Inverse Problems for Philosophers

Bridging the gap between agent-based models and behavioral data

Lucas Gautheron[1,2] (Ph.D. candidate)

[1]Interdisciplinary Center for Science and Technology Studies, Wuppertal, Germany
[2]Département d'Études Cognitives, École Normale Supérieure, Paris, France

University of Bochum, December 2024

Summary

1. Inverse problems for philosophers and agent-based modelers

2. A case-study of conventions: the metric signature in particle physics
   - How do physicists choose which convention to use in their own papers?
   - How do scientists resolve conflicting preferences in collaborations?
   - How do physicists' preferences get formed?

Why should philosophers care about data?

- Verify that models capture what is actually going on in the situations we are interested in.

Why should philosophers care about data?

- Verify that models capture what is actually going on in the situations we are interested in.
- Robustness (insensitivity to model assumptions/parameters) is a fallible validation procedure: what if the outcome really is contingent on certain circumstances (the values of underlying parameters, the topology of some relevant network, etc.)

Why should philosophers care about data?

- Verify that models capture what is actually going on in the situations we are interested in.
- Robustness (insensitivity to model assumptions/parameters) is a fallible validation procedure: what if the outcome really is contingent on certain circumstances (the values of underlying parameters, the topology of some relevant network, etc.)
- Insights from models without connection to data may not be translatable into interventions/policies.

Why should philosophers care about data?

- Verify that models capture what is actually going on in the situations we are interested in.
- Robustness (insensitivity to model assumptions/parameters) is a fallible validation procedure: what if the outcome really is contingent on certain circumstances (the values of underlying parameters, the topology of some relevant network, etc.)
- Insights from models without connection to data may not be translatable into interventions/policies.
- Unvalidated models should maybe not provide guidance for policy-making.

Why should philosophers care about data?

- Verify that models capture what is actually going on in the situations we are interested in.
- Robustness (insensitivity to model assumptions/parameters) is a fallible validation procedure: what if the outcome really is contingent on certain circumstances (the values of underlying parameters, the topology of some relevant network, etc.)
- Insights from models without connection to data may not be translatable into interventions/policies.
- Unvalidated models should maybe not provide guidance for policy-making.

$\Rightarrow$ inverse problems are a promising candidate for bridging the formal/empirical gap.

Summary

1. Inverse problems for philosophers and agent-based modelers

2. A case-study of conventions: the metric signature in particle physics
   - How do physicists choose which convention to use in their own papers?
   - How do scientists resolve conflicting preferences in collaborations?
   - How do physicists' preferences get formed?

What are inverse problems

- Inverse problems seek to **infer the invisible causes underlying a set of observations**.

Rules governing agents' behavior $\xrightarrow{\text{“Forward problem”}}$ Outcome of agents' behavior

What are inverse problems

- Inverse problems seek to **infer the invisible causes underlying a set of observations**.
- In the context of Agent-Based Modeling:

Rules governing agents' behavior $\xrightarrow{\text{“Forward problem”}}$ Outcome of agents' behavior

What are inverse problems

- Inverse problems seek to **infer the invisible causes underlying a set of observations**.
- In the context of Agent-Based Modeling:

$$\text{Rules governing agents' behavior} \underset{\text{"Inverse problem"}}{\overset{\text{"Forward problem"}}{\rightleftarrows}} \text{Outcome of agents' behavior}$$

What are inverse problems

- Inverse problems seek to **infer the invisible causes underlying a set of observations**.
- In the context of Agent-Based Modeling:

$$\begin{tikzcd} \text{Rules governing agents' behavior} \arrow[r, shift left, "\text{``Forward problem''}"] & \text{Outcome of agents' behavior} \arrow[l, shift left, "\text{``Inverse problem''}"] \end{tikzcd}$$

- Inverse problems are **hard**:

What are inverse problems

- Inverse problems seek to **infer the invisible causes underlying a set of observations**.
- In the context of Agent-Based Modeling:

Rules governing agents' behavior	"Forward problem" $\longrightarrow$ $\longleftarrow$ "Inverse problem"	Outcome of agents' behavior

- Inverse problems are **hard**:
  1. **Identifiability problems** (underdetermination): many causes could have produced a given outcome

What are inverse problems

- Inverse problems seek to **infer the invisible causes underlying a set of observations**.
- In the context of Agent-Based Modeling:

Rules governing agents' behavior	"Forward problem" $\longrightarrow$ $\longleftarrow$ "Inverse problem"	Outcome of agents' behavior

- Inverse problems are **hard**:
  1. **Identifiability problems** (underdetermination): many causes could have produced a given outcome
  2. **Misspecification problems**: inverse problems may produce misleading results when modeling assumptions are "too wrong".

What are inverse problems

- Inverse problems seek to **infer the invisible causes underlying a set of observations**.
- In the context of Agent-Based Modeling:

$$\text{Rules governing agents' behavior} \;\underset{\text{“Inverse problem”}}{\overset{\text{“Forward problem”}}{\rightleftarrows}}\; \text{Outcome of agents' behavior}$$

- Inverse problems are **hard**:
  1. **Identifiability problems** (underdetermination): many causes could have produced a given outcome
  2. **Misspecification problems**: inverse problems may produce misleading results when modeling assumptions are “too wrong”.
  3. **Computational problems**: solving inverse problems often involves intractable computations and requires approximation schemes.

Bayesian inference for inverse problems

- Both forward models and inverse problems have a stochastic/probabilistic component (random initialization, partially random decisions, uncertainty quantification. . .)
- We appeal to **probabilities** and **Bayesian inference**.

$$\begin{array}{ccc} \text{Rules governing} & \text{“Forward problem”} & \text{Outcome of} \\ \text{agents' behavior} & \xrightarrow{\text{Estimate } \boldsymbol{P(O|M)}} & \text{agents' behavior} \\ \textbf{Model } \boldsymbol{M} & \xleftarrow[\text{Estimate } \boldsymbol{P(M|O)}]{} & \textbf{Outcome } \boldsymbol{O} \\ & \text{“Inverse problem”} & \end{array}$$

$$P(M|O) = \frac{P(O|M)\overbrace{P(M)}^{\text{Prior}}}{P(O)} \tag{1}$$

$$P(M|O) = \frac{P(O|M)\,\overbrace{P(M)}^{\text{Prior}}}{P(O)} \tag{2}$$

Model comparison and parameter estimation

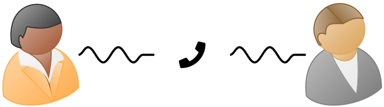
$$P(\theta|O) = \frac{P(O|\theta)\,\overbrace{P(\theta)}^{\text{Prior}}}{P(O)} \qquad (2)$$

Summary

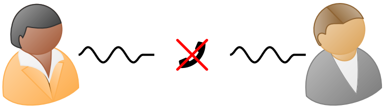
1. Inverse problems for philosophers and agent-based modelers

2. A case-study of conventions: the metric signature in particle physics
    - How do physicists choose which convention to use in their own papers?
    - How do scientists resolve conflicting preferences in collaborations?
    - How do physicists' preferences get formed?

- **Coordination problems** arise when individuals would benefit from acting in a mutually compatible way, but it is somehow non-trivial to do so (Lewis, 1969).

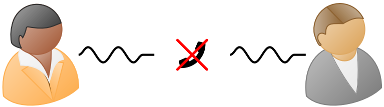


- **Coordination problems** arise when individuals would benefit from acting in a mutually compatible way, but it is somehow non-trivial to do so (Lewis, 1969).



Conventions

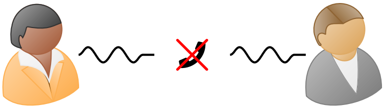
- **Coordination problems** arise when individuals would benefit from acting in a mutually compatible way, but it is somehow non-trivial to do so (Lewis, 1969).



	Bob calls back	Bob awaits
Alice calls back	0,0	1,1
Alice awaits	1,1	0,0

Conventions

- **Coordination problems** arise when individuals would benefit from acting in a mutually compatible way, but it is somehow non-trivial to do so (Lewis, 1969).

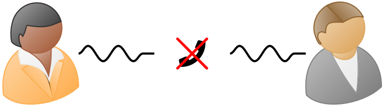


	Bob calls back	**Bob awaits**
Alice calls back	0,0	1,1
Alice awaits	1,1	0,0

- **"Conventions"** are cultural tools for solving coordination problems by providing individuals with expectations about how others will behave. These expectations suggest particular courses of action.

Conventions

- **Coordination problems** arise when individuals would benefit from acting in a mutually compatible way, but it is somehow non-trivial to do so (Lewis, 1969).

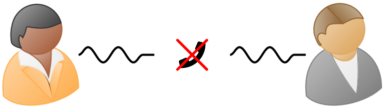


	Bob calls back	**Bob awaits**
Alice calls back	0,0	1,1
Alice awaits	1,1	0,0

- **"Conventions"** are cultural tools for solving coordination problems by providing individuals with expectations about how others will behave. These expectations suggest particular courses of action.
  - Example: left-hand or right-hand traffic.

Conventions

- **Coordination problems** arise when individuals would benefit from acting in a mutually compatible way, but it is somehow non-trivial to do so (Lewis, 1969).



	Bob calls back	**Bob awaits**
Alice calls back	0,0	1,1
Alice awaits	1,1	0,0

- **"Conventions"** are cultural tools for solving coordination problems by providing individuals with expectations about how others will behave. These expectations suggest particular courses of action.
  - Example: left-hand or right-hand traffic.
  - Language!

Conventions in the literature

- Can conventions emerge spontaneously from dyadic interactions alone? (Centola and Baronchelli, 2015; Hawkins, Goodman, and Goldstone, 2019)
- How does the topology of social networks influence the propagation of conventions via dyadic interactions? (Pujol et al., 2005; Delgado, 2002)
- How to measure the degree of conventionality of a convention? (O'Connor, 2020)
- What is the role of leadership in addressing coordination problems? (Calvert, 1992)

A case-study from high-energy physics

- Relativity: unified description of spacetime.

A case-study from high-energy physics

- Relativity: unified description of spacetime.
- The metric tensor ($g_{\mu\nu}$) captures the metric properties of spacetime; e.g. the pseudo- distance between events (t_1, x_1, y_1, z_1) and (t_2, x_2, y_2, z_2). **Two possible descriptions**:

A case-study from high-energy physics

- Relativity: unified description of spacetime.
- The metric tensor ($g_{\mu\nu}$) captures the metric properties of spacetime; e.g. the pseudo- distance between events (t_1, x_1, y_1, z_1) and (t_2, x_2, y_2, z_2). **Two possible descriptions**:

$$\begin{pmatrix} +1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \text{ or } \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & +1 & 0 & 0 \\ 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & +1 \end{pmatrix}?$$

A case-study from high-energy physics

- Relativity: unified description of spacetime.
- The metric tensor $(g_{\mu\nu})$ captures the metric properties of spacetime; e.g. the pseudo- distance between events (t_1, x_1, y_1, z_1) and (t_2, x_2, y_2, z_2). **Two possible descriptions**:

$$\begin{pmatrix} +1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \text{ or } \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & +1 & 0 & 0 \\ 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & +1 \end{pmatrix}?$$

$$\text{“mostly minus” (-1) or “mostly plus” } (+1) \tag{3}$$

A case-study from high-energy physics

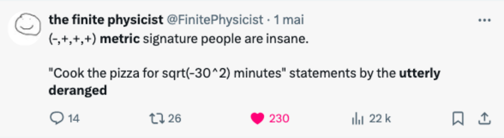
- Relativity: unified description of spacetime.
- The metric tensor ($g_{\mu\nu}$) captures the metric properties of spacetime; e.g. the pseudo- distance between events (t_1, x_1, y_1, z_1) and (t_2, x_2, y_2, z_2). **Two possible descriptions**:

$$\begin{pmatrix} +1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \text{ or } \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & +1 & 0 & 0 \\ 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & +1 \end{pmatrix}?$$

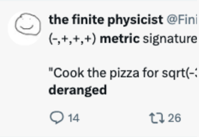
$$\text{“mostly minus” (-1) or “mostly plus” } (+1) \tag{3}$$

- Both choices are legitimate, as long as one remains consistent.

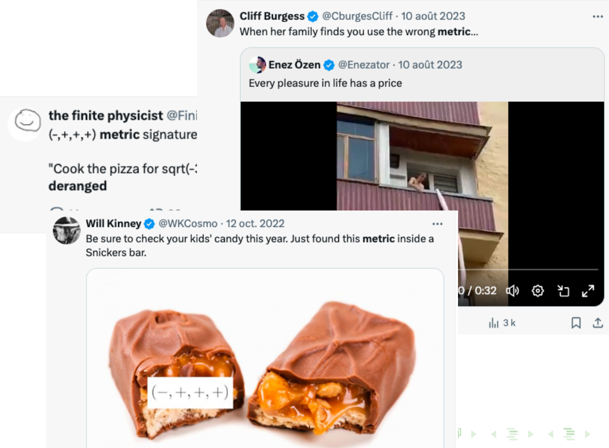
A heated debate



A heated debate

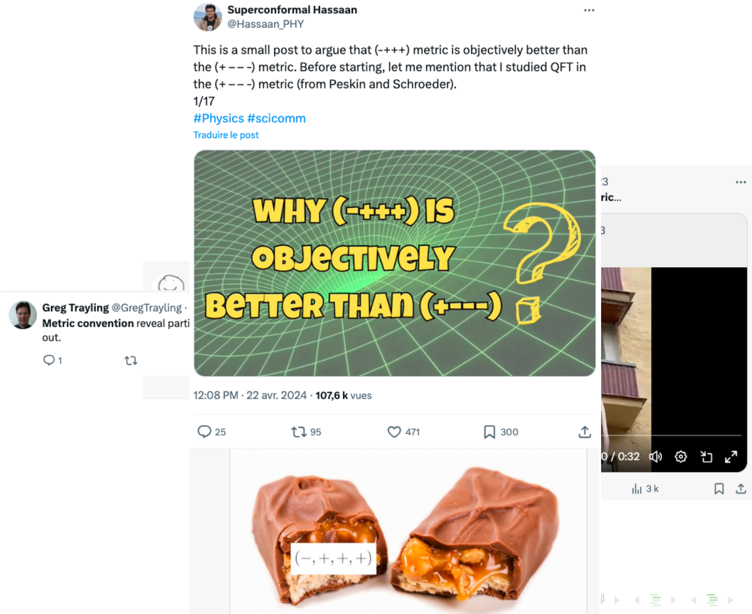


A heated debate



A heated debate

A heated debate



Inverse problems and conventions

- Let's use inverse problems to infer:

Inverse problems and conventions

- Let's use inverse problems to infer:
  1. How do scientists decide to choose one convention or the other in a paper?

Inverse problems and conventions

- Let's use inverse problems to infer:
  1. How do scientists decide to choose one convention or the other in a paper?
  2. How do they resolve conflicting preferences in collaborations?

Inverse problems and conventions

- Let's use inverse problems to infer:
  1. How do scientists decide to choose one convention or the other in a paper?
  2. How do they resolve conflicting preferences in collaborations?
  3. What factors shape scientists' preferences?

Data

- Data collected from **Inspire HEP** (authorship/citation metadata) and **arXiv** (LaTeX source)
- Categories: hep-th (high-energy physics theory), hep-ph (phenomenology), gr-qc (gravitation and cosmology), astro-ph (astrophysics)
- 22 500 papers classified according to their metric signature (mostly plus or mostly minus) using regular expressions.

1. Inverse problems for philosophers and agent-based modelers

2. A case-study of conventions: the metric signature in particle physics
   - How do physicists choose which convention to use in their own papers?
   - How do scientists resolve conflicting preferences in collaborations?
   - How do physicists' preferences get formed?

How do physicists choose which convention to use in their own papers?

Individuals' attitude towards a convention may be shaped by:

How do physicists choose which convention to use in their own papers?

Individuals' attitude towards a convention may be shaped by:

1. **Coordination costs** (the cost of disagreeing with peers).

How do physicists choose which convention to use in their own papers?

Individuals' attitude towards a convention may be shaped by:

1. **Coordination costs** (the cost of disagreeing with peers).
2. **Inconsistency costs** (the cost of switching between different conventions).

How do physicists choose which convention to use in their own papers?

Individuals' attitude towards a convention may be shaped by:

1. **Coordination costs** (the cost of disagreeing with peers).
2. **Inconsistency costs** (the cost of switching between different conventions).
3. **Maladaptation costs** (the cost of using a convention that is a poor fit/suboptimal in a given context). Degrees of conventionality (O'Connor, 2020)

How do physicists choose which convention to use in their own papers?

Individuals' attitude towards a convention may be shaped by:

1. **Coordination costs** (the cost of disagreeing with peers).
2. **Inconsistency costs** (the cost of switching between different conventions).
3. **Maladaptation costs** (the cost of using a convention that is a poor fit/suboptimal in a given context). Degrees of conventionality (O'Connor, 2020)

$\Rightarrow$ What are their implications for the diffusion of conventions? Are these involved in the context of the metric signature?

Inconsistency and maladaptation costs

- Is physicists' attitude towards the convention dictated by consistency or adaptation (fitness) to their research?

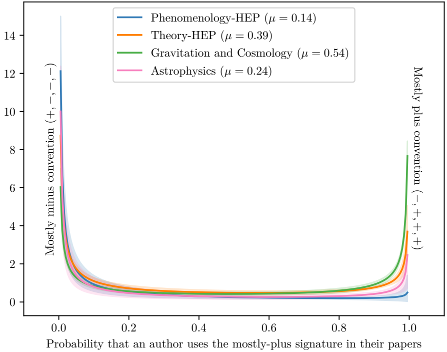


Figure: Physicists tend to always be using the same convention

Inconsistency and maladaptation costs

-  publishes in category $c_d \in \{\text{phenomenology, theory}, \dots\}$. What is the probability that they use the mostly plus convention?

$P(\sigma_d = +1 | a_d =$  $, c_d) = f\big(\theta($  $)$ $+ \underbrace{b(c_d)}_{\text{Effect of research area } c_d} \big)$ (4)

Author's preference

Inconsistency and maladaptation costs

-  publishes in category $c_d \in \{\text{phenomenology, theory}, \dots\}$. What is the probability that they use the mostly plus convention?

$P(\sigma_d = +1 | a_d =$  $, c_d) = f\big($ $\theta($  $)$ $+ \underbrace{b(c_d)}_{\text{Effect of research area } c_d} \big)$ (4)

Author's preference

- "Item-response model": recover invisible traits/factors that may account for observed behaviors.

Inconsistency and maladaptation costs

-  publishes in category $c_d \in \{\text{phenomenology, theory, } \dots\}$. What is the probability that they use the mostly plus convention?

$P(\sigma_d = +1 | a_d =$  $, c_d) = f(\underbrace{\theta(}_{}$  $\underbrace{)}_{\text{Author's preference}} + \underbrace{b(c_d)}_{\text{Effect of research area } c_d})$ (4)

- "Item-response model": recover invisible traits/factors that may account for observed behaviors.
- $\theta(i)$ is a latent (unobserved) parameter measuring the preference of each author i. $\theta(i) > 0$ indicates a preference for the mostly plus signature

Inconsistency and maladaptation costs

-  publishes in category $c_d \in \{\text{phenomenology, theory, } \ldots\}$. What is the probability that they use the mostly plus convention?

$P(\sigma_d = +1|a_d =$  $, c_d) = f\Big(\underbrace{\theta(}$  $\underbrace{)}_{\text{Author's preference}} + \underbrace{b(c_d)}_{\text{Effect of research area } c_d}\Big)$ (4)

- "Item-response model": recover invisible traits/factors that may account for observed behaviors.
- $\theta(i)$ is a latent (unobserved) parameter measuring the preference of each author i. $\theta(i) > 0$ indicates a preference for the mostly plus signature
- b_c is the unobserved bias associated with research area c

Inconsistency and maladaptation costs

-  publishes in category $c_d \in \{\text{phenomenology, theory}, \ldots\}$. What is the probability that they use the mostly plus convention?

$P(\sigma_d = +1 | a_d =$  $, c_d) = f(\underbrace{\theta(}$  $\underbrace{)}_{\text{Author's preference}} + \underbrace{b(c_d)}_{\text{Effect of research area } c_d}) \quad (4)$

- "Item-response model": recover invisible traits/factors that may account for observed behaviors.
- $\theta(i)$ is a latent (unobserved) parameter measuring the preference of each author i. $\theta(i) > 0$ indicates a preference for the mostly plus signature
- b_c is the unobserved bias associated with research area c
- If $|\theta| \gg |b|$ then individual preferences dominate the need to adapt to a given research area

Inconsistency and maladaptation costs

- publishes in category $c_d \in \{\text{phenomenology, theory, } \ldots\}$. What is the probability that they use the mostly plus convention?

$$P(\sigma_d = +1 | a_d = \;, c_d) = f(\underbrace{\theta(\;)}_{\text{Author's preference}} + \underbrace{b(c_d)}_{\text{Effect of research area } c_d}) \quad (4)$$

- "Item-response model": recover invisible traits/factors that may account for observed behaviors.
- $\theta(i)$ is a latent (unobserved) parameter measuring the preference of each author i. $\theta(i) > 0$ indicates a preference for the mostly plus signature
- b_c is the unobserved bias associated with research area c
- If $|\theta| \gg |b|$ then individual preferences dominate the need to adapt to a given research area
- **Given physicists' choices in their solo-authored papers, we can infer back θ and b using Bayesian inference.**

Inconsistency and maladaptation costs

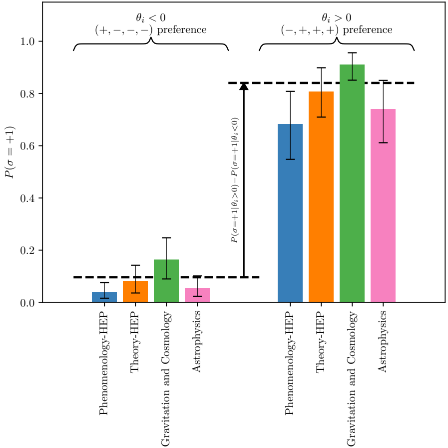


Figure: Consistency matters most, but adaptation to the context can occur.

1. Inverse problems for philosophers and agent-based modelers

2. A case-study of conventions: the metric signature in particle physics
   - How do physicists choose which convention to use in their own papers?
   - How do scientists resolve conflicting preferences in collaborations?
   - How do physicists' preferences get formed?

Decision optimality and decision costs in the resolution of conflicts

- In scientific collaborations, the resolution of disagreement involves two factors:

1 The "optimality" of the decision (i.e., truth-value – if relevant –, collective satisfaction, appropriateness of the solution etc.).

2 The cost of reaching an "optimal" decision.

- Leadership is a tool for reducing "transaction" and decision costs in organizations (Calvert, 1992). Does it play a similar role in the case of the metric signature?

Inferring preference-aggregation mechanisms in conflicts

- Focusing on co-authored papers for which:

Inferring preference-aggregation mechanisms in conflicts

- Focusing on co-authored papers for which:
  (i) The metric signature $S_d \in \{-1, +1\}$ of the paper is observed

Inferring preference-aggregation mechanisms in conflicts

- Focusing on co-authored papers for which:
  (i) The metric signature $S_d \in \{-1, +1\}$ of the paper is observed
  (ii) The preference of each author $(\sigma_1, \dots, \sigma_n) \in \{\pm 1\}^n$ is known independently from at least one solo-authored publication

Inferring preference-aggregation mechanisms in conflicts

- Focusing on co-authored papers for which:
  (i) The metric signature $S_d \in \{-1, +1\}$ of the paper is observed
  (ii) The preference of each author $(\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n$ is known independently from at least one solo-authored publication
- We can assume different preference aggregation strategies (A_k):

Inferring preference-aggregation mechanisms in conflicts

- Focusing on co-authored papers for which:
  (i) The metric signature $S_d \in \{-1, +1\}$ of the paper is observed
  (ii) The preference of each author $(\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n$ is known independently from at least one solo-authored publication
- We can assume different preference aggregation strategies (A_k):
  - Dictatorial strategies (the first author, the last author, or another author decides)

Inferring preference-aggregation mechanisms in conflicts

- Focusing on co-authored papers for which:
  (i) The metric signature $S_d \in \{-1, +1\}$ of the paper is observed
  (ii) The preference of each author $(\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n$ is known independently from at least one solo-authored publication
- We can assume different preference aggregation strategies (A_k):
  - Dictatorial strategies (the first author, the last author, or another author decides)
  - Majoritarian strategy (the majority preference prevails)

Inferring preference-aggregation mechanisms in conflicts

- Focusing on co-authored papers for which:
  (i) The metric signature $S_d \in \{-1, +1\}$ of the paper is observed
  (ii) The preference of each author $(\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n$ is known independently from at least one solo-authored publication
- We can assume different preference aggregation strategies (A_k):
  - Dictatorial strategies (the first author, the last author, or another author decides)
  - Majoritarian strategy (the majority preference prevails)
  - Conventional strategy (the signature most common in the target research area prevails)

Inferring preference-aggregation mechanisms in conflicts

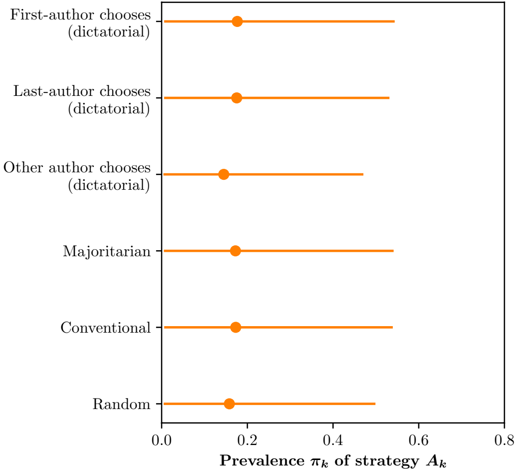
- Focusing on co-authored papers for which:
  (i) The metric signature $S_d \in \{-1, +1\}$ of the paper is observed
  (ii) The preference of each author $(\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n$ is known independently from at least one solo-authored publication
- We can assume different preference aggregation strategies (A_k):
  - Dictatorial strategies (the first author, the last author, or another author decides)
  - Majoritarian strategy (the majority preference prevails)
  - Conventional strategy (the signature most common in the target research area prevails)
  - Random/coin-flip (both individual preferences and context are ignored)

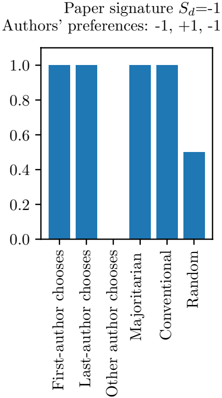
Inferring preference-aggregation mechanisms in conflicts

- Focusing on co-authored papers for which:
  (i) The metric signature $S_d \in \{-1, +1\}$ of the paper is observed
  (ii) The preference of each author $(\sigma_1, \dots, \sigma_n) \in \{\pm 1\}^n$ is known independently from at least one solo-authored publication
- We can assume different preference aggregation strategies (A_k):
  - Dictatorial strategies (the first author, the last author, or another author decides)
  - Majoritarian strategy (the majority preference prevails)
  - Conventional strategy (the signature most common in the target research area prevails)
  - Random/coin-flip (both individual preferences and context are ignored)
- We can estimate the prevalence of each strategy (π_k) given that they predict different outcomes (different probabilities $P(S_d | \sigma_1, \dots, \sigma_n, A_k)$)

Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

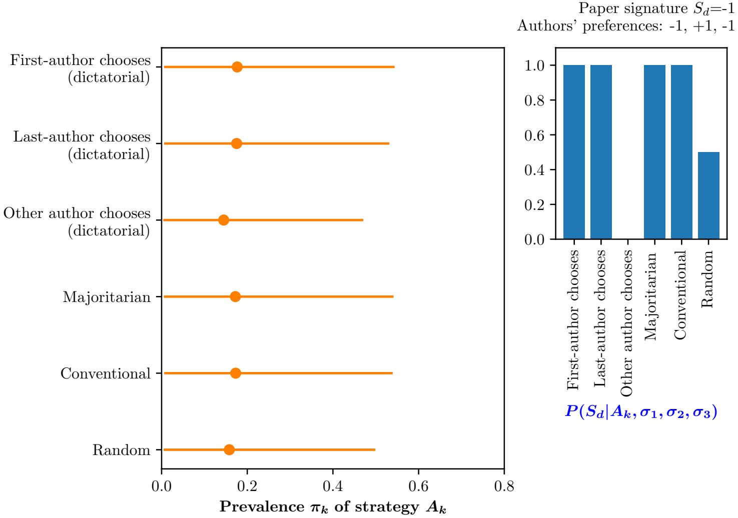




$P(S_d|A_k, \sigma_1, \sigma_2, \sigma_3)$

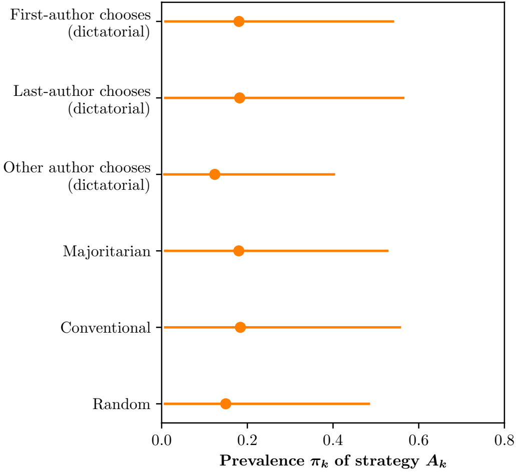
Inferring preference-aggregation mechanisms

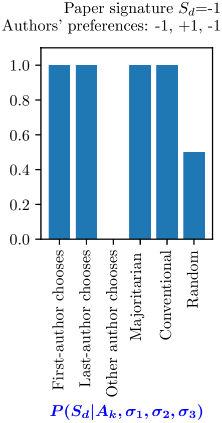
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



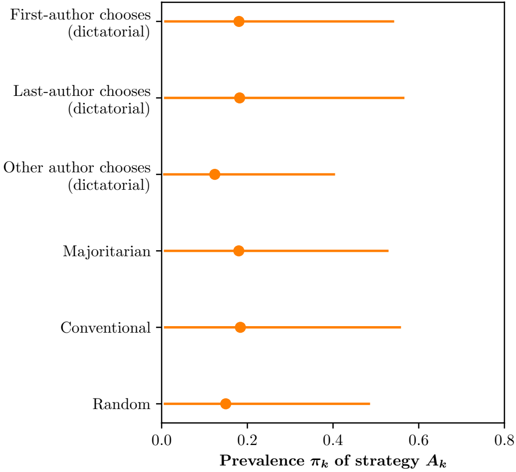
Inferring preference-aggregation mechanisms

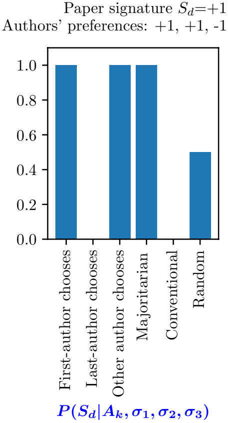
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





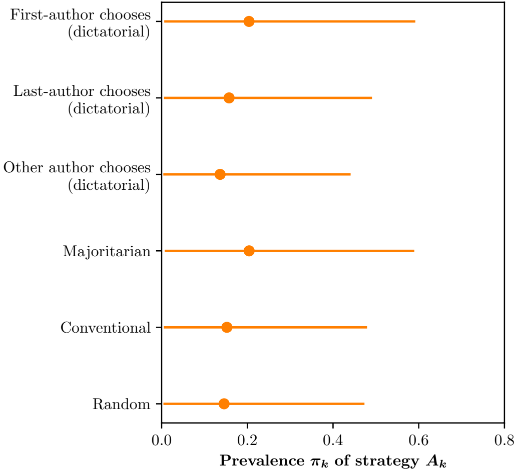
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

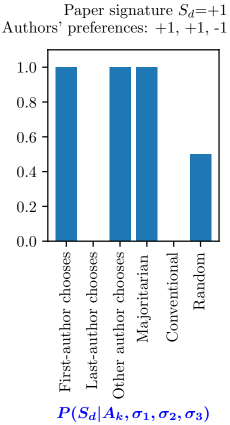




Inferring preference-aggregation mechanisms

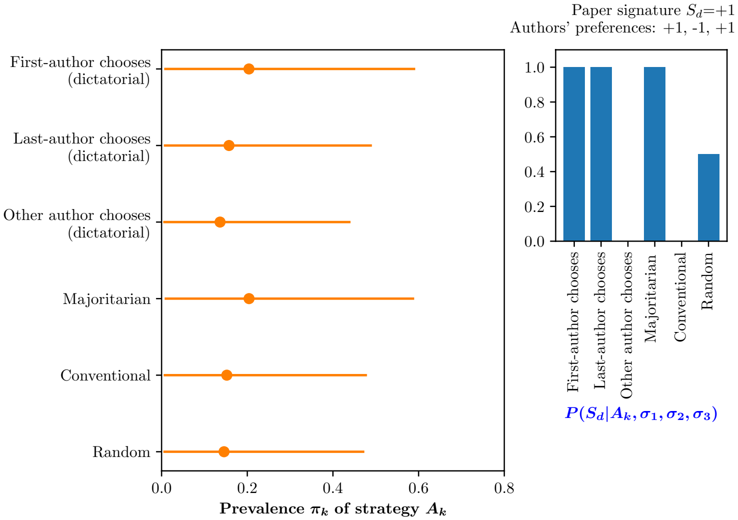
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





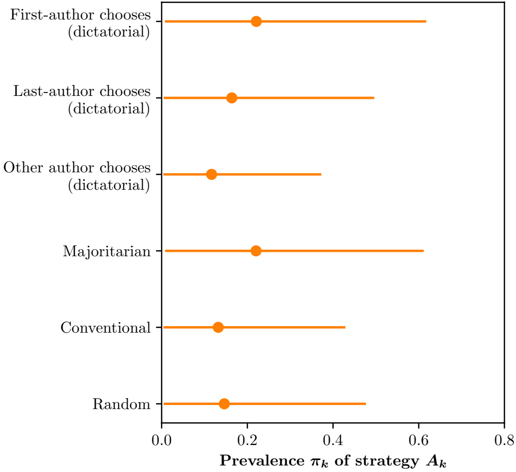
Inferring preference-aggregation mechanisms

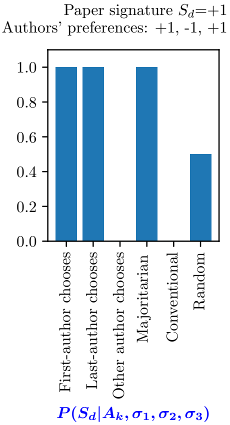
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



Inferring preference-aggregation mechanisms

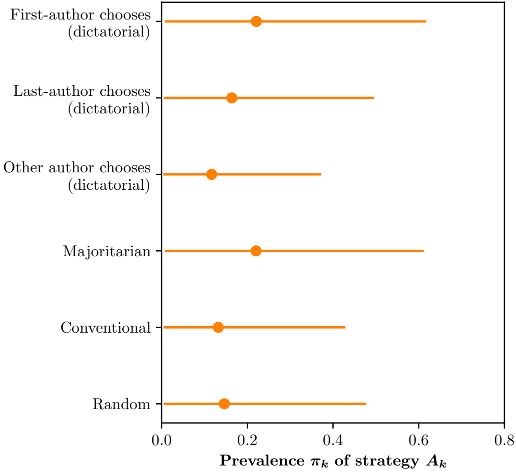
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

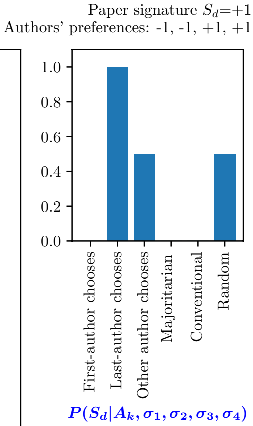




Inferring preference-aggregation mechanisms

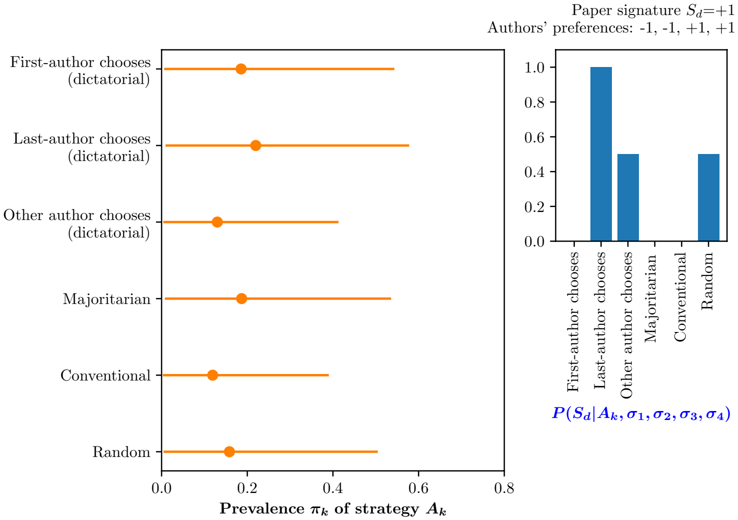
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





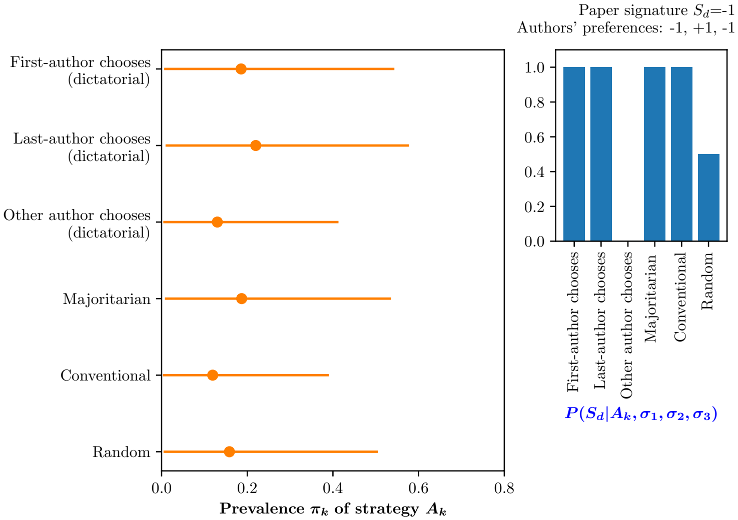
Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



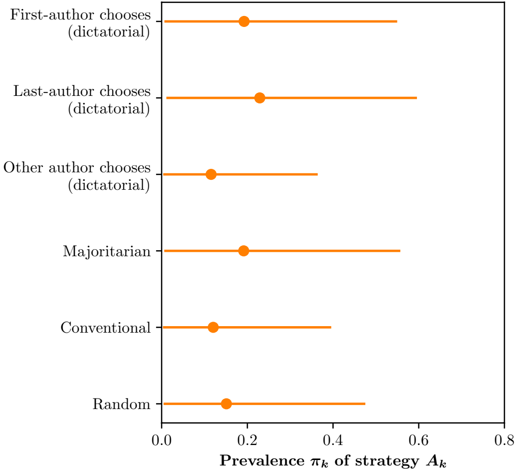
Inferring preference-aggregation mechanisms

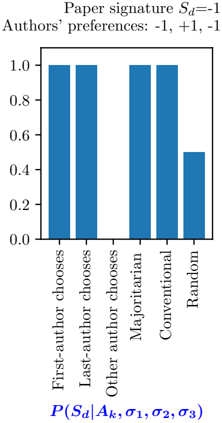
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



Inferring preference-aggregation mechanisms

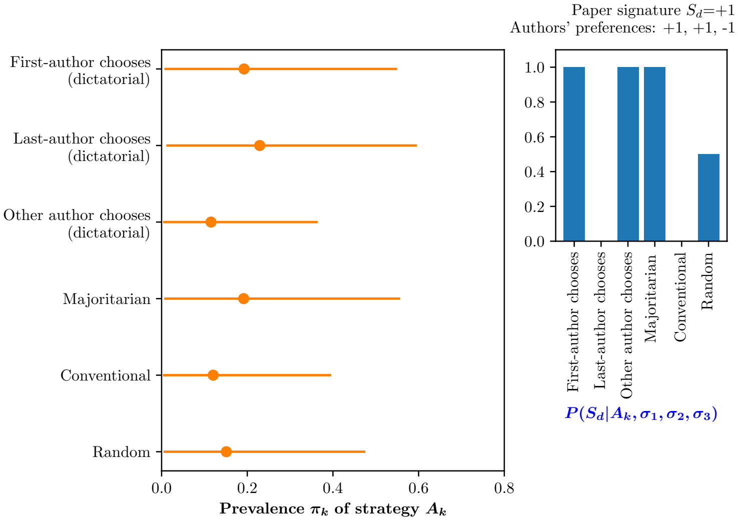
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





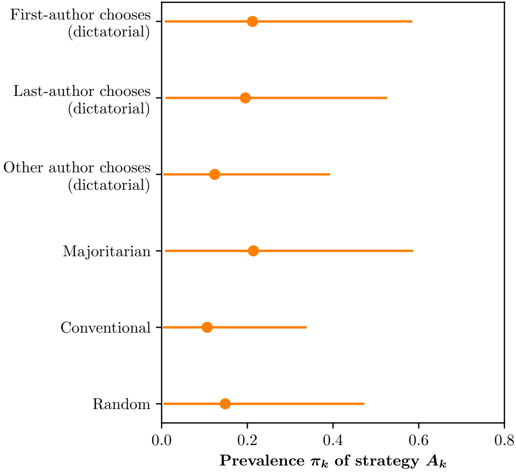
Inferring preference-aggregation mechanisms

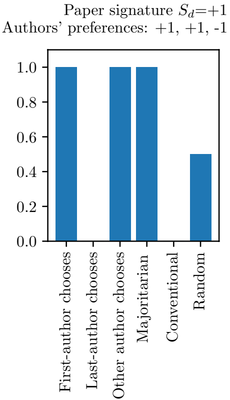
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

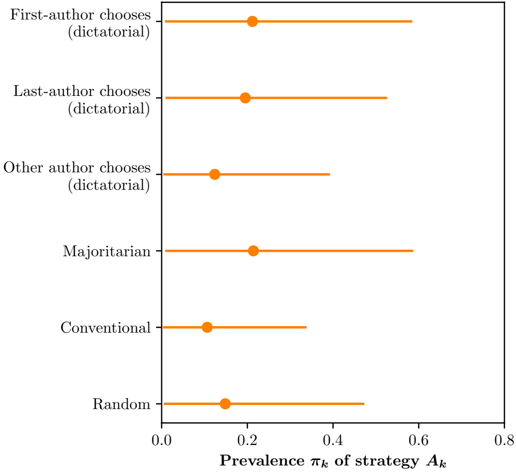


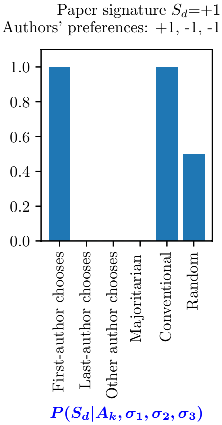


$P(S_d|A_k, \sigma_1, \sigma_2, \sigma_3)$

Inferring preference-aggregation mechanisms

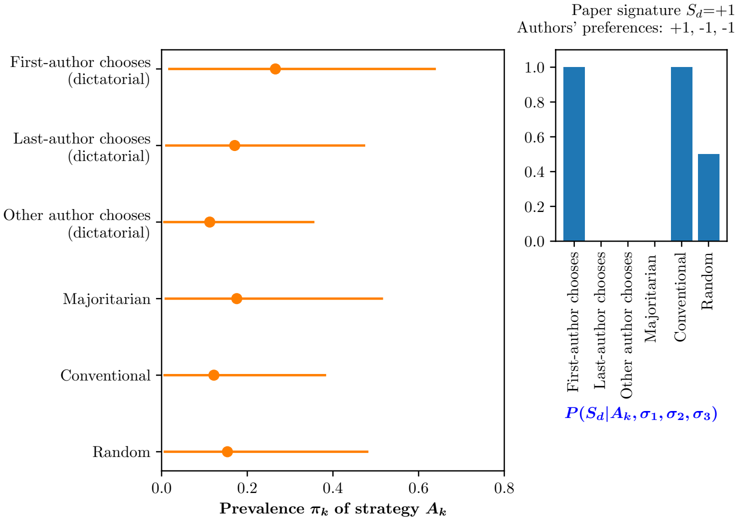
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





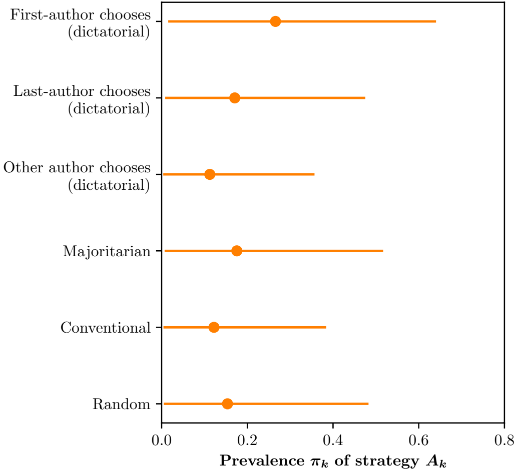
Inferring preference-aggregation mechanisms

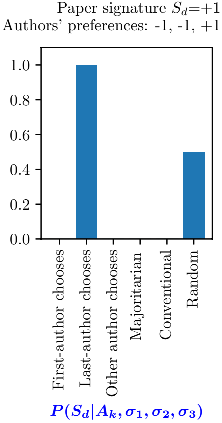
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



Inferring preference-aggregation mechanisms

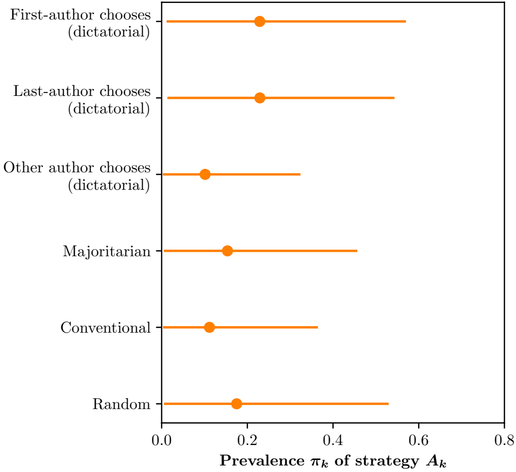
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

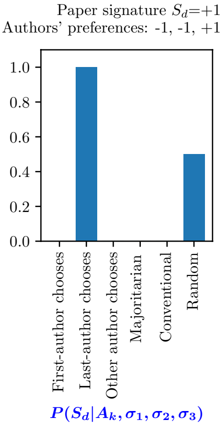




Inferring preference-aggregation mechanisms

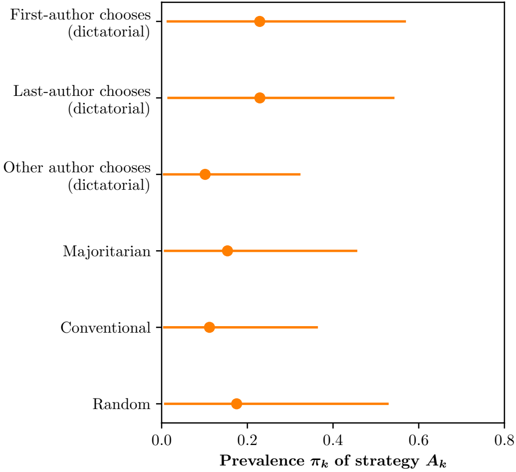
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

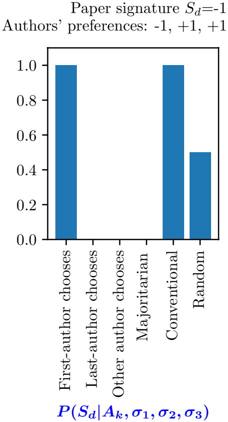




Inferring preference-aggregation mechanisms

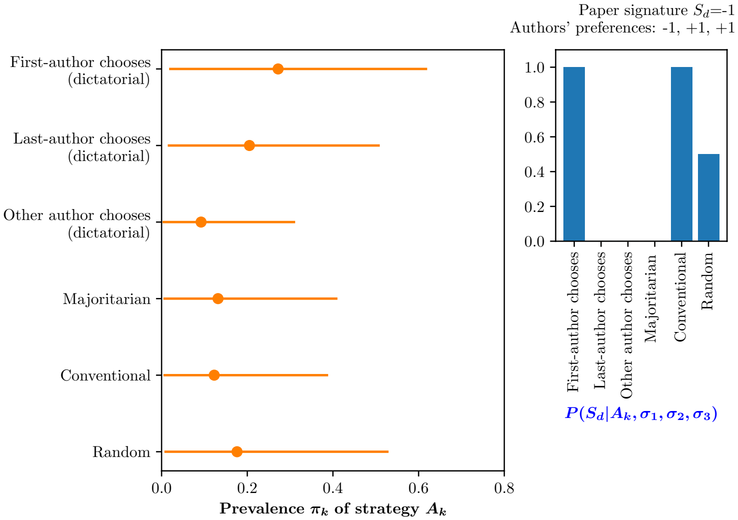
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





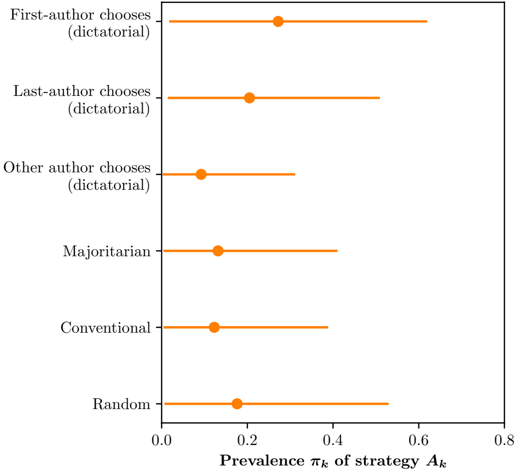
Inferring preference-aggregation mechanisms

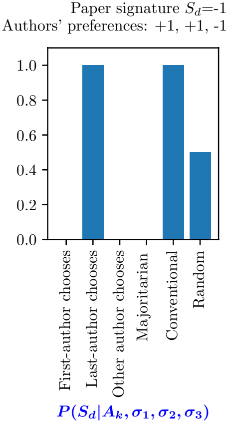
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



Inferring preference-aggregation mechanisms

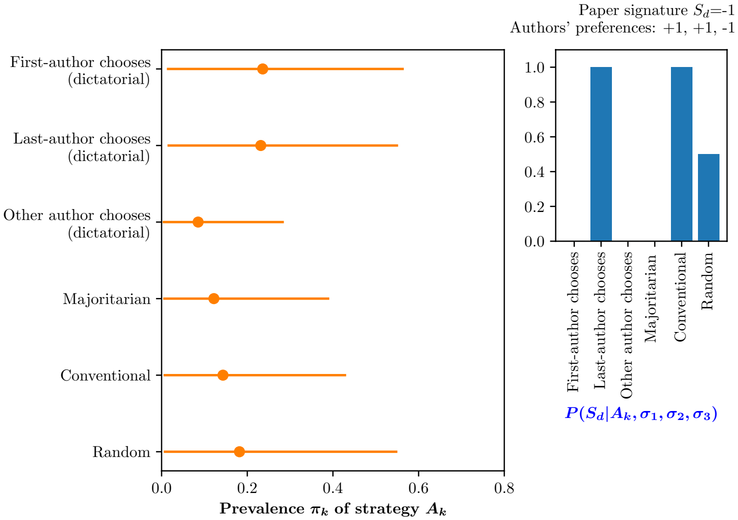
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





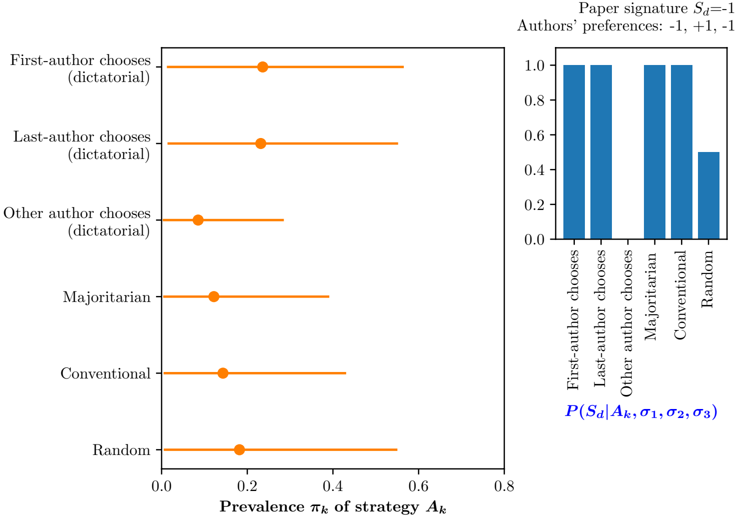
Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



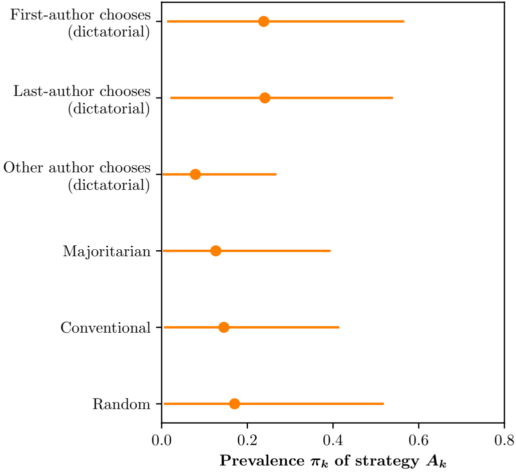
Inferring preference-aggregation mechanisms

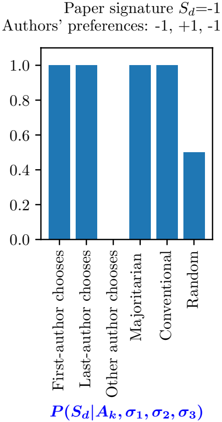
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



Inferring preference-aggregation mechanisms

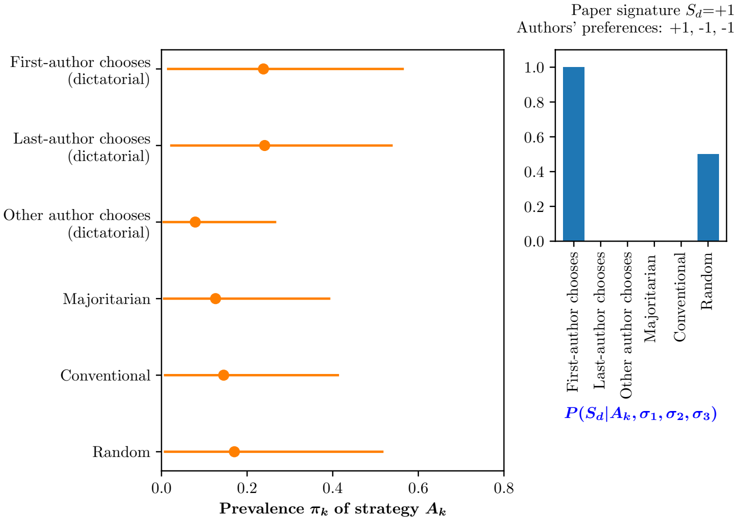
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





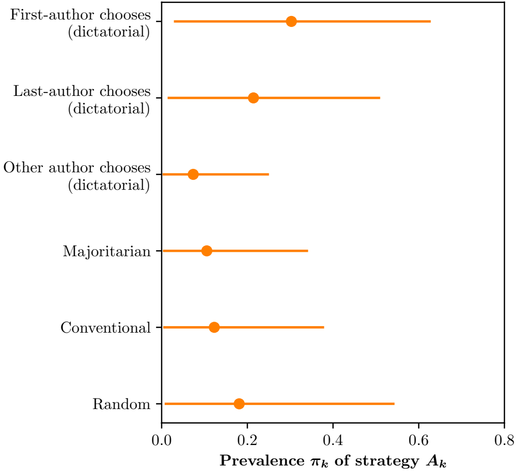
Inferring preference-aggregation mechanisms

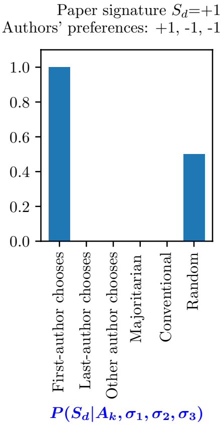
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



Inferring preference-aggregation mechanisms

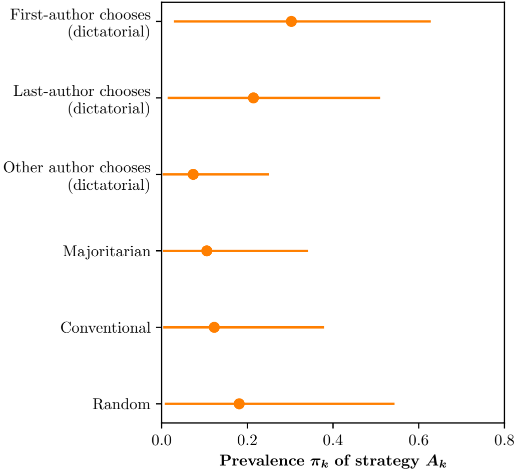
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

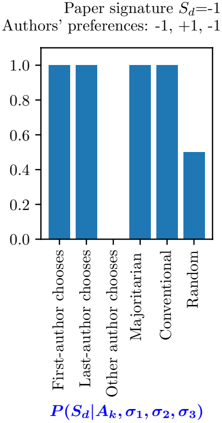




Inferring preference-aggregation mechanisms

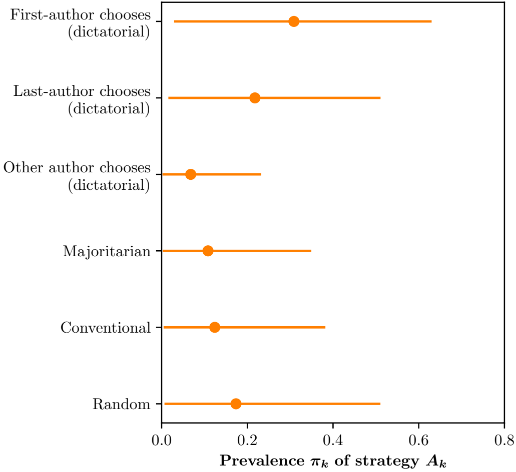
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

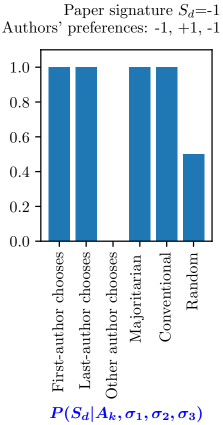




Inferring preference-aggregation mechanisms

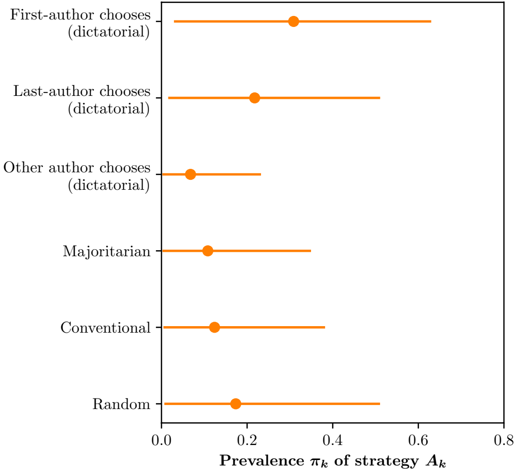
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

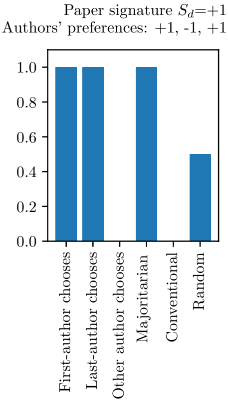




Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

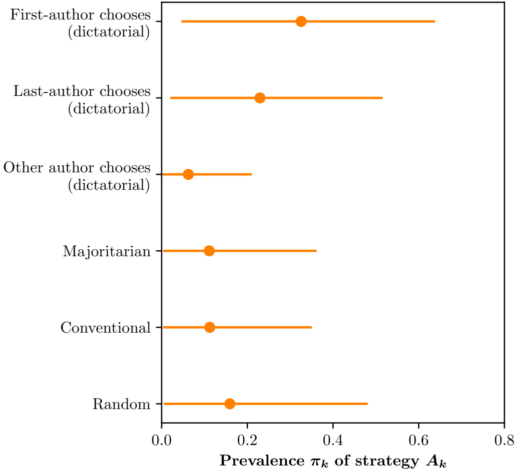


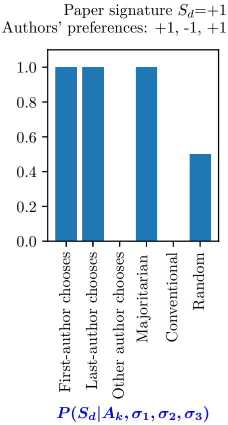


$P(S_d|A_k, \sigma_1, \sigma_2, \sigma_3)$

Inferring preference-aggregation mechanisms

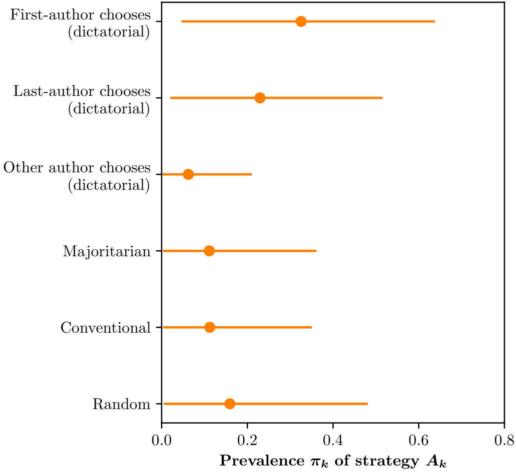
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

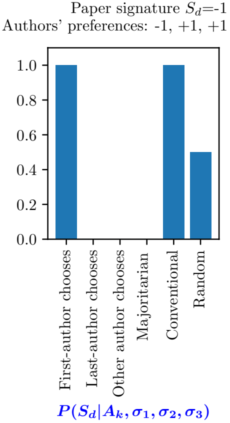




Inferring preference-aggregation mechanisms

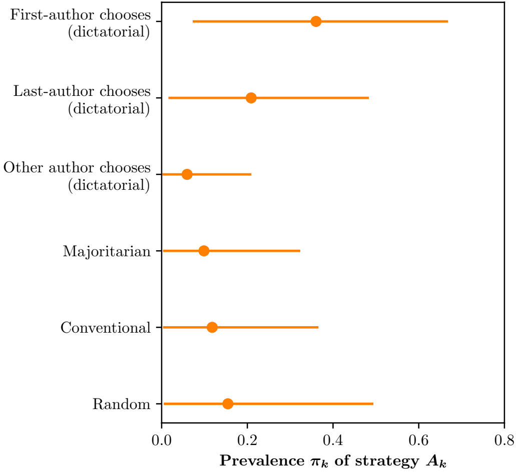
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

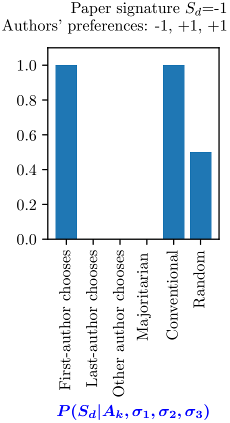




Inferring preference-aggregation mechanisms

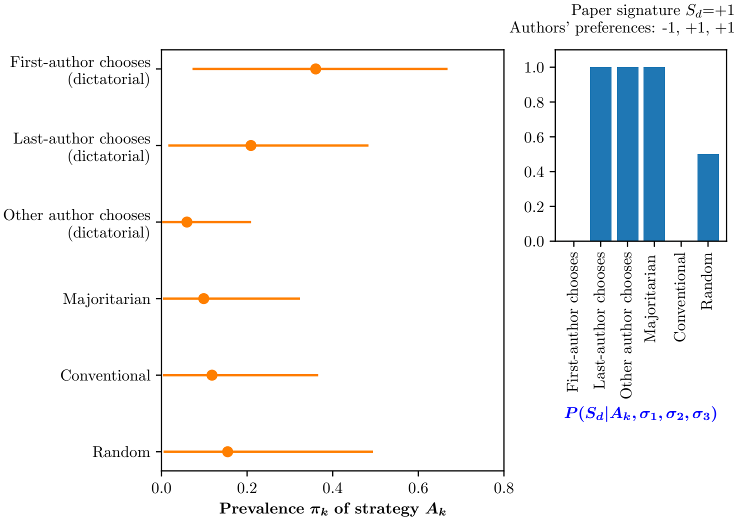
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





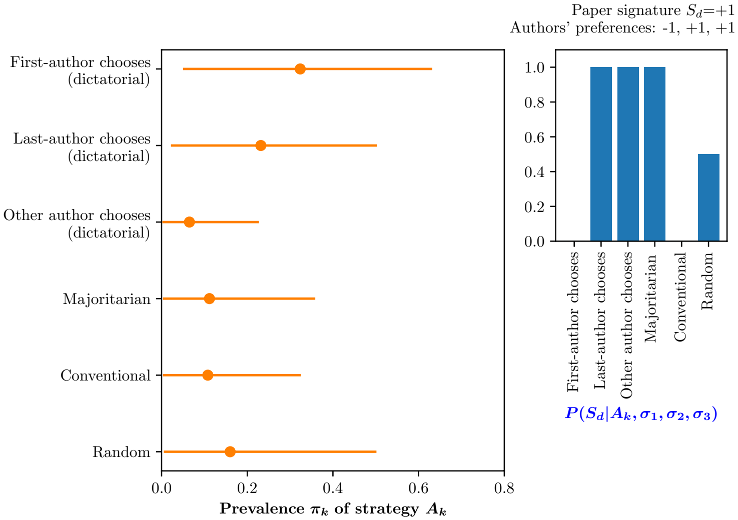
Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



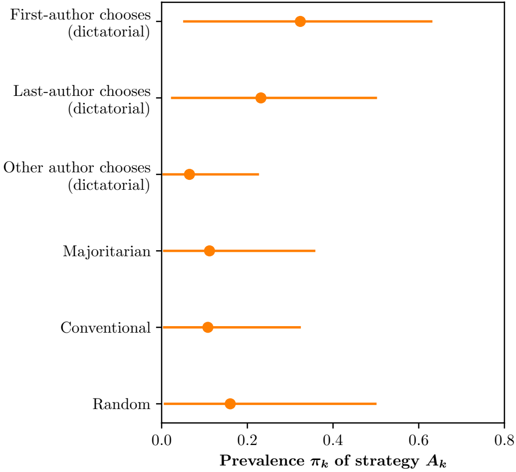
Inferring preference-aggregation mechanisms

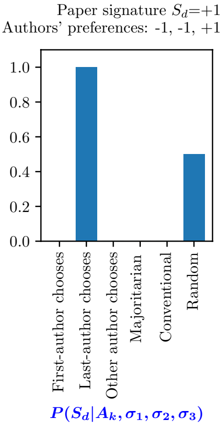
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



Inferring preference-aggregation mechanisms

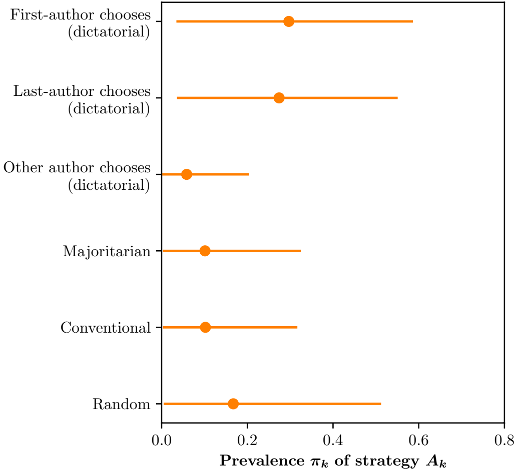
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

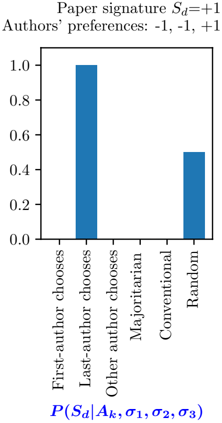




Inferring preference-aggregation mechanisms

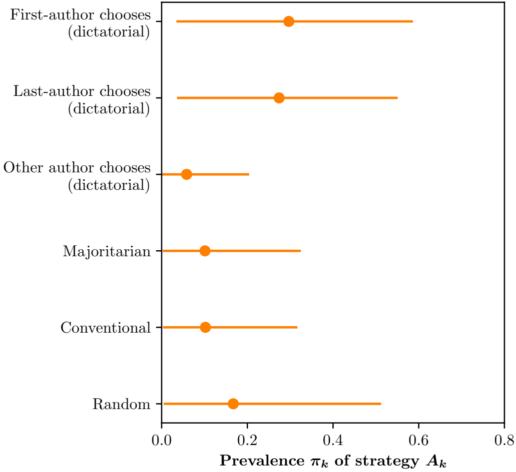
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

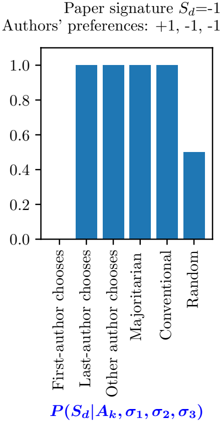




Inferring preference-aggregation mechanisms

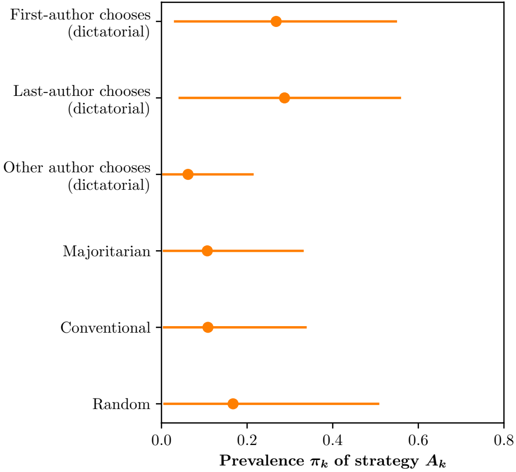
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

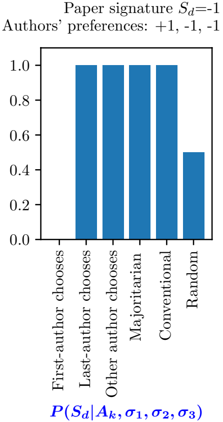




Inferring preference-aggregation mechanisms

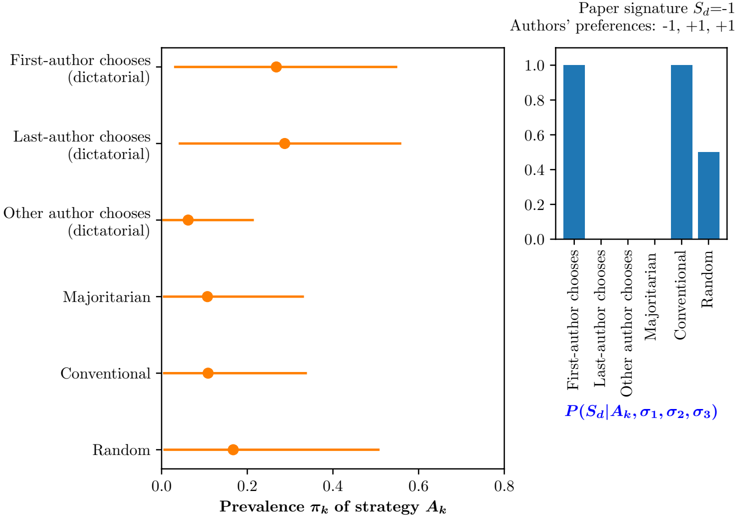
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





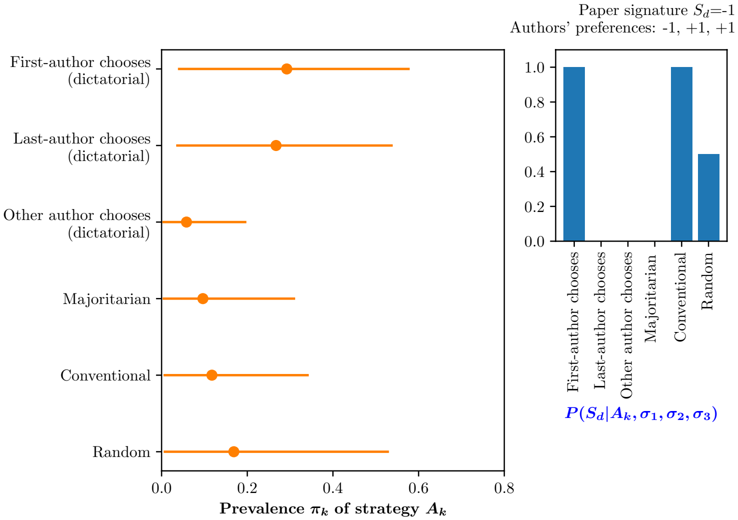
Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



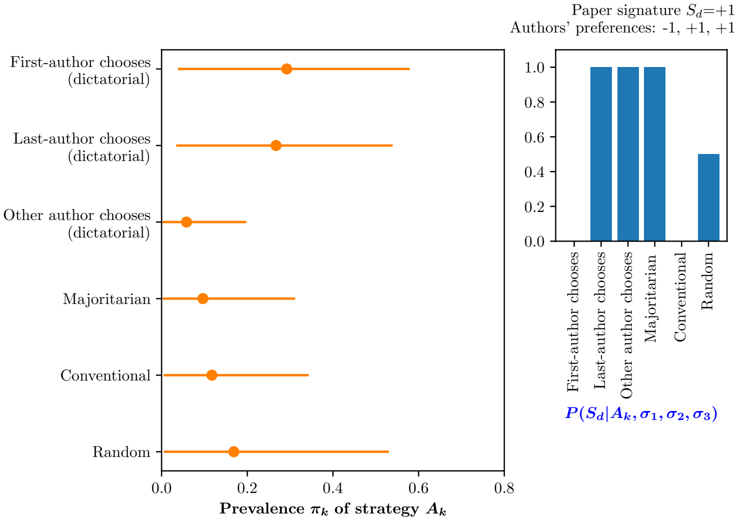
Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



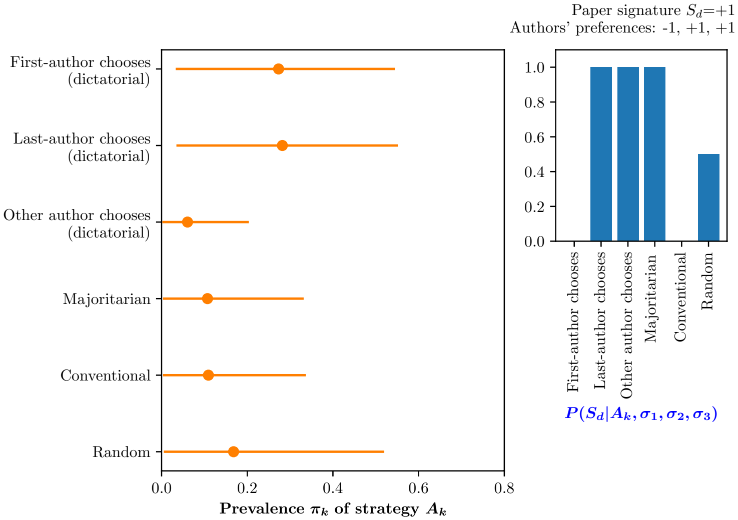
Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



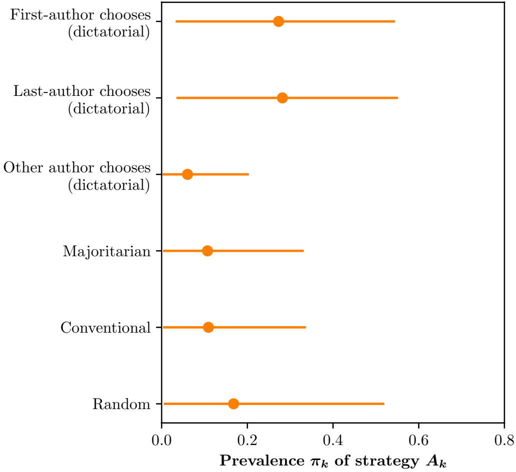
Inferring preference-aggregation mechanisms

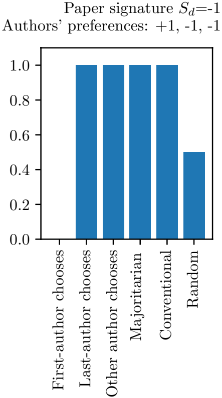
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

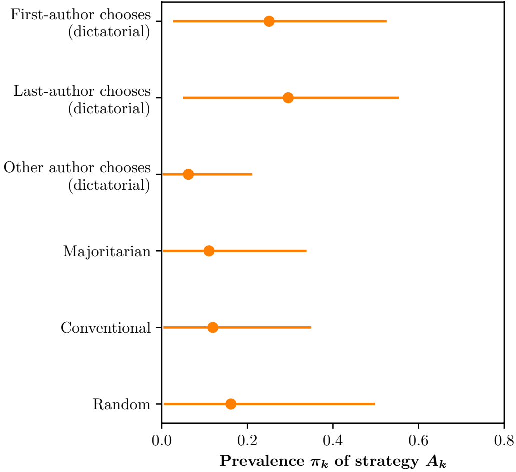


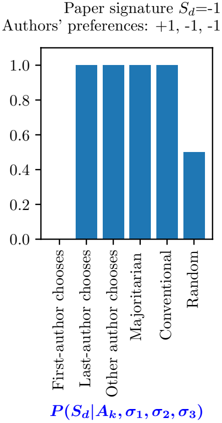


$P(S_d|A_k, \sigma_1, \sigma_2, \sigma_3)$

Inferring preference-aggregation mechanisms

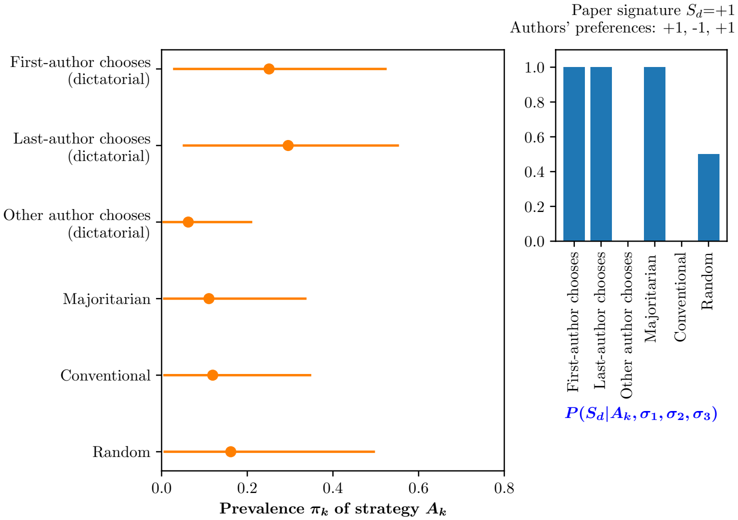
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





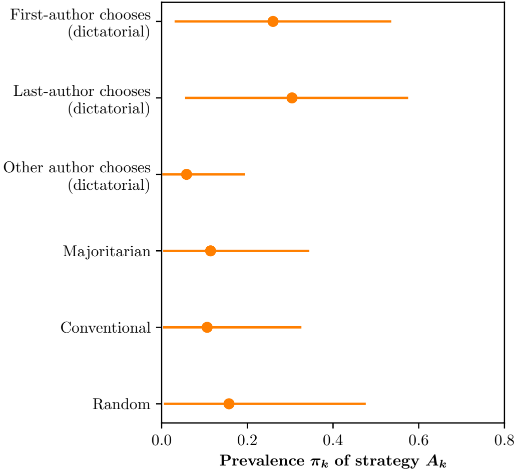
Inferring preference-aggregation mechanisms

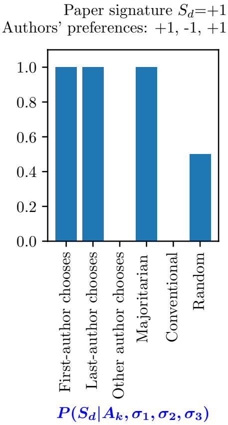
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



Inferring preference-aggregation mechanisms

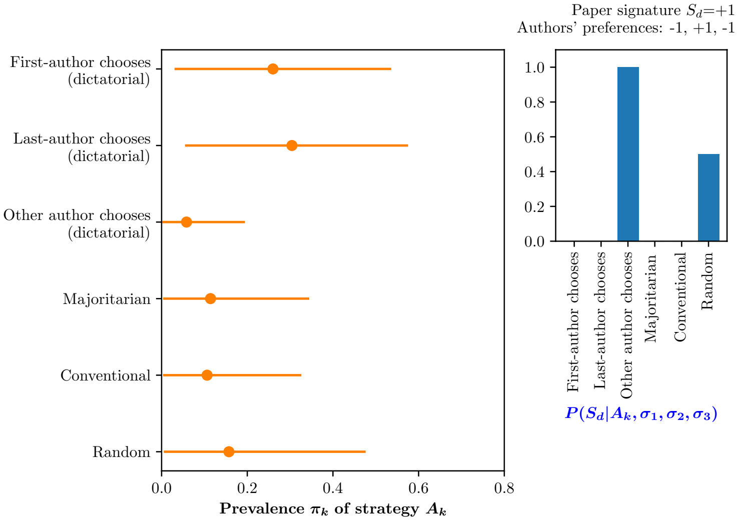
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





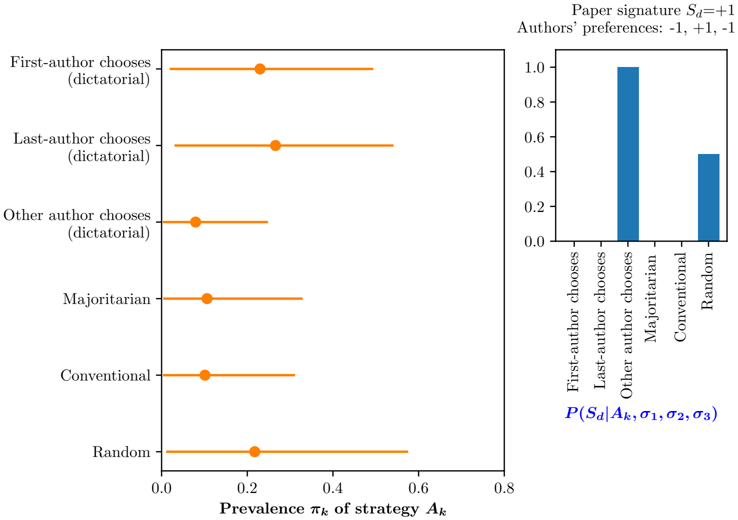
Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



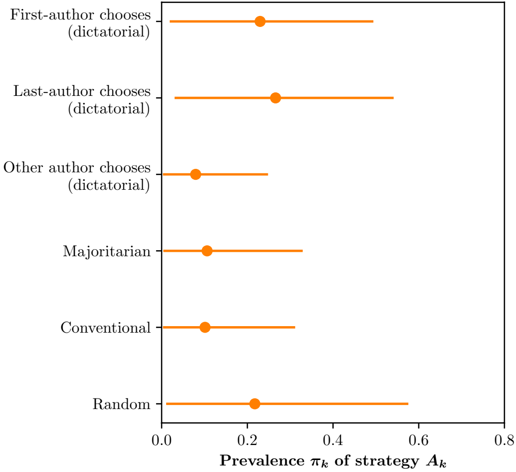
Inferring preference-aggregation mechanisms

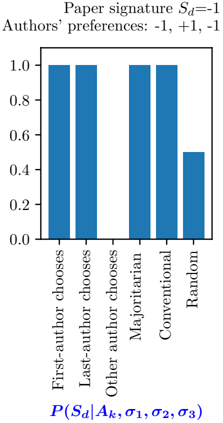
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



Inferring preference-aggregation mechanisms

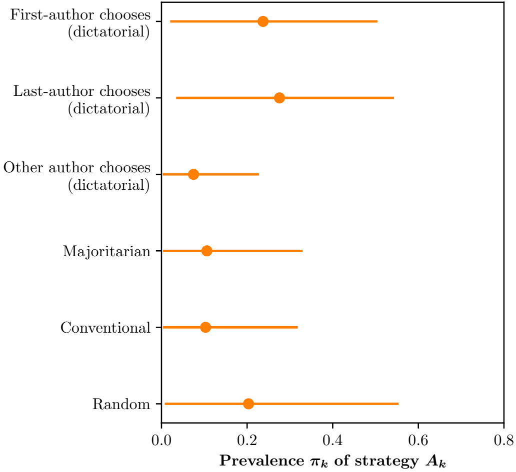
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

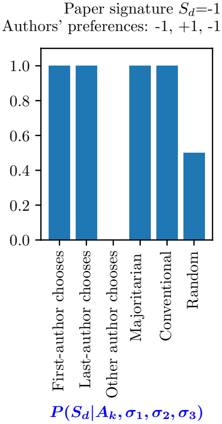




Inferring preference-aggregation mechanisms

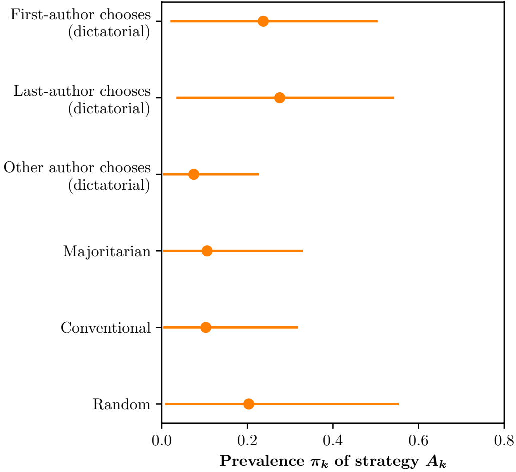
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

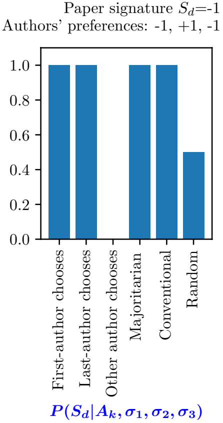




Inferring preference-aggregation mechanisms

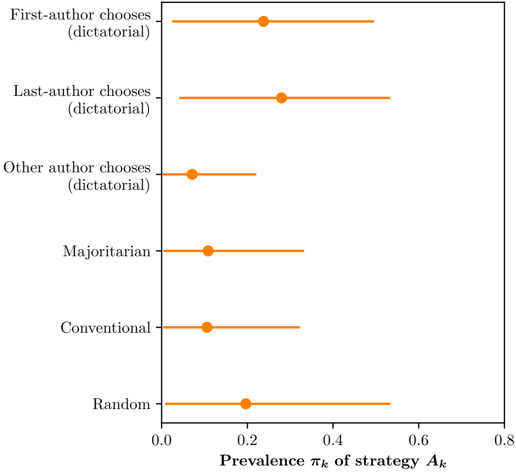
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

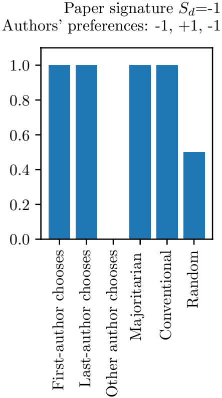




Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

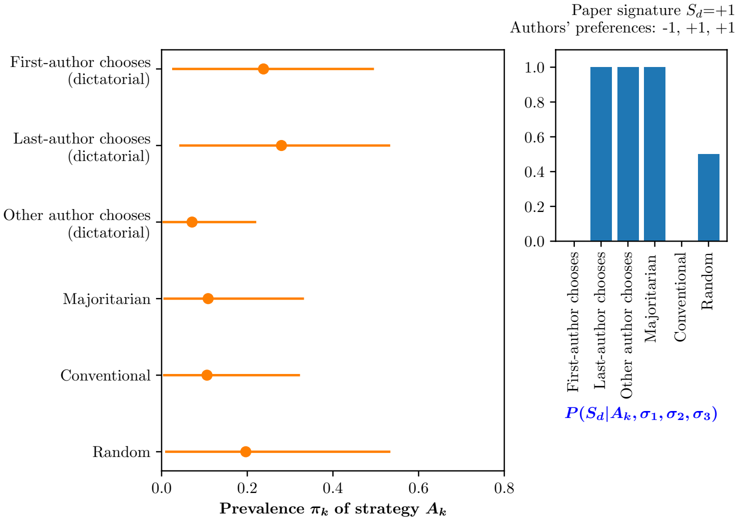




$P(S_d|A_k, \sigma_1, \sigma_2, \sigma_3)$

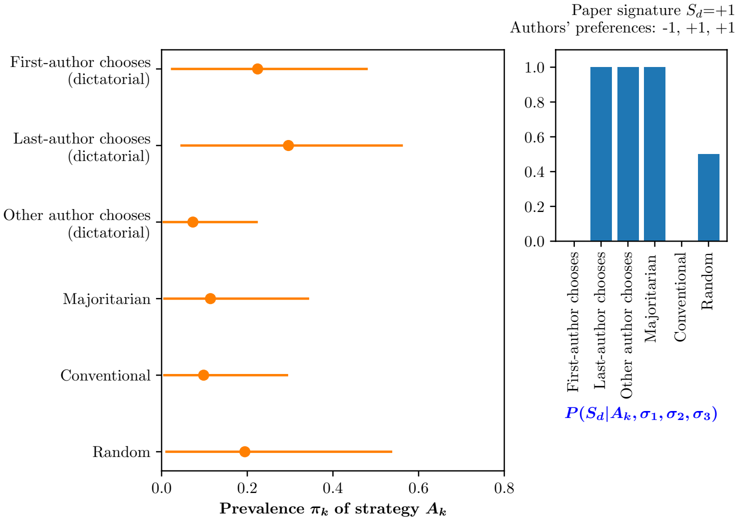
Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



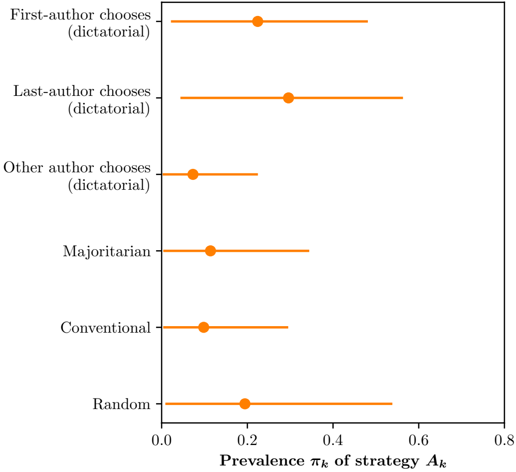
Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

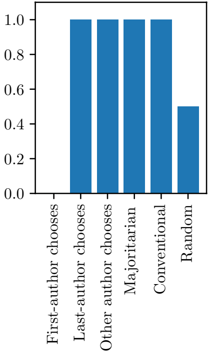


Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



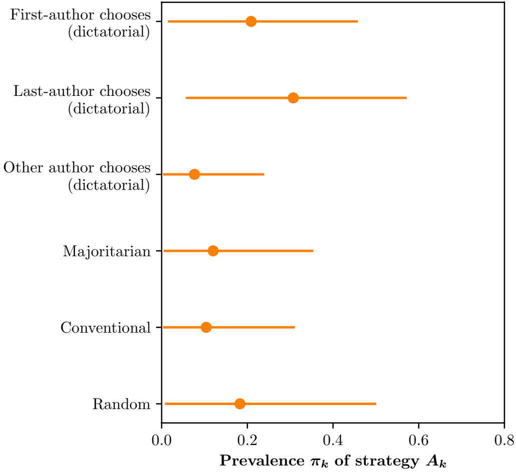
Paper signature S_d=+1
Authors' preferences: -1, +1, +1

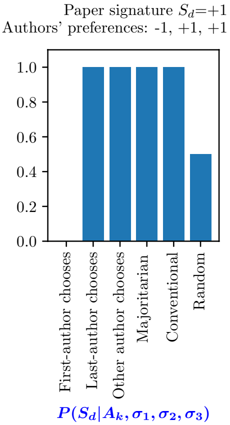


$P(S_d|A_k, \sigma_1, \sigma_2, \sigma_3)$

Inferring preference-aggregation mechanisms

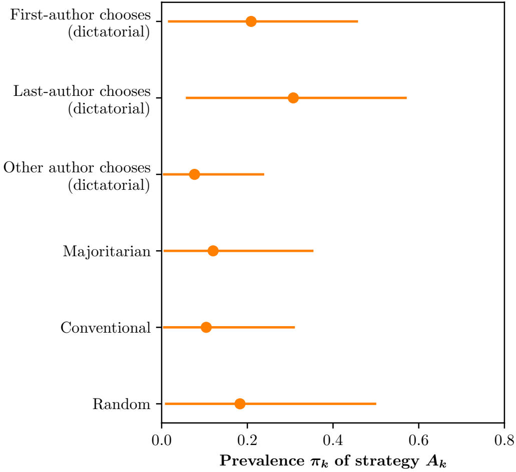
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

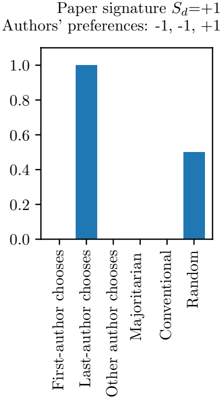




Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.

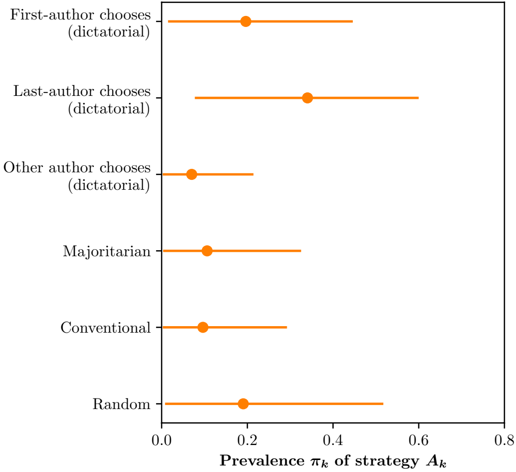


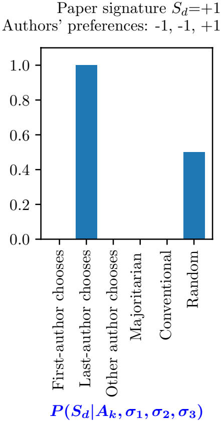


$P(S_d | A_k, \sigma_1, \sigma_2, \sigma_3)$

Inferring preference-aggregation mechanisms

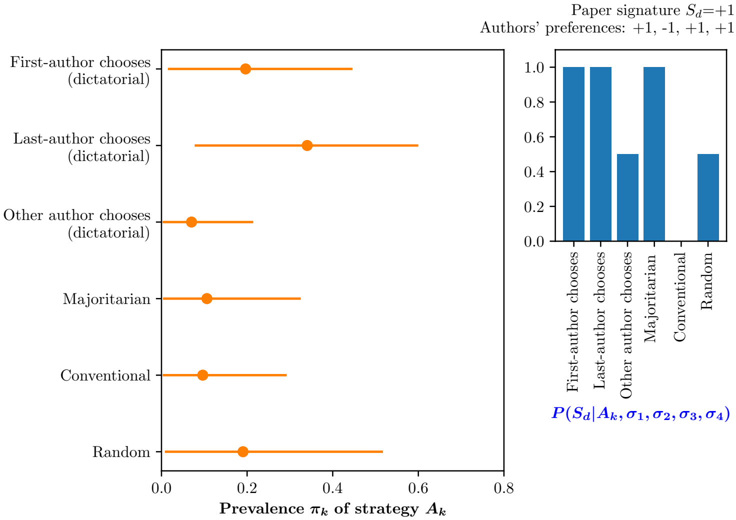
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





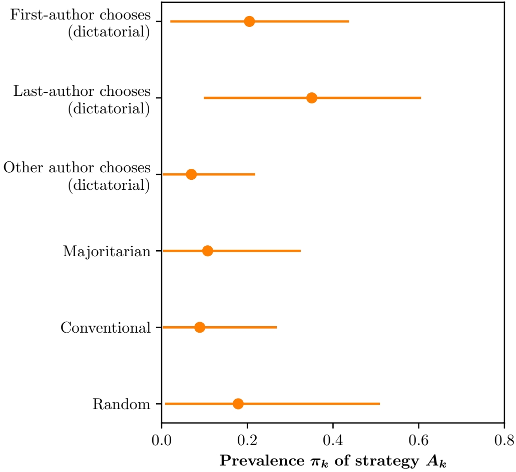
Inferring preference-aggregation mechanisms

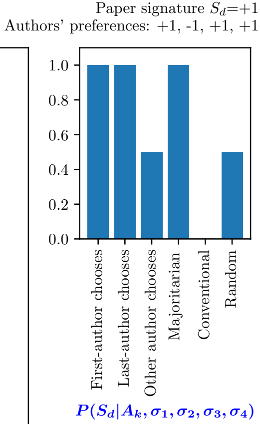
Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.



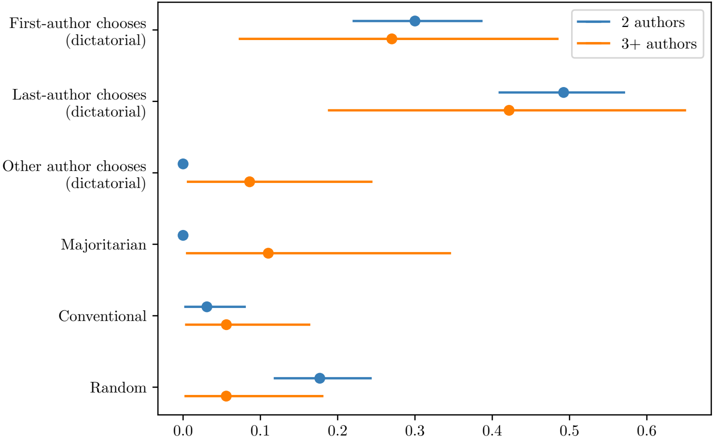
Inferring preference-aggregation mechanisms

Each paper brings a bit more information about π_k, the prevalence of an aggregation strategy A_k.





Prevalence of each preference-aggregation strategy



1. Inverse problems for philosophers and agent-based modelers

2. A case-study of conventions: the metric signature in particle physics
   - How do physicists choose which convention to use in their own papers?
   - How do scientists resolve conflicting preferences in collaborations?
   - How do physicists' preferences get formed?

Authors' preferences ($n = 2\,277$)

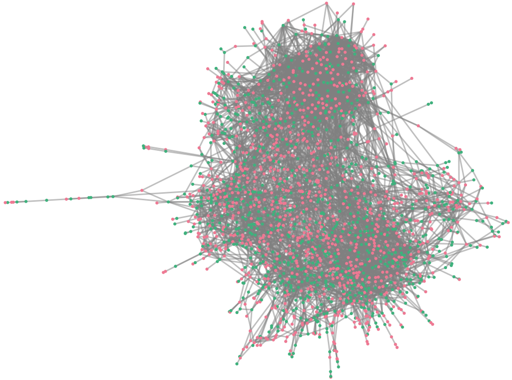
How do physicists' preferences get formed?

- Let's assume three models of the formation of physicists' preference towards the convention:
  1. A **"strategic agent" model** (M_1) assuming that individuals navigate three costs (coordination costs, inconsistency costs, and maladaptation costs) depending on their collaborators' preferences and the research areas in which they publish.
  2. A **global cultural transmission model** (M_2), in which physicists settle once and for all for a specific convention with a certain probability that depends on their primary research area (textbooks?)
  3. A **local cultural transmission model** (M_3), in which physicists copy the preference of their first collaborator.
- Which of these is more plausible given the observed patterns of preferences?

Example: the strategic agent model

The model has multiple unknown parameters:

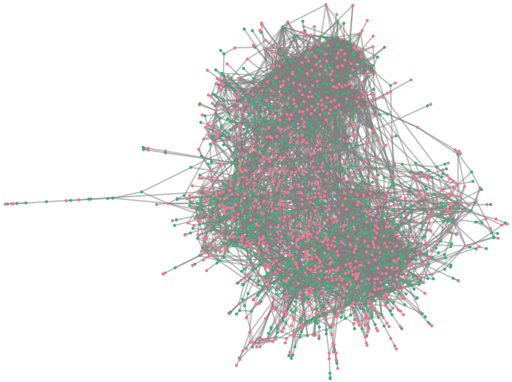
- c_s: the cost of switching from one convention to another
- c_c: the cost of disagreeing with co-authors
- c_r the cost of using a suboptimal convention in a given research area



Example: the strategic agent model

The model has multiple unknown parameters:

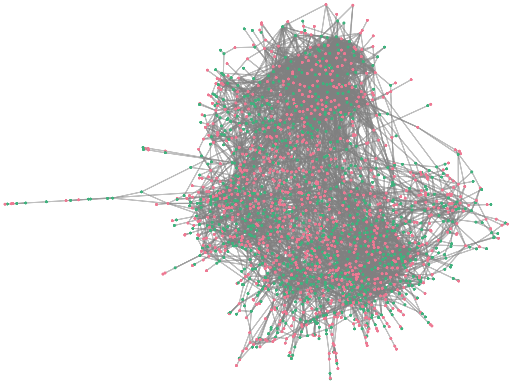
- c_s: the cost of switching from one convention to another
- c_c: the cost of disagreeing with co-authors
- c_r the cost of using a suboptimal convention in a given research area



Example: the strategic agent model

The model has multiple unknown parameters:

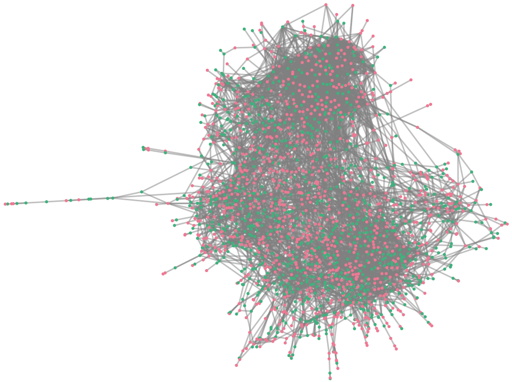
- c_s: the cost of switching from one convention to another
- c_c: the cost of disagreeing with co-authors
- c_r the cost of using a suboptimal convention in a given research area



Example: the strategic agent model

The model has multiple unknown parameters:

- c_s: the cost of switching from one convention to another
- c_c: the cost of disagreeing with co-authors
- c_r the cost of using a suboptimal convention in a given research area



Simulation-based inference

$$P(M_1|O) = \frac{\overbrace{P(O|M_1)}\,P(M_1)}{P(O)} \tag{5}$$

Simulation-based inference

$$P(M_1|O) = \frac{\overbrace{P(O|M_1)}^{\substack{\text{Unknown}\\ \text{in ABMs!}}} P(M_1)}{P(O)} \quad (5)$$

Simulation-based inference

$$P(M_1|O) = \frac{\overbrace{P(O|M_1)}^{\text{Unknown in ABMs!}} P(M_1)}{P(O)} \tag{5}$$

Let us draw N configurations O_s, each of them assuming a model $M_s \in \{1, 2, 3\}$. Then, $P(O|M_1)$ is approximated by the fraction of draws from model M_1 that match the data:

$$P(O|M_1) = \lim_{N\to\infty} \frac{\sum_{s=1}^{N} \mathbb{1}(O_s = O, M_s = 1)}{\sum_{s=1}^{N} \mathbb{1}(M_s = 1)} \tag{6}$$

Simulation-based inference

$$P(M_1|O) = \frac{\overbrace{P(O|M_1)}^{\text{Unknown in ABMs!}} P(M_1)}{P(O)} \tag{5}$$

Let us draw N configurations O_s, each of them assuming a model $M_s \in \{1, 2, 3\}$. Then, $P(O|M_1)$ is approximated by the fraction of draws from model M_1 that match the data:

$$P(O|M_1) = \lim_{N\to\infty} \frac{\sum_{s=1}^{N} \mathbb{1}(O_s = O, M_s = 1)}{\sum_{s=1}^{N} \mathbb{1}(M_s = 1)} \tag{6}$$

$$P(M_1|O) = \frac{P(O|M_1)P(M_1)}{P(O)} \tag{7}$$

$$P(M_1|O) = \frac{\overbrace{P(O|M_1)}^{\text{Unknown in ABMs!}} P(M_1)}{P(O)} \tag{5}$$

Let us draw N configurations O_s, each of them assuming a model $M_s \in \{1, 2, 3\}$. Then, $P(O|M_1)$ is approximated by the fraction of draws from model M_1 that match the data:

$$P(O|M_1) = \lim_{N\to\infty} \frac{\sum_{s=1}^{N} \mathbb{1}(O_s = O, M_s = 1)}{\sum_{s=1}^{N} \mathbb{1}(M_s = 1)} \tag{6}$$

$$P(M_1|O) = \frac{P(O|M_1)P(M_1)}{P(O)} = \lim_{N\to\infty} \frac{\sum_{s=1}^{N} \mathbb{1}(O_s = O, M_s = 1)}{\sum_{s=1}^{N} \mathbb{1}(O_s = O)} \tag{7}$$

Curse of dimensionality in simulation-based inference

- This approach requires that at least some draws O_s match the observed outcome O. These draws must match exactly each and everyone's of the 2 277 authors' preferences!

Curse of dimensionality in simulation-based inference

- This approach requires that at least some draws O_s match the observed outcome O. These draws must match exactly each and everyone's of the 2 277 authors' preferences!
- The probability for this to happen is virtually 0 (if a model has a probability $p = 0.95$ to predict the correct preference for any author independently, the joint probability would be 2×10^{-51})

Curse of dimensionality in simulation-based inference

- This approach requires that at least some draws O_s match the observed outcome O. These draws must match exactly each and everyone's of the 2 277 authors' preferences!
- The probability for this to happen is virtually 0 (if a model has a probability $p = 0.95$ to predict the correct preference for any author independently, the joint probability would be 2×10^{-51})
- It would take an absurd amount of time so that at least one simulation matches the observed outcome (many times the age of the universe, even with millions of draws per second)

Curse of dimensionality in simulation-based inference

- This approach requires that at least some draws O_s match the observed outcome O. These draws must match exactly each and everyone's of the 2 277 authors' preferences!
- The probability for this to happen is virtually 0 (if a model has a probability $p = 0.95$ to predict the correct preference for any author independently, the joint probability would be 2×10^{-51})
- It would take an absurd amount of time so that at least one simulation matches the observed outcome (many times the age of the universe, even with millions of draws per second)
- This is because because the data has *too many dimensions* $\Rightarrow$ **"curse of dimensionality"**

Curse of dimensionality in simulation-based inference

- This approach requires that at least some draws O_s match the observed outcome O. These draws must match exactly each and everyone's of the 2 277 authors' preferences!
- The probability for this to happen is virtually 0 (if a model has a probability $p = 0.95$ to predict the correct preference for any author independently, the joint probability would be 2×10^{-51})
- It would take an absurd amount of time so that at least one simulation matches the observed outcome (many times the age of the universe, even with millions of draws per second)
- This is because because the data has *too many dimensions* $\Rightarrow$ **"curse of dimensionality"**
- The solution: "conditioning" on **summary statistics** rather than the entire data.

Curse of dimensionality in simulation-based inference

- This approach requires that at least some draws O_s match the observed outcome O. These draws must match exactly each and everyone's of the 2 277 authors' preferences!
- The probability for this to happen is virtually 0 (if a model has a probability $p = 0.95$ to predict the correct preference for any author independently, the joint probability would be 2×10^{-51})
- It would take an absurd amount of time so that at least one simulation matches the observed outcome (many times the age of the universe, even with millions of draws per second)
- This is because because the data has *too many dimensions* $\Rightarrow$ **"curse of dimensionality"**
- The solution: "conditioning" on **summary statistics** rather than the entire data.
- Summary statistics are **low-dimensional descriptions of the data** that capture their essential features. e.g.:

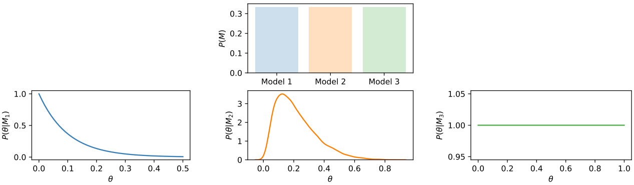
$$m = \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \right| \tag{8}$$

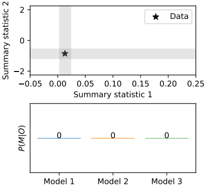
Summary statistics in simulation-based inference

There are two main approaches for choosing adequate summary statistics:

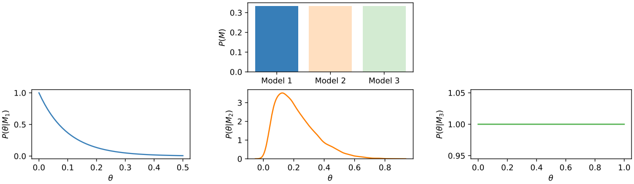
1. Hand-picking interpretable summary statistics based on our own intuitions.
2. Using sophisticated methods to learn statistically optimal (but potentially un-interpretable) summary statistics. Optimal summary statistics reduce our posterior uncertainty given a fixed amount of data.

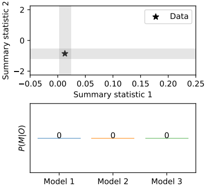
Simulation-based inference with summary statistics



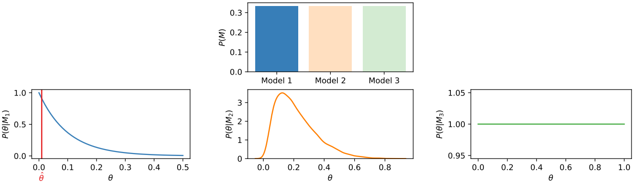


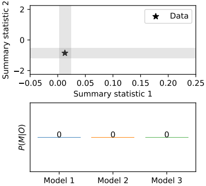
Simulation-based inference with summary statistics



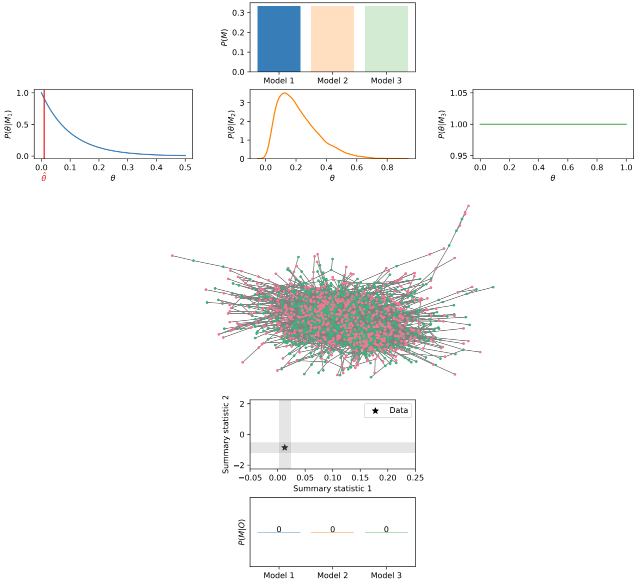


Simulation-based inference with summary statistics

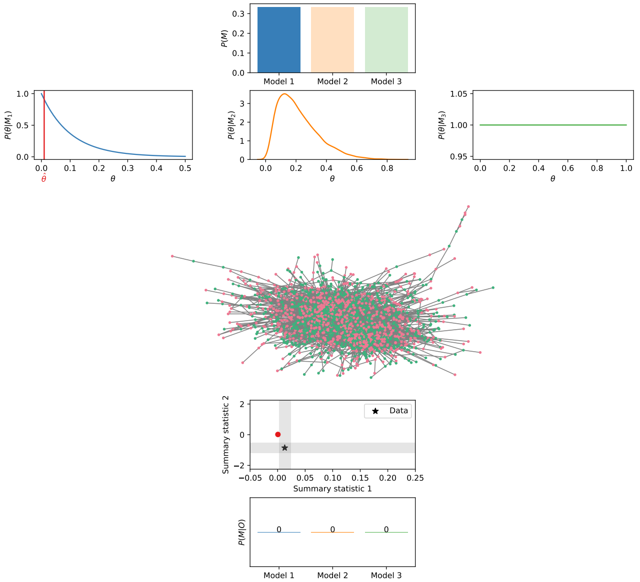




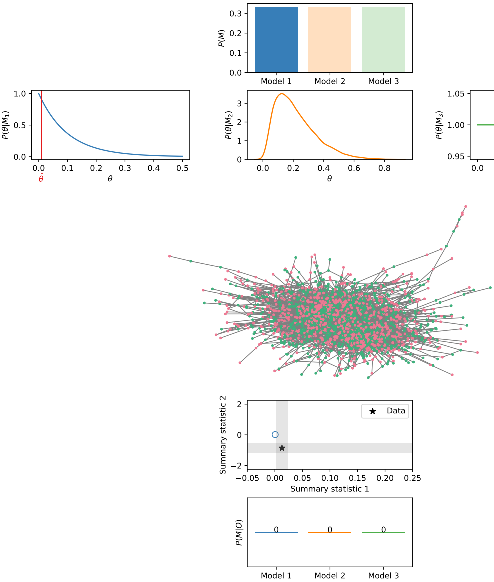
Simulation-based inference with summary statistics



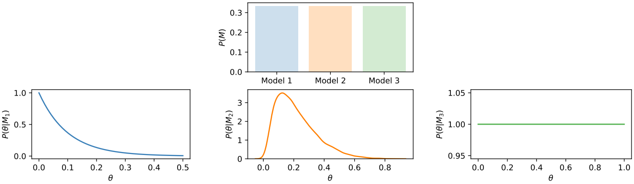
Simulation-based inference with summary statistics

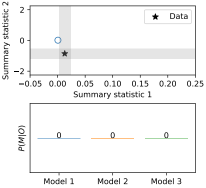


Simulation-based inference with summary statistics

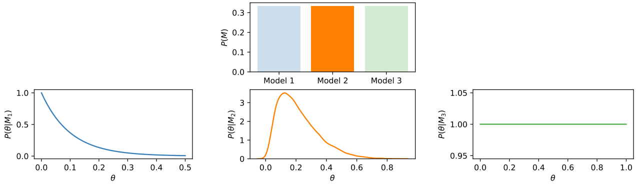


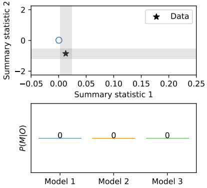
Simulation-based inference with summary statistics



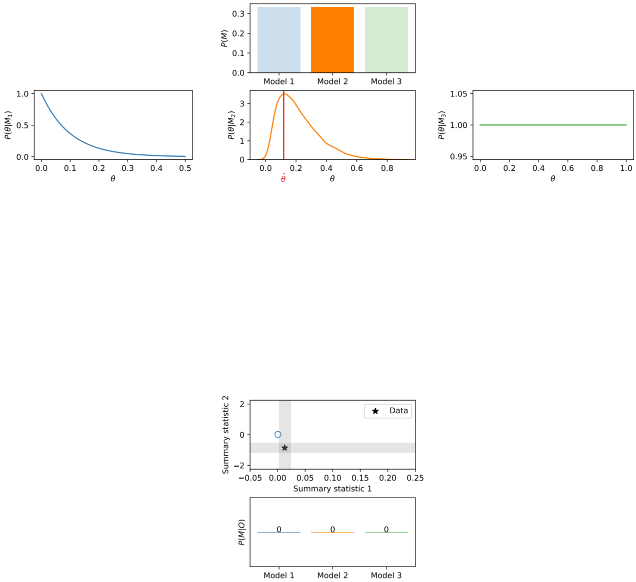


Simulation-based inference with summary statistics

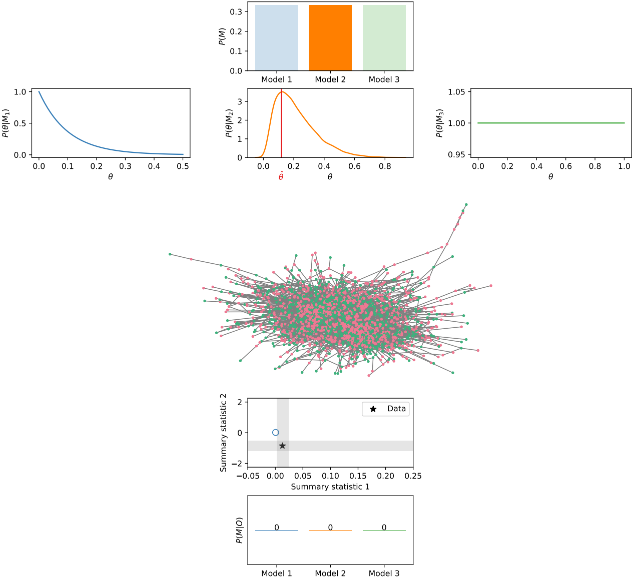




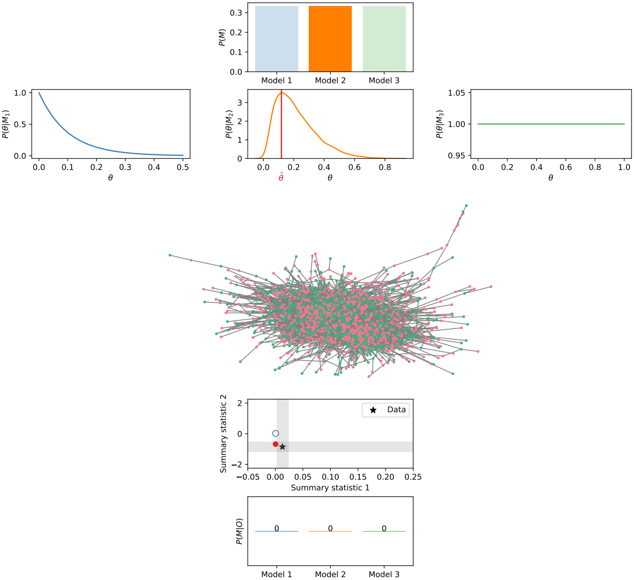
Simulation-based inference with summary statistics



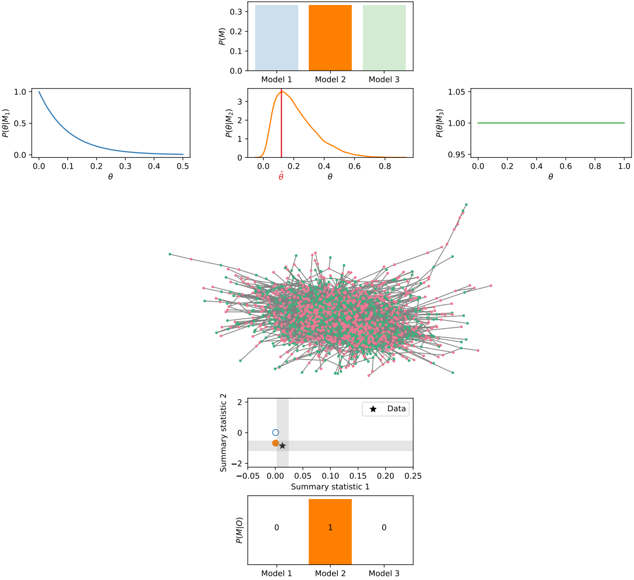
Simulation-based inference with summary statistics



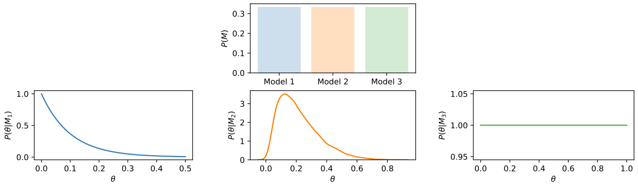
Simulation-based inference with summary statistics

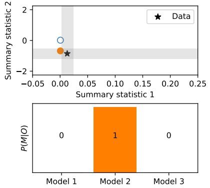


Simulation-based inference with summary statistics

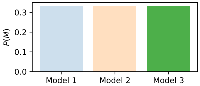


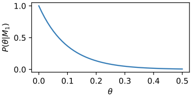
Simulation-based inference with summary statistics

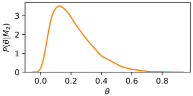


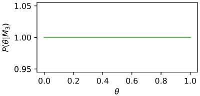


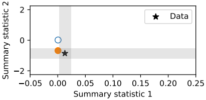
Simulation-based inference with summary statistics

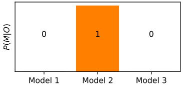




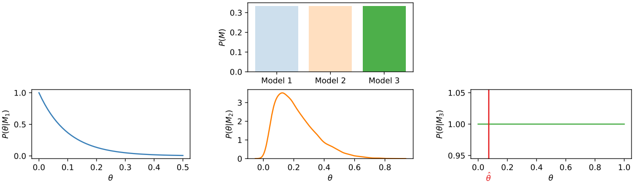


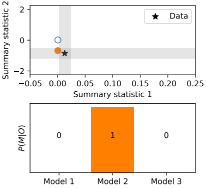




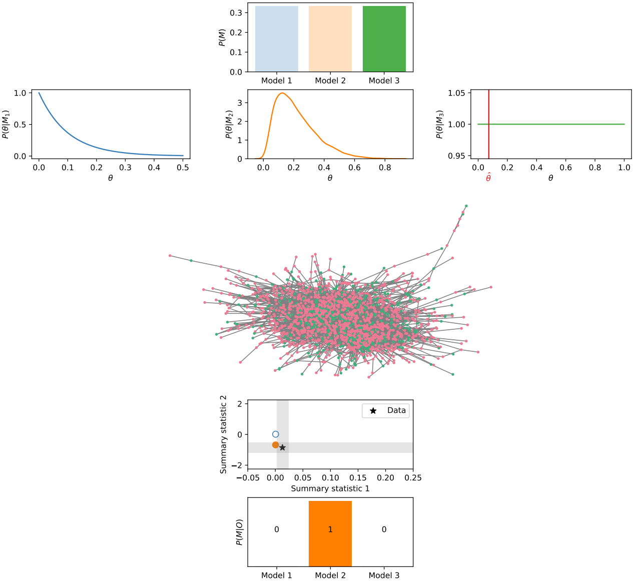


Simulation-based inference with summary statistics

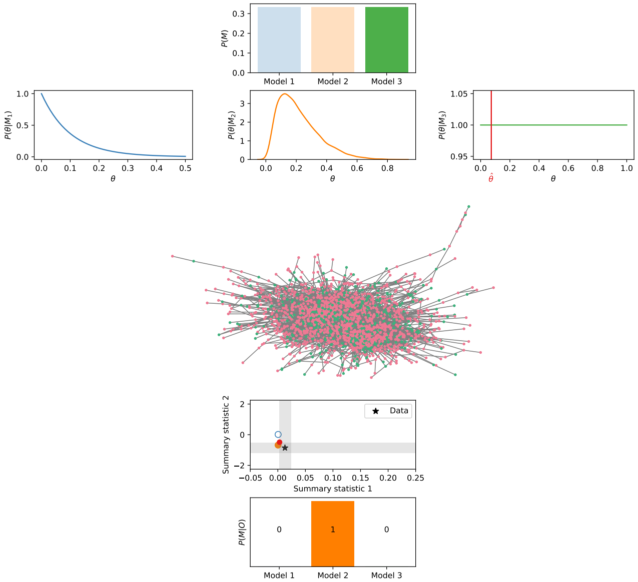




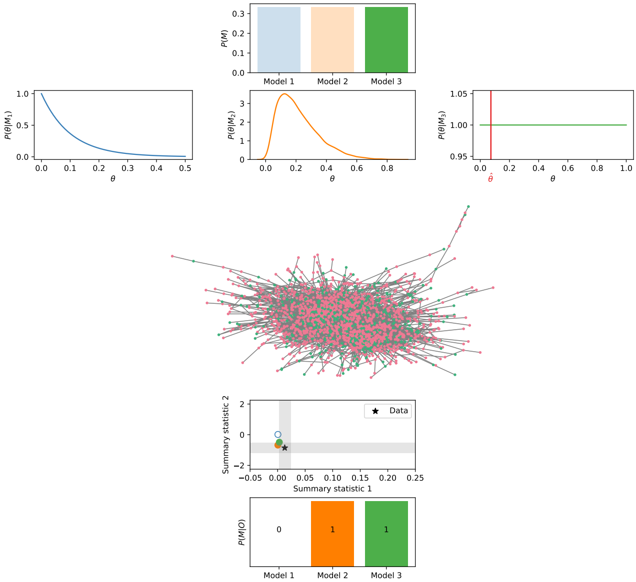
Simulation-based inference with summary statistics



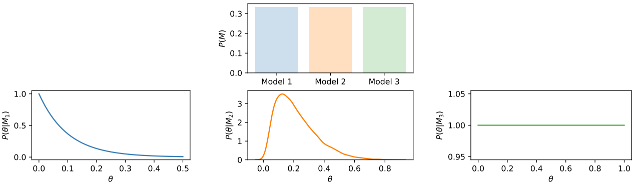
Simulation-based inference with summary statistics

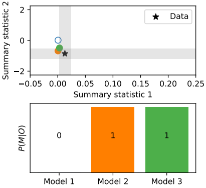


Simulation-based inference with summary statistics

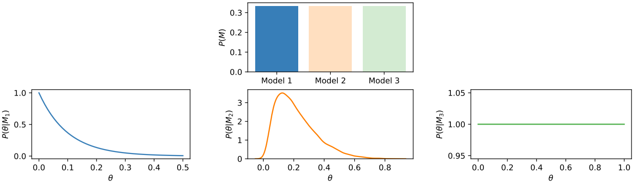


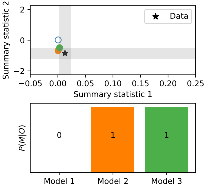
Simulation-based inference with summary statistics



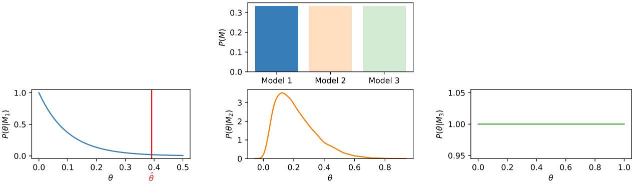


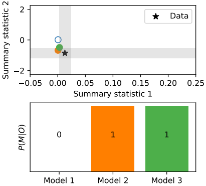
Simulation-based inference with summary statistics



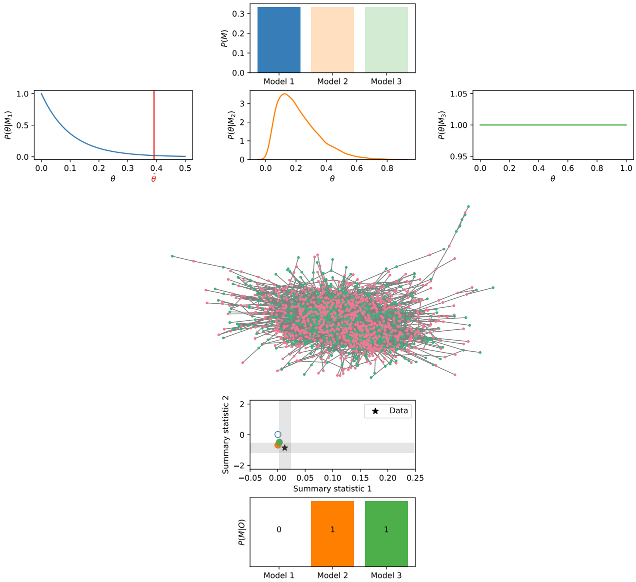


Simulation-based inference with summary statistics

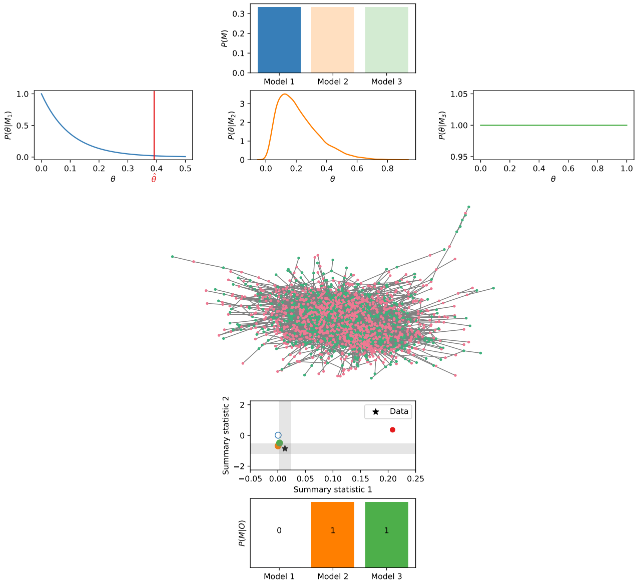




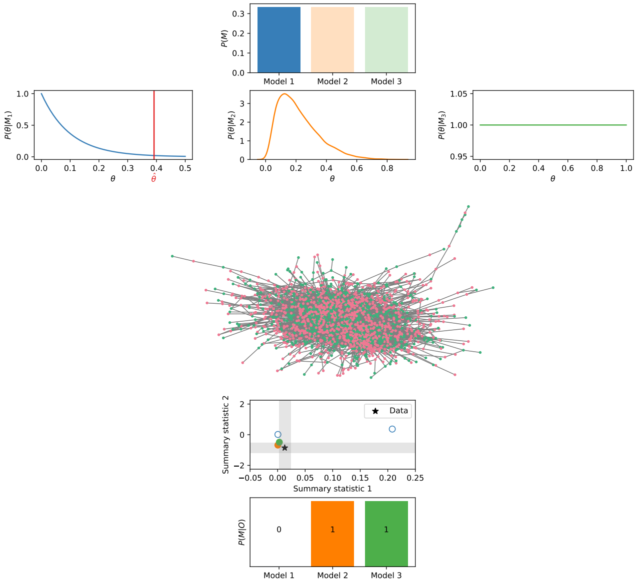
Simulation-based inference with summary statistics



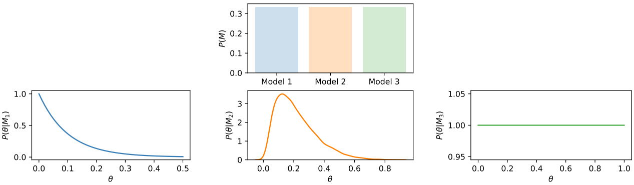
Simulation-based inference with summary statistics

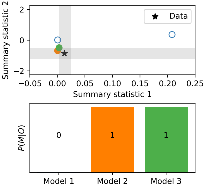


Simulation-based inference with summary statistics

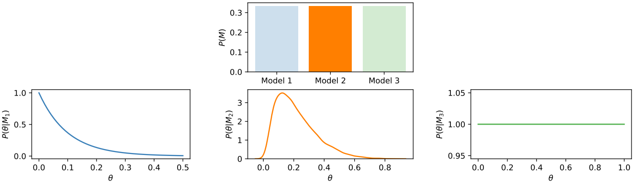


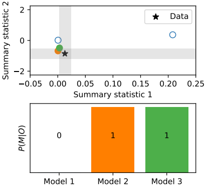
Simulation-based inference with summary statistics



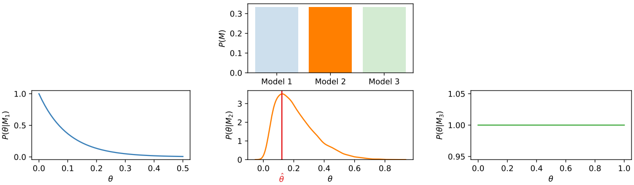


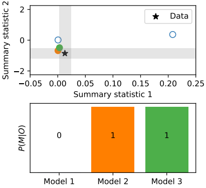
Simulation-based inference with summary statistics



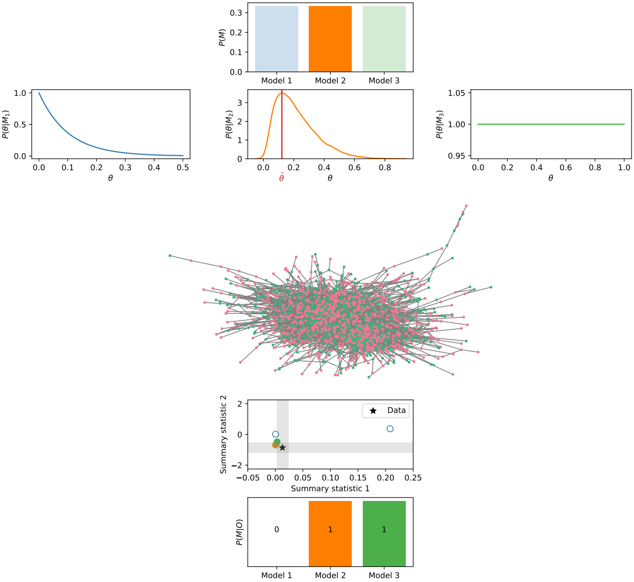


Simulation-based inference with summary statistics

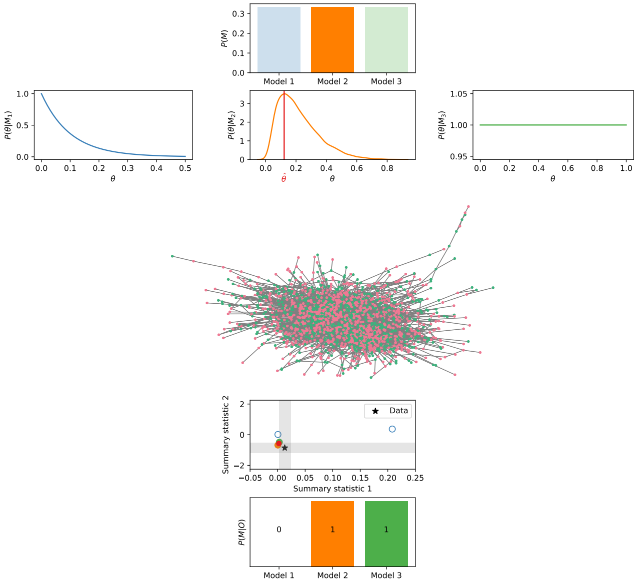




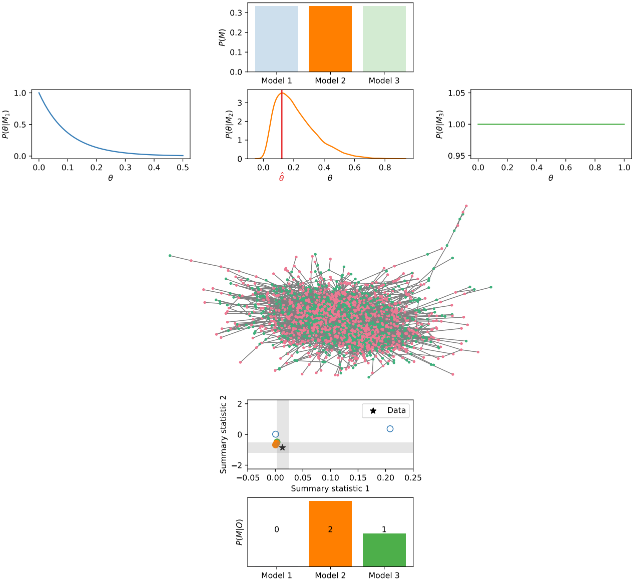
Simulation-based inference with summary statistics



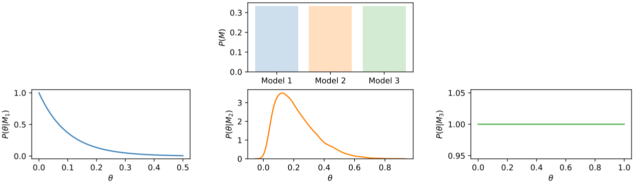
Simulation-based inference with summary statistics

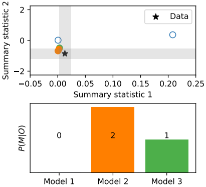


Simulation-based inference with summary statistics

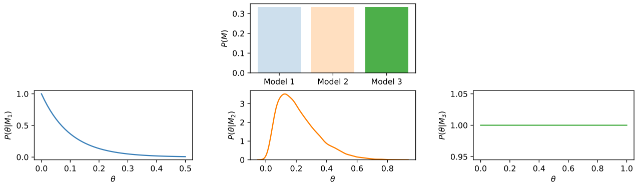


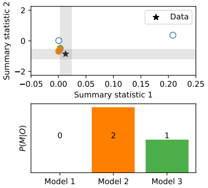
Simulation-based inference with summary statistics



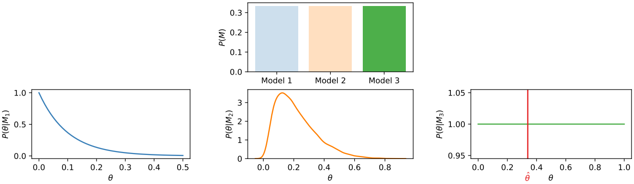


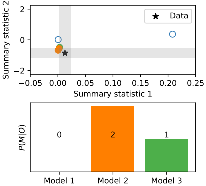
Simulation-based inference with summary statistics



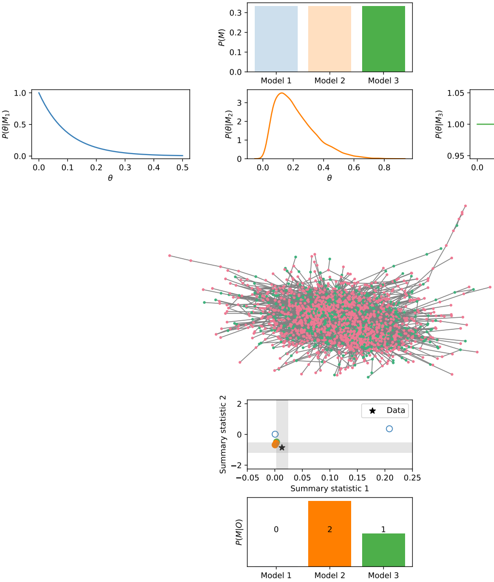


Simulation-based inference with summary statistics

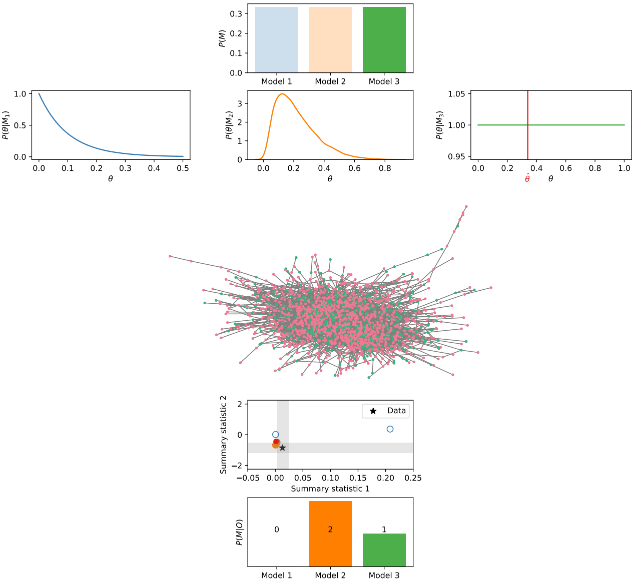




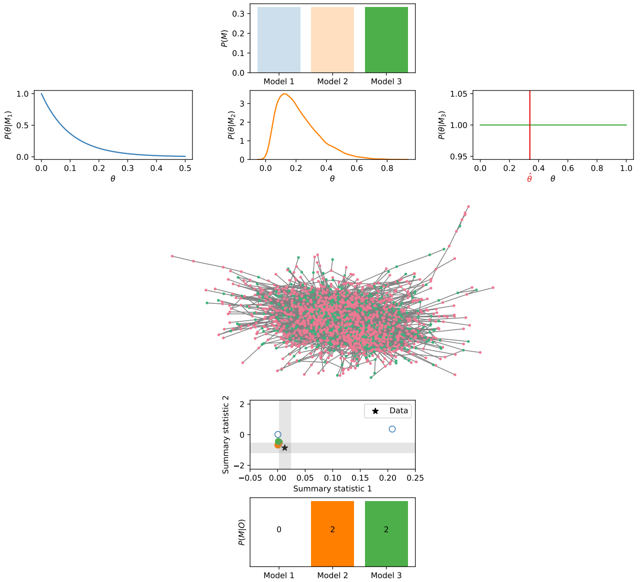
Simulation-based inference with summary statistics



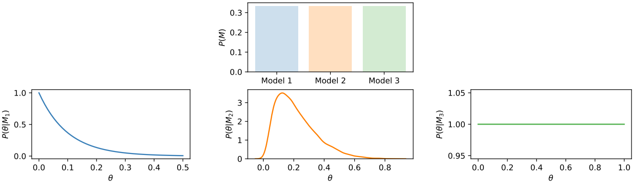
Simulation-based inference with summary statistics

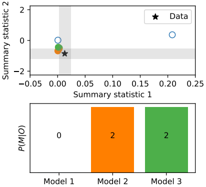


Simulation-based inference with summary statistics

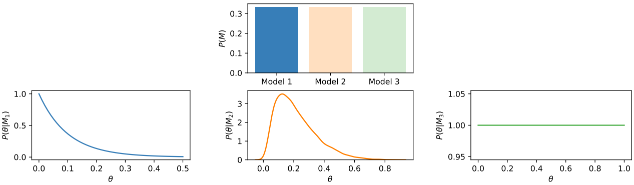


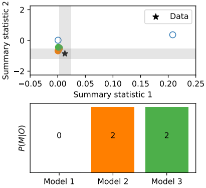
Simulation-based inference with summary statistics



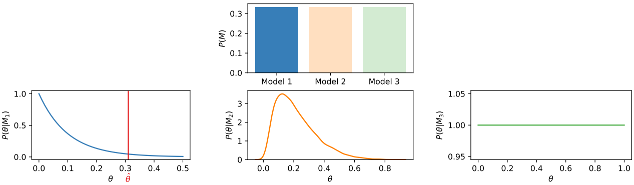


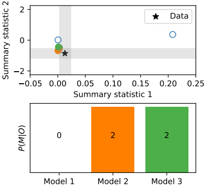
Simulation-based inference with summary statistics



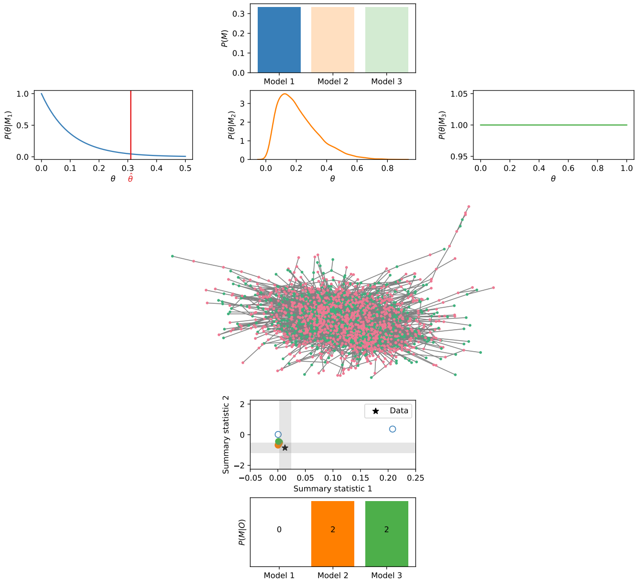


Simulation-based inference with summary statistics

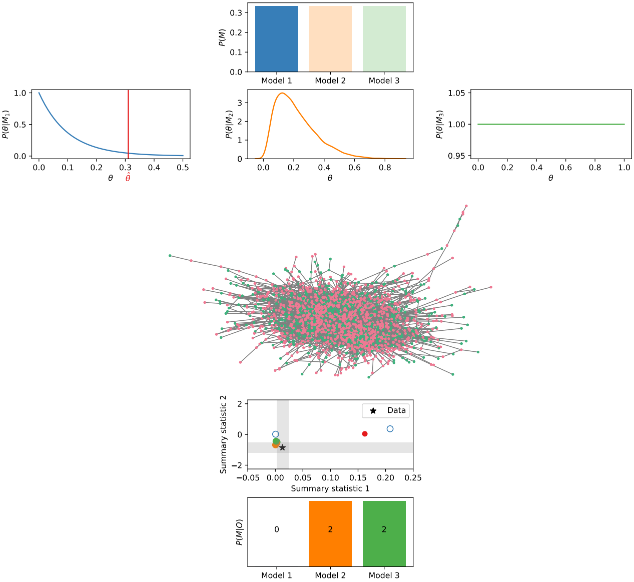




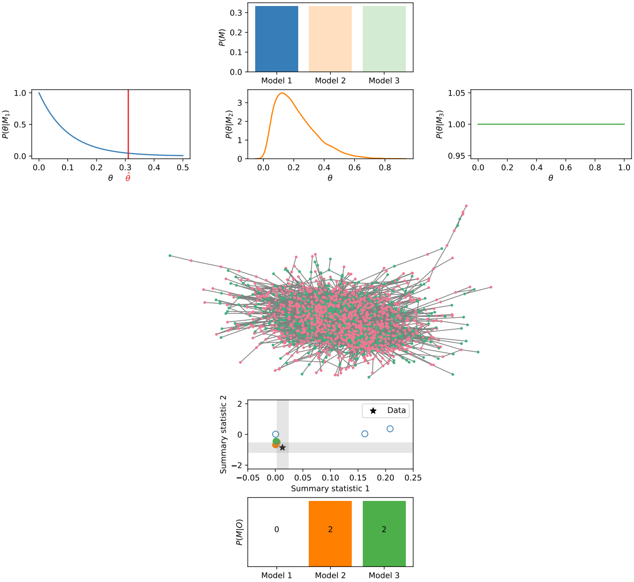
Simulation-based inference with summary statistics



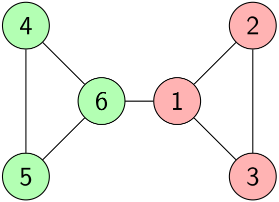
Simulation-based inference with summary statistics



Simulation-based inference with summary statistics



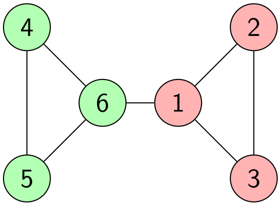
Local versus global mechanisms of coordination

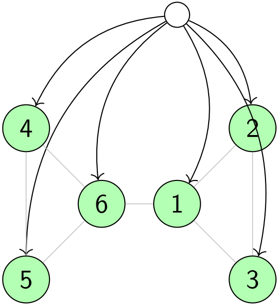


Local coordination

Strategic alignment,
imitation of peers. . .
$\boldsymbol{J}$

Local versus global mechanisms of coordination





Local coordination

Strategic alignment,
imitation of peers. . .
J

Global coordination

Adaptation to research purposes,
or shared culture ("disciplinary matrix")
B

The Ising model as an intermediate idealized model

- Atomic magnetic spins in a material can be in two states: $\uparrow$ $(+1)$ or $\downarrow$ (-1).
- Magnetic spins prefer to be aligned to their neighbors ($\uparrow\uparrow$ or $\downarrow\downarrow$)
- Can local interactions between spins at the microscopic level lead to macroscopic alignment?

$$P(\{\sigma_i\}|J,\boldsymbol{B}) = \frac{1}{Z(J,\boldsymbol{B})} e^{-H(\{\sigma_i\},J,\boldsymbol{B})}, \text{ and } H = - \underbrace{\sum_{i,j} J w_{ij} \sigma_i \sigma_j}_{\substack{\text{local}\\ \text{pairwise interactions}}} \underbrace{- \sum_i B_{C_i} \sigma_i}_{\substack{\text{external}\\ \text{magnetic field}}} \tag{9}$$

`https://mattbierbaum.github.io/ising.js/`

Inverse Ising problem: $P(J, J^{\text{cit}}, \boldsymbol{B}|\{\sigma_i\})$

Local coordination in multi-layered graphs

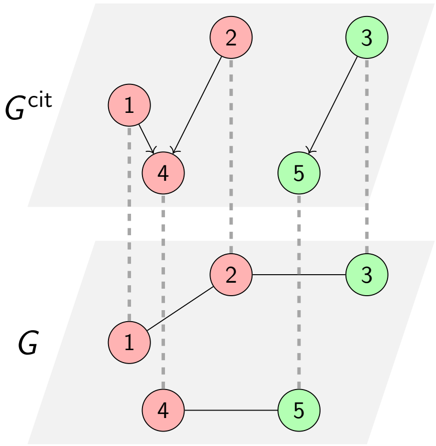


Figure: **Illustration of local coordination in multilayered social networks**. Nodes can be connected through different kinds of relationships (for instance, authors can be related via collaborations (G) or citations (G^{cit})).

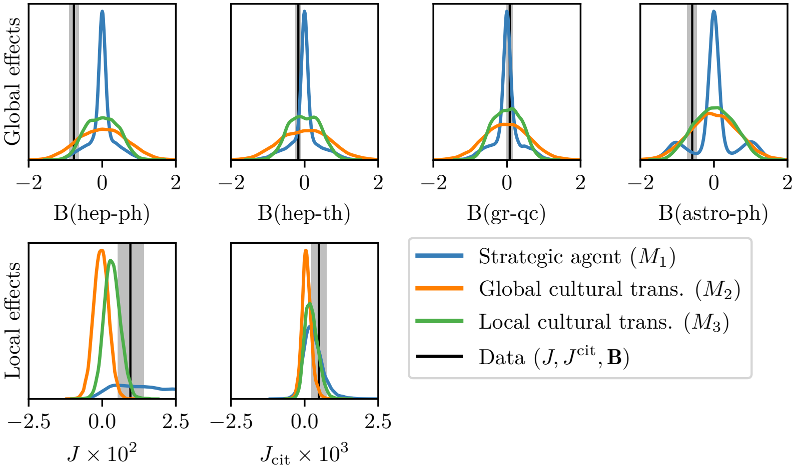
Local versus global coordination

Table: Parameters of the Ising model.

Parameter	Effect size	$\mathrm{CI}_{95\%}$	Effect size	$\mathrm{CI}_{95\%}$
J	$+0.013$	[$+0.009$, $+0.017$]	$+0.0095$	[$+0.0052$, $+0.014$]
J^{cit}	-	-	$+0.00049$	[$+0.00023$, $+0.00075$]
$B(\mathrm{hep-ph})$	-0.86	[-0.99, -0.73]	-0.77	[-0.91, -0.64]
$B(\mathrm{hep-th})$	-0.22	[-0.29, -0.15]	-0.17	[-0.24, -0.095]
$B(\mathrm{gr-qc})$	$+0.075$	[-0.0069, $+0.16$]	$+0.076$	[-0.0066, $+0.16$]
$B(\mathrm{astro})$	-0.6	[-0.74, -0.47]	-0.59	[-0.73, -0.46]

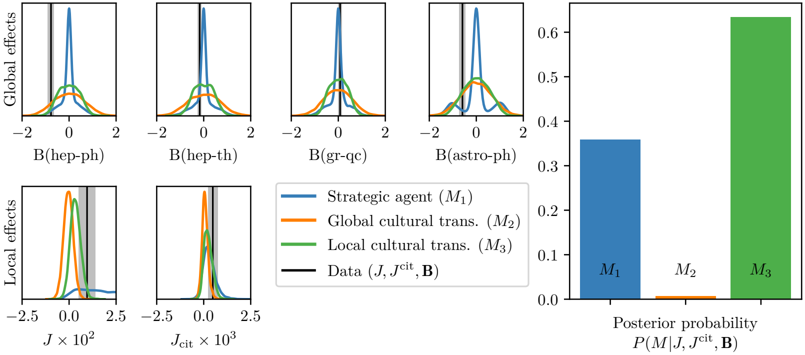
Local versus global coordination

What values of $\boldsymbol{J}$ and $\boldsymbol{B}$ do our models predict? In other words, what is the probability $P(J, J^{\text{cit}}, \boldsymbol{B}|M_i)$ for each model M_i?



Local versus global coordination

Given $P(J, J^{\mathrm{cit}}, \boldsymbol{B}|M_i)$, and the true values of $\boldsymbol{J}$ and $\boldsymbol{B}$, what is $P(M_i|J, J^{\mathrm{cit}}, \boldsymbol{B})$? After a bit of computational trickery – "amortized simulation-based model comparison with neural networks" with BayesFlow –:



Thank you!

Calvert, Randall (1992). "Leadership and its basis in problems of social coordination". In: *International Political Science Review* 13.1.

Centola, Damon and Andrea Baronchelli (Feb. 2015). "The spontaneous emergence of conventions: An experimental study of cultural evolution". In: *Proceedings of the National Academy of Sciences* 112.7.

Delgado, Jordi (2002). "Emergence of social conventions in complex networks". In: *Artificial intelligence* 141.1-2.

Hawkins, Robert XD, Noah D Goodman, and Robert L Goldstone (2019). "The emergence of social norms and conventions". In: *Trends in cognitive sciences* 23.2.

Lewis, David (Jan. 1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

O'Connor, Cailin (June 2020). "Measuring Conventionality". In: *Australasian Journal of Philosophy* 99.3.

Pujol, Josep M et al. (2005). "The role of clustering on the emergence of efficient social conventions". In: *Proceedings of the 19th international joint conference on Artificial intelligence*.

Radev, Stefan T et al. (2021). "Amortized bayesian model comparison with evidential deep learning". In: *IEEE Transactions on Neural Networks and Learning Systems* 34.8.

Amortized simulation-based inference

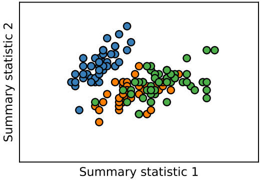
- Even with summary statistics, simulation-based inference is difficult because no simulated sample will *exactly* match the observed data.
- Solution:

Amortized simulation-based inference

- Even with summary statistics, simulation-based inference is difficult because no simulated sample will *exactly* match the observed data.
- Solution:
  - Use amortized inference with neural networks $\Rightarrow$ train a neuralnet to predict the probability of each model M_i given one or more observed outcomes. The neuralnet is trained with many simulated training samples (M_s, O_s) (Radev et al., 2021)

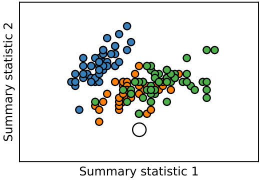
Amortized simulation-based inference

- Even with summary statistics, simulation-based inference is difficult because no simulated sample will *exactly* match the observed data.
- Solution:
  - Use amortized inference with neural networks ⇒ train a neuralnet to predict the probability of each model M_i given one or more observed outcomes. The neuralnet is trained with many simulated training samples (M_s, O_s) (Radev et al., 2021)



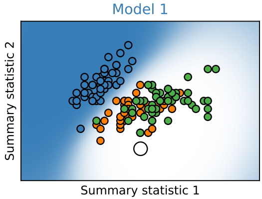
Amortized simulation-based inference

- Even with summary statistics, simulation-based inference is difficult because no simulated sample will *exactly* match the observed data.
- Solution:
  - Use amortized inference with neural networks $\Rightarrow$ train a neuralnet to predict the probability of each model M_i given one or more observed outcomes. The neuralnet is trained with many simulated training samples (M_s, O_s) (Radev et al., 2021)



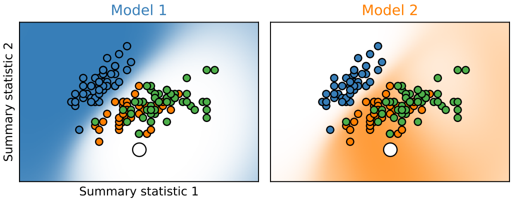
Amortized simulation-based inference

- Even with summary statistics, simulation-based inference is difficult because no simulated sample will *exactly* match the observed data.
- Solution:
    - Use amortized inference with neural networks $\Rightarrow$ train a neuralnet to predict the probability of each model M_i given one or more observed outcomes. The neuralnet is trained with many simulated training samples (M_s, O_s) (Radev et al., 2021)



Amortized simulation-based inference

- Even with summary statistics, simulation-based inference is difficult because no simulated sample will *exactly* match the observed data.
- Solution:
  - Use amortized inference with neural networks $\Rightarrow$ train a neuralnet to predict the probability of each model M_i given one or more observed outcomes. The neuralnet is trained with many simulated training samples (M_s, O_s) (Radev et al., 2021)



Amortized simulation-based inference

- Even with summary statistics, simulation-based inference is difficult because no simulated sample will *exactly* match the observed data.
- Solution:
  - Use amortized inference with neural networks $\Rightarrow$ train a neuralnet to predict the probability of each model M_i given one or more observed outcomes. The neuralnet is trained with many simulated training samples (M_s, O_s) (Radev et al., 2021)