - event times are (generalized) exponentially distributed
   - event types are multinomially distributed
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 **logistic regression** with
     - only successes
     - no intercept
     - covariates: $\Delta x_i := x_{s_i r_i} - x_{s_i^* r_i^*}$