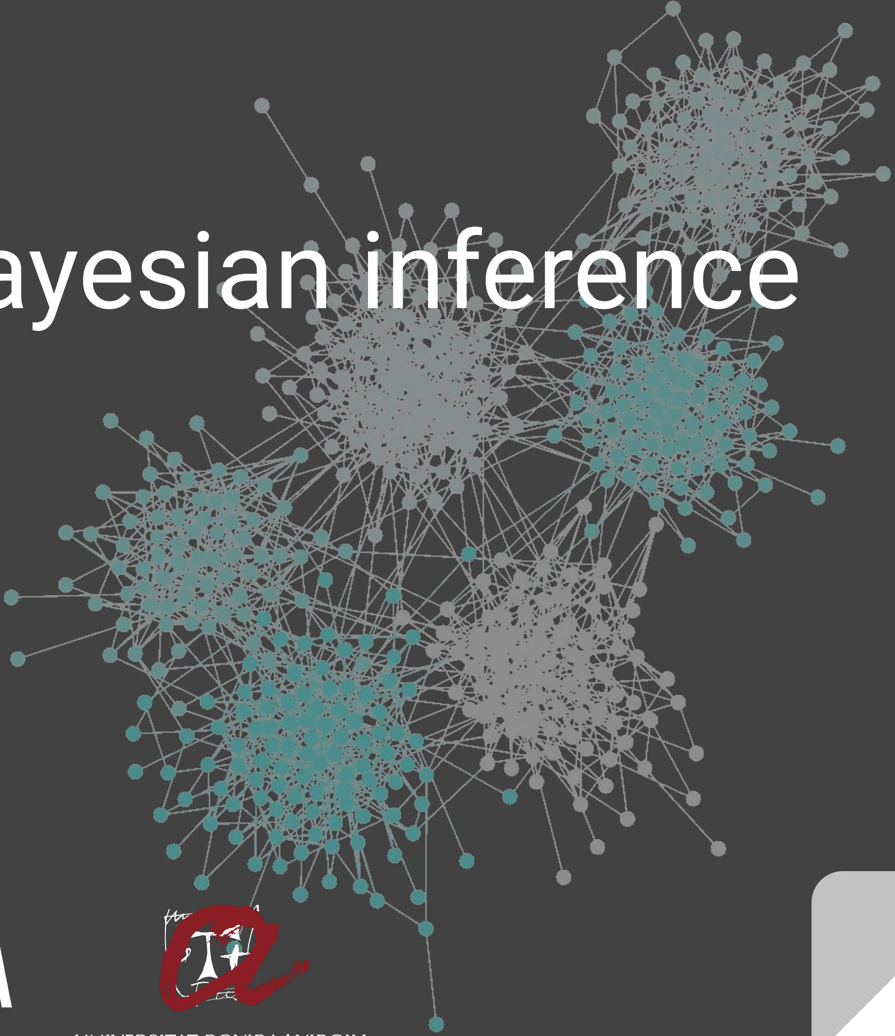
Introduction to Bayesian inference

Theory and practice



Roger Guimerà
ICREA & Univ. Rovira i Virgili, Catalonia

XI Jornada Complexitat
Barcelona, June 5th, 2024

COMPLEXITAT ICREA UNIVERSITAT ROVIRA i VIRGILI

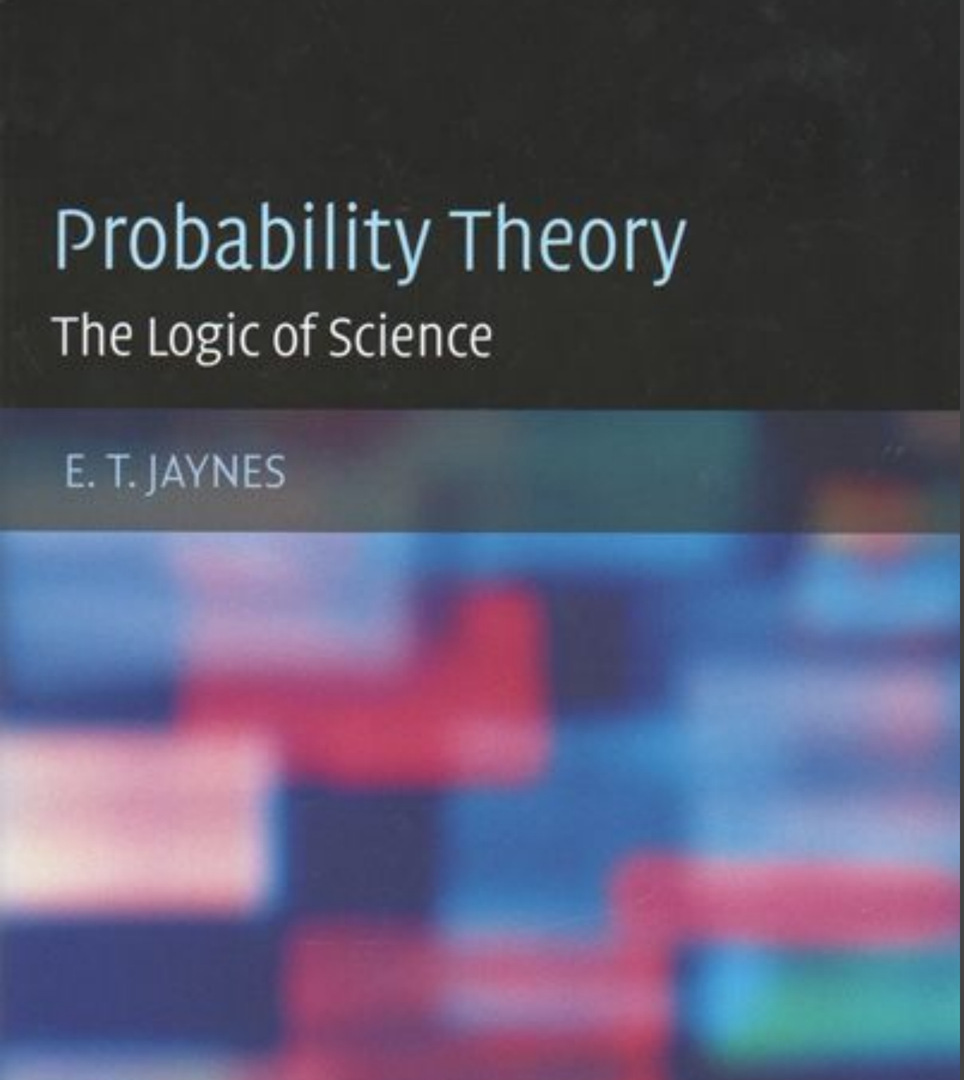
Plan for the lecture

Theory of Bayesian inference:

- Bayesian interpretation of probabilities
- Bayesian model selection
- Bayesian prediction
- Markov chain Monte Carlo (MCMC)

Applications:

- Inferential community detection in complex networks
- Bayesian machine scientist



Theory of Bayesian inference:

- Bayesian interpretation of probabilities
- Bayesian model selection
- Bayesian prediction
- Markov chain Monte Carlo (MCMC)

Applications:

- Inferential community detection in complex networks
- Bayesian machine scientist

What should the policeman do?

Aristotelian logic only deals with absolute certainty; we aim to extend it to plausible reasoning

Two strong syllogisms:

- If A is true, then B is true; A is true, therefore B is true
- If A is true, then B is true; B is false, therefore A is false

Aristotelian logic only deals with absolute certainty; we aim to extend it to plausible reasoning

Two strong syllogisms:

- If A is true, then B is true; A is true, therefore B is true
- If A is true, then B is true; B is false, therefore A is false

Often, however, we have to fall back to weaker syllogisms:

- If A is true, then B is true; B is true, therefore A becomes more plausible
  - A := it will start raining at 10am
  - B := the sky will become cloudy before 10am

Aristotelian logic only deals with absolute certainty; we aim to extend it to plausible reasoning

Two strong syllogisms:

- If A is true, then B is true; A is true, therefore B is true
- If A is true, then B is true; B is false, therefore A is false

Often, however, we have to fall back to weaker syllogisms:

- If A is true, then B is true; B is true, therefore A becomes more plausible
  - A := it will start raining at 10am
  - B := the sky will become cloudy before 10am
- If A is true, then B is true; A is false, therefore B becomes less plausible

Aristotelian logic only deals with absolute certainty; we aim to extend it to plausible reasoning

Two strong syllogisms:

- If A is true, then B is true; A is true, therefore B is true
- If A is true, then B is true; B is false, therefore A is false

Often, however, we have to fall back to weaker syllogisms:

- If A is true, then B is true; B is true, therefore A becomes more plausible
  - A := it will start raining at 10am
  - B := the sky will become cloudy before 10am
- If A is true, then B is true; A is false, therefore B becomes less plausible
- If A is true, then B becomes more plausible; B is true, therefore A becomes more plausible
  - A := it is a robbery
  - B := there is a broken glass, thief is wearing a mask...

We want do design a “thinking robot” that reasons (that is, deals quantitatively with plausibility, extending logic) according to definite rules

We aim to design a “thinking robot”

We want do design a robot that reasons (that is, deals quantitatively with plausibility) according to definite rules

Our robot reasons about propositions {A, B, …} that are either true or false and have unambiguous meaning

We aim to design a “thinking robot”

We want do design a robot that reasons (that is, deals quantitatively with plausibility) according to definite rules

Our robot reasons about propositions {A, B, …} that are either true or false and have unambiguous meaning

These propositions obey the rules of usual symbolic logic (Boolean algebra):

- AB := A and B are both true
- $A+B$:= at least A or B are true
- $\underline{A}$:= A is false

We aim to design a “thinking robot”

We want do design a robot that reasons (that is, deals quantitatively with plausibility) according to definite rules

Our robot reasons about propositions {A, B, …} that are either true or false and have unambiguous meaning

These propositions obey the rules of usual symbolic logic (Boolean algebra):

- AB := A and B are both true
- $A+B$:= at least A or B are true
- $\underline{A}$:= A is false

- **Commutativity** $AB = BA$; $A+B = B+A$
- **Associativity** $A(BC)=(AB)C=ABC$; $A + (B+C) = (A+B) + C = A+B+C$
- **Distributivity** $A (B+C) = AB + AC$; $A + (BC) = (A+B) (A+C)$
- **Duality** If $C = AB$, then $\underline{C} = \underline{A}+\underline{B}$; If $D = A+B$, then $\underline{D} = \underline{A}\ \underline{B}$

Basic desiderata for our thinking robot

I. Degrees of plausibility are represented by real numbers

The ***plausibility*** that the robot assigns to some proposition A will, in general, depend on whether we told the robot that some other proposition B is true. Therefore, we represent this plausibility as:

$A|B$

This stands for a **real number**, and so do other symbols:

$A|BC$ is the plausibility that A is true given that B and C are true

$A+B|C$ is the plausibility that A or B are true given that C is true

Basic desiderata for our thinking robot

II. Qualitative correspondence with common sense

If our robot has old information C which get updated to C' in such a way that the plausibility of A increases:

$$A|C' > A|C$$

but the plausibility of B given A doesn't change:

$$B|AC' = B|AC$$

Basic desiderata for our thinking robot

II. Qualitative correspondence with common sense

If our robot has old information C which get updated to C' in such a way that the plausibility of A increases:

$$A|C' > A|C$$

but the plausibility of B given A doesn't change:

$$B|AC' = B|AC$$

then, it must be true that:

$$\underline{A}|C' < \underline{A}|C$$

$$AB|C' \geq AB|C$$

Basic desiderata for our thinking robot

III. Consistency

IIIa. If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result

IIIb. The robot always takes into account all of the evidence it has relevant to a question

IIIc. The robot always represents equivalent states of knowledge by equivalent plausibility assignments

Cox's Theorem

These conditions uniquely determine the rules by which our robot must reason

I. Degrees of plausibility are represented by real numbers
II. Qualitative correspondence with common sense
III. Consistency

Cox's Theorem

The product rule

We seek to relate $AB|C$ to the plausibilities $A|C$ and $B|C$ separately

Using (I) + (II) + (IIIa) one can prove that there is an increasing monotonic function w of the plausibility that verifies

- $w(AB|C) = w(A|BC)\, w(B|C) = w(B|AC)\, w(A|C)$

Cox's Theorem

The product rule

We seek to relate AB|C to the plausibilities A|C and B|C separately

Using (I) + (II) + (IIIa) one can prove that there is an increasing monotonic function w of the plausibility that verifies

- $w(AB|C) = w(A|BC)\, w(B|C) = w(B|AC)\, w(A|C)$
- Certainty of A|C corresponds to $w(A|C) = 1$

Cox's Theorem

The product rule

We seek to relate AB|C to the plausibilities A|C and B|C separately

Using (I) + (II) + (IIIa) one can prove that there is an increasing monotonic function w of the plausibility that verifies

- $w(AB|C) = w(A|BC)\ w(B|C) = w(B|AC)\ w(A|C)$
- Certainty of A|C corresponds to $w(A|C) = 1$
- Certainty of $\underline{A}|C$ corresponds to $w(A|C) = 0$

Cox’s Theorem

The sum rule

We start by noting that, since $A + \underline{A}$ is always true, the plausibility of A must depend on the plausibility of A:

$$w(\underline{A}|B) = S[w(A|B)]$$

Cox’s Theorem
The sum rule

We start by noting that, since A + $\underline{A}$ is always true, the plausibility of A must depend on the plausibility of A:

$$w(\underline{A}|B) = S[w(A|B)]$$

Product rule together with (IIIa) then imply that:

$$w^m(A|B) + w^m(\underline{A}|B) = 1 \text{ with arbitrary positive } m$$

Cox's Theorem
Putting it all together

From our basic desiderata, we have been able to conclude that there must be a positive monotonic increasing function of the plausibility that verifies:

1. $w(AB|C) = w(A|BC)\ w(B|C) = w(B|AC)\ w(A|C)$
2. $w^m(A|B) + w^m(\underline{A}|B) = 1$ with arbitrary positive m
3. Certainty of $A|C$ corresponds to $w(A|C) = 1$
4. Certainty of $\underline{A}|C$ corresponds to $w(A|C) = 0$

Cox's Theorem

Putting it all together

We can now define $p(x) := w^m(x)$ and our rules take the form:

1. $p(AB|C) = p(A|BC)\, p(B|C) = p(B|AC)\, p(A|C)$
2. $p(A|B) + p(\underline{A}|B) = 1$
3. Certainty of $A|C$ corresponds to $p(A|C) = 1$
4. Certainty of $\underline{A}|C$ corresponds to $p(A|C) = 0$

We still don't know what actual numerical values of plausibility should be assigned at the beginning of the problem so that the robot can get started!

We can solve this problem by invoking (IIIb) and (IIIc) (which we haven't used, yet!)

Numerical values

Consider $p(A_1+A_2+...+A_N|B)$, where A_i are mutually exclusive and exhaustive (that is, one and only one of them must be true)

B does not favor any of the propositions A_i

Applying the rules we have so far one can prove that

$$\sum_{i=1}^{N} p(A_i|B) = 1$$

Numerical values

Consider $p(A_1+A_2+...+A_N|B)$, where A_i are mutually exclusive and exhaustive (that is, one and only one of them must be true)

B does not favor any of the propositions A_i

Applying the rules we have so far one can prove that

$$\sum_{i=1}^{N} p(A_i|B) = 1$$

Now, using (IIIb) (the robot always takes into account all of the evidence) and (IIIc) (the robot always represents equivalent states of knowledge by equivalent plausibility assignments), one can prove that

$$p(A_i|B) = \frac{1}{N}$$

and we have arrived at definite numerical values!!

Cox's Theorem leads to the quantitative and precise definition of probability

We can now define $p(x) := w^m(x)$ and our rules take the form:

1. $p(AB|C) = p(A|BC)\ p(B|C) = p(B|AC)\ p(A|C)$
2. $p(A|B) + p(\underline{A}|B) = 1$
3. Certainty of $A|C$ corresponds to $p(A|C) = 1$
4. Certainty of $\underline{A}|C$ corresponds to $p(A|C) = 0$

Information given to the robot (we've seen one case but it can be generalized) determines completely the values of the quantities $p(A|B)$ and allows the robot to start

Since p is fixed by the data (not $A|B$) we can just turns thing around and:

- say that $A|B$ is a monotonic function of p (instead of the opposite)
- call p **probability and make it the object of our study**
- let the plausibility $A|B$ fade

So what is **probability**?

…and what it is *not*?

Probability is a representation of the plausibility of a proposition in the “mind” of our robot

Note that we have made no reference whatsoever to frequencies

Probability theory (as we know it) is therefore the extended logic we were looking for

Plan for the lecture

Theory of Bayesian inference:

- Bayesian interpretation of probabilities
- Bayesian model selection
- Bayesian prediction
- Markov chain Monte Carlo (MCMC)

Applications:

- Inferential community detection in complex networks
- Bayesian machine scientist

Probability theory and Bayesian model selection: A simple example with coins

You know I have two coins:

Coin A

Coin B

I select one of the coins without letting you know whether it is A or B

I toss the coin and get ***tails***

What is the probability that the coin I selected is B?

Probability theory and Bayesian model selection: A *not so* simple example with coins

You know I have two coins:

Coin A

Coin B

I select one of the coins without letting you know whether it is A or B

I toss the coin and get ***heads***

What is the probability that the coin I selected is B?

Probability theory and Bayesian model selection: Yet another example with coins

You know I have two coins:

Coin A

Coin B

Coin C

I select one of the coins without letting you know whether it is A, B, or C

I toss the coin and get ***heads***

What is the probability that the coin I selected is B?

Intuition from the examples with coins

“Final” probability of the model is proportional to the probability of generating the observed data with the model

“Final” probability of the model is also proportional to the probability of the model *a priori*, that is, before seeing any data

This is a consequence of the application of Bayes theorem to model selection

$$\begin{aligned} p(A,\, B) &= p(A|B)\, p(B) \\ &= p(B|A)\, p(A) \end{aligned} \quad \Rightarrow \quad p(A|B) = \frac{p(B|A)\, p(A)}{p(B)}$$

Suppose we have some data *D* and we want to say something about a model *M*. What is the plausibility of model *M*?

This is a consequence of the application of Bayes theorem to model selection

$$p(A,\, B) = p(A|B)\, p(B) = p(B|A)\, p(A) \quad \Rightarrow \quad p(A|B) = \frac{p(B|A)\, p(A)}{p(B)}$$

Suppose we have some data *D* and we want to say something about a model *M*. What is the plausibility of model *M*?

$$p(M|D) = \frac{p(D|M)\, p(M)}{p(D)}$$

This is a consequence of the application of Bayes theorem to model selection

$$\begin{aligned} p(A,\, B) &= p(A|B)\, p(B) \\ &= p(B|A)\, p(A) \end{aligned} \quad \Rightarrow \quad p(A|B) = \frac{p(B|A)\, p(A)}{p(B)}$$

Suppose we have some data *D* and we want to say something about a model *M*. What is the plausibility of model *M*?

$$\underbrace{p(M|D)}_{\text{posterior}} = \frac{\overbrace{p(D|M)}^{\text{likelihood}}\, \overbrace{p(M)}^{\text{prior}}}{\underbrace{p(D)}_{\text{evidence}}}$$

Typically, our models have parameters θ, and our model selection approach needs to take this fact into consideration

Without parameters:

$$p(M|D) = \frac{p(D|M)\,p(M)}{p(D)}$$

With parameters:

Typically, our models have parameters θ, and our model selection approach needs to take this fact into consideration

Without parameters:

$$p(M|D) = \frac{p(D|M)\,p(M)}{p(D)}$$

With parameters:

$$p(M,\theta|D) = \frac{p(D|M,\theta)\,p(M,\theta)}{p(D)} = \frac{p(D|M,\theta)\,p(\theta|M)\,p(M)}{p(D)}$$

$$p(M|D) = \int_\Theta d\theta\; p(M,\theta|D) = \frac{1}{p(D)} \underbrace{\int_\Theta d\theta\; p(D|M,\theta)\,p(\theta|M)p(M)}_{\text{integrated likelihood}}$$

Information theoretic interpretation

The most plausible model given the data has the shortest description length

The posterior can always be written as:

$$
\begin{aligned}
p(M|D) &= \frac{1}{p(D)} \int_{\Theta} d\theta \, p(D|M,\theta) \, p(\theta|M) \, p(M) \\
&= \frac{\mathrm{e}^{-\mathcal{L}(M,D)}}{p(D)}
\end{aligned}
$$

Information theoretic interpretation

The most plausible model given the data has the shortest description length

The posterior can always be written as:

$$
\begin{aligned}
p(M|D) &= \frac{1}{p(D)} \int_\Theta d\theta \, p(D|M,\theta) \, p(\theta|M) \, p(M) \\
&= \frac{\mathrm{e}^{-\mathcal{L}(M,D)}}{p(D)}
\end{aligned}
$$

with the ***description length***:

$$
\mathcal{L}(M,D) = -\log p(M,D) = -\log \int_\Theta d\theta \, p(D|M,\theta) \, p(\theta|M) - \log p(M)
$$

But why do we call -log $p(M, D)$ the ***description length***?

Information theoretic interpretation

The most plausible model given the data has the shortest description length

The posterior can always be written as:

$$
\begin{aligned}
p(M|D) &= \frac{1}{p(D)} \int_\Theta d\theta \, p(D|M,\theta) \, p(\theta|M) \, p(M) \\
&= \frac{\mathrm{e}^{-\mathcal{L}(M,D)}}{p(D)}
\end{aligned}
$$

with the ***description length***:

$$
\mathcal{L}(M,D) = -\log p(M,D) = -\log \int_\Theta d\theta \, p(D|M,\theta) \, p(\theta|M) - \log p(M)
$$

Statistical physics interpretation

The most plausible model given the data is the ground state

The posterior can always be written as:

$$p(M|D) = \frac{1}{p(D)} \int_\Theta d\theta\, p(D|M,\theta)\, p(\theta|M)\, p(M)$$
$$= \frac{\mathrm{e}^{-\mathcal{L}(M,D)}}{p(D)}$$

with the ***energy***:

$$\mathcal{L}(M,D) = -\log p(M,D) = -\log \int_\Theta d\theta\, p(D|M,\theta)\, p(\theta|M) - \log p(M)$$

So far…

Bayesian model selection:

- ***Probabilistic interpretation*** - Most plausible model given the data
- ***Information theoretic interpretation*** - Shortest (or, equivalently, most compressive) description of the data
- ***Statistical physics interpretation*** - Ground state of a system whose “states” are the models

For models with parameters, we must integrate them and use the ***integrated likelihood*** instead of the “regular” likelihood

Arguments for a probabilistic approach

Cox-type argument: Any alternative way to assign plausibilities to models must violate some of the very basic conditions in the desiderata

Dutch book-type argument: Betting on models using any alternative assignment of plausibility results in sets of bets that one would be willing to accept but that result in certain loss

Consistency argument: Any alternative that does not coincide with the probabilistic approach in the large *N* limit will **not** select the true generating model in this limit

Information theory argument: Any alternative way of selecting models will lead to models that compress the data less

Arguments for a probabilistic approach

Cox-type argument: Any alternative way to assign plausibilities to models must violate some of the very basic conditions in the desiderata

Dutch book-type argument: Betting on models using any alternative assignment of plausibility results in sets of bets that one would be willing to accept but that result in certain loss

Consistency argument: Any alternative that does not coincide with the probabilistic approach in the large N limit will **not** select the true generating model in this limit

Information theory argument: Any alternative way of selecting models will lead to models that compress the data less

Arguments for a probabilistic approach

Cox-type argument: Any alternative way to assign plausibilities to models must violate some of the very basic conditions in the desiderata

Dutch book-type argument: Betting on models using any alternative assignment of plausibility results in sets of bets that one would be willing to accept but that result in certain loss

Consistency argument: Any alternative that does not coincide with the probabilistic approach in the large *N* limit will **not** select the true generating model in this limit

Information theory argument: Any alternative way of selecting models will lead to models that compress the data less

Arguments for a probabilistic approach

Cox-type argument: Any alternative way to assign plausibilities to models must violate some of the very basic conditions in the desiderata

Dutch book-type argument: Betting on models using any alternative assignment of plausibility results in sets of bets that one would be willing to accept but that result in certain loss

Consistency argument: Any alternative that does not coincide with the probabilistic approach in the large N limit will **not** select the true generating model in this limit

Information theory argument: Any alternative way of selecting models will lead to models that compress the data less

Plan for the lecture

Theory of Bayesian inference:

- Bayesian interpretation of probabilities
- Bayesian model selection
- Bayesian prediction
- Markov chain Monte Carlo (MCMC)

Applications:

- Inferential community detection in complex networks
- Bayesian machine scientist

Probability theory and Bayesian inference: last example with coins

Imagine that we toss a coin 5 times and get D := {H,H,T,H,T}

So, what is the probability that the next toss gives H?

Probability theory and Bayesian inference: last example with coins

Imagine that we toss a coin 5 times and get D := {H,H,T,H,T}

So, what is the probability that the next toss gives H?

Bernoulli process At each toss, independently of previous ones, **probability of getting H is *h***. The model is fully specified by *h* (therefore, *M* := *h*)

Then, the probability of getting {H,H,T,H,T} is

$$p(\{H,H,T,H,T\}|h) = h \times h \times (1-h) \times h \times (1-h) = h^3(1-h)^2$$

Probability theory and Bayesian inference: last example with coins

Imagine that we toss a coin 5 times and get D := {H,H,T,H,T}

So, what is the probability that the next toss gives H?

Bernoulli process At each toss, independently of previous ones, **probability of getting H is *h***. The model is fully specified by *h* (therefore, *M* := *h*)

Then, the probability of getting {H,H,T,H,T} is

$$p(\{H,H,T,H,T\}|h) = h \times h \times (1-h) \times h \times (1-h) = h^3(1-h)^2$$

If, a priori, we don't know anything about the right value of h, we can assume that the prior is uniform

$$p(h) = 1, h \in [0,1]$$

Probability theory and Bayesian inference: last example with coins

Imagine that we toss a coin 5 times and get D := {H,H,T,H,T}

So, what is the probability that the next toss gives H?

Bernoulli process At each toss, independently of previous ones, **probability of getting H is *h***. The model is fully specified by *h* (therefore, *M* := *h*)

Then, the probability of getting {H,H,T,H,T} is

$$p(\{H,H,T,H,T\}|h) = h \times h \times (1-h) \times h \times (1-h) = h^3(1-h)^2$$

If, a priori, we don't know anything about the right value of h, we can assume that the prior is uniform

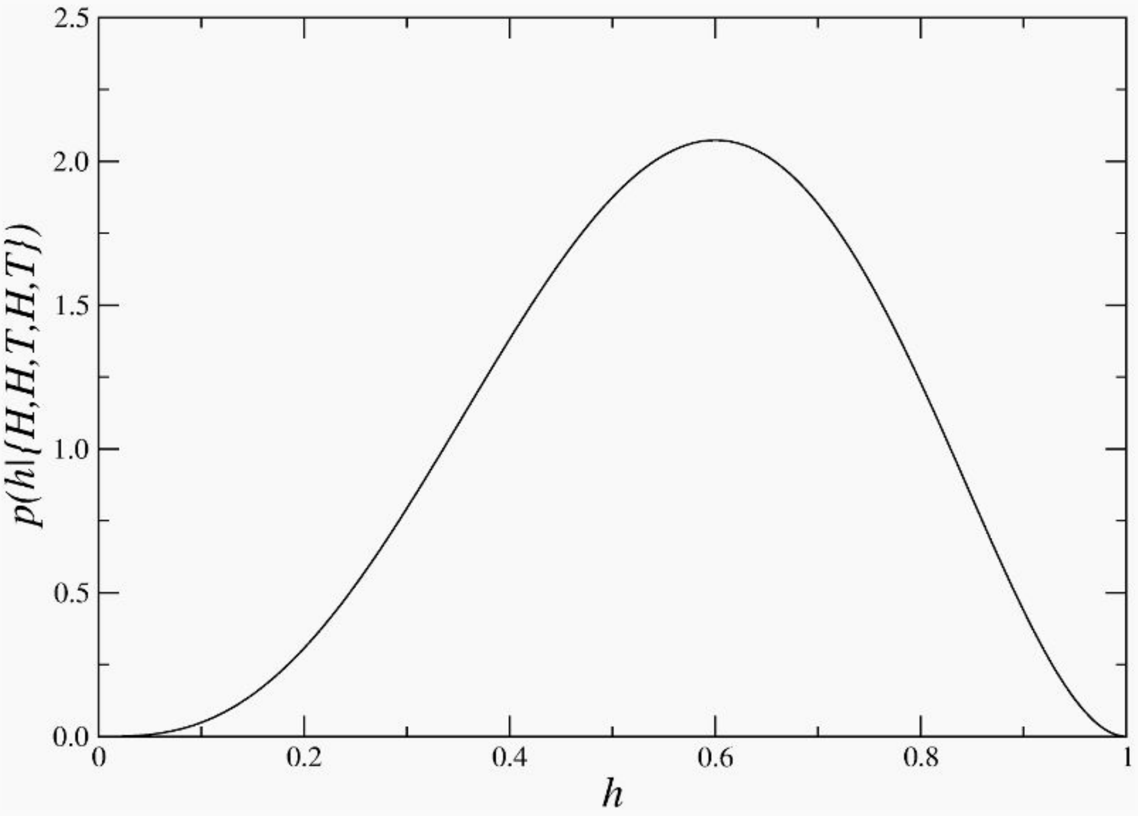
$$p(h) = 1, h \in [0,1]$$

Then, we finally have that

$$p(h|\{H,H,T,H,T\}) \propto p(\{H,H,T,H,T\}|h)\, p(h) = h^3(1-h)^2$$

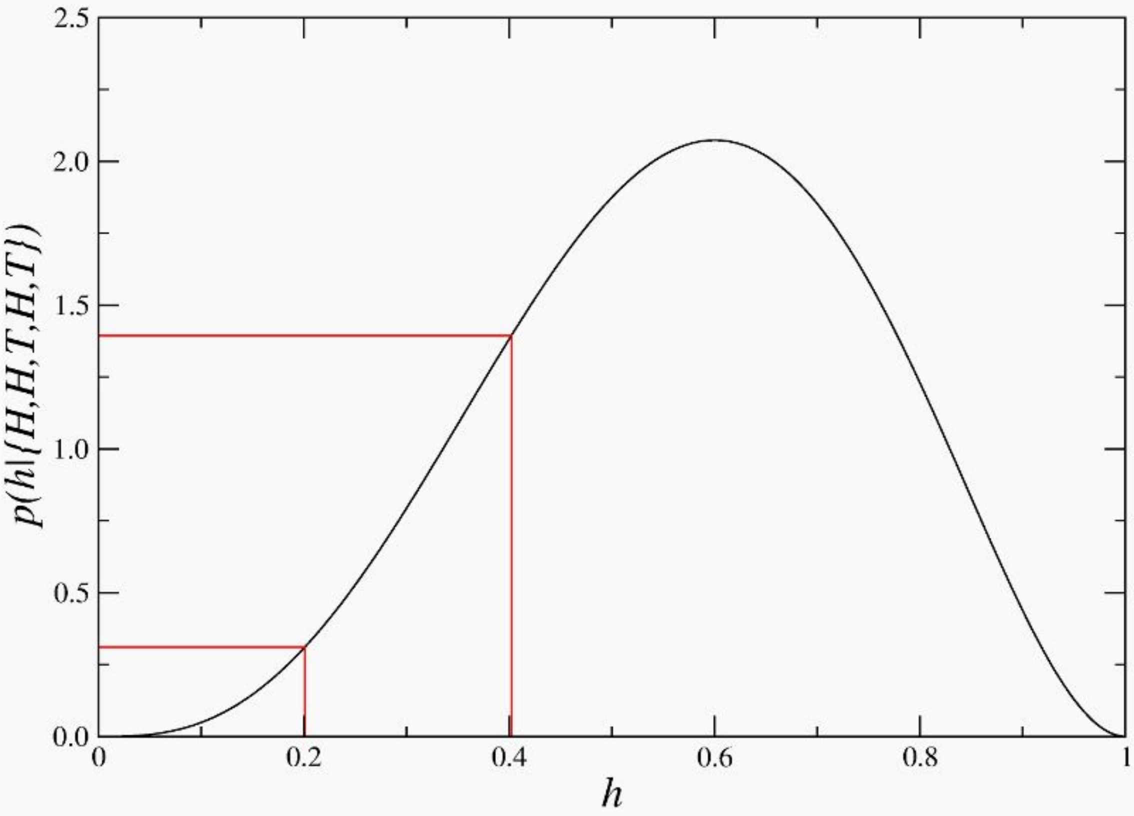
Probability theory and Bayesian inference: last example with coins

$$p(h|\{H, H, T, H, T\}) \propto h^3(1-h)^2$$



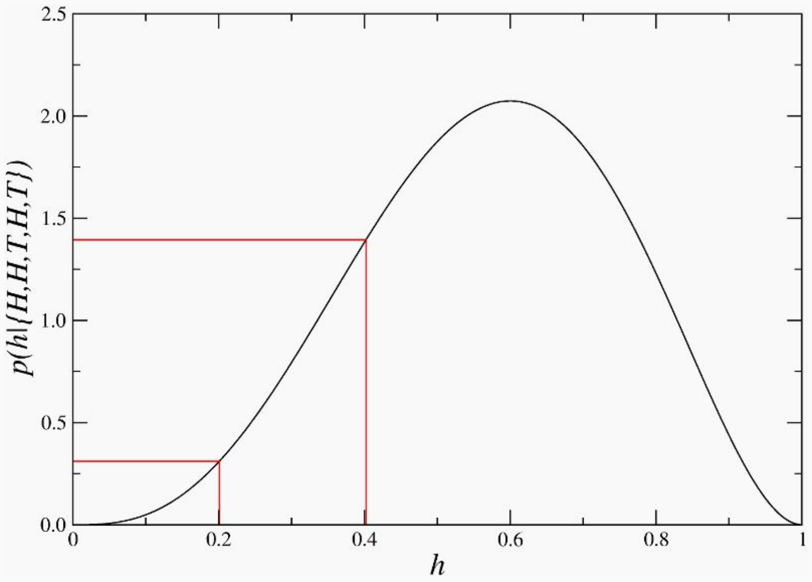
Within the Bayesian approach, we can and should consider all evidence at hand

$$p(h|\{H,H,T,H,T\}) \propto h^3(1-h)^2$$



So, what is the probability that the next toss gives H?

$$p(h|\{H,H,T,H,T\}) \propto h^3(1-h)^2$$



$$p(\text{next toss} = H|\{H,H,T,H,T\}) = \int_0^1 h \times p(h|\{H,H,T,H,T\})dh = \frac{4}{7}$$

In general, carrying out integrals like this one is not straightforward

In the previous example, we were interested in calculating some property using our complete probabilistic description of the parameter of the model (the posterior)

$$p(\text{next toss} = H|\{H,H,T,H,T\}) = \int_0^1 h \times p(h|\{H,H,T,H,T\})dh = \frac{4}{7}$$

This is, in fact, a very common situation

$$\langle f(M) \rangle = \int f(M) \times p(M|D)dM$$

Unfortunately, unlike in the coin example, more often than not these integrals cannot be calculated exactly

Plan for the lecture

Theory of Bayesian inference:

- Bayesian interpretation of probabilities
- Bayesian model selection
- Bayesian prediction
- Markov chain Monte Carlo (MCMC)

Applications:

- Inferential community detection in complex networks
- Bayesian machine scientist

In general, carrying out integrals like this one is not straightforward

$$\langle f(M) \rangle = \int f(M) \times p(M|D) dM$$

In general, carrying out integrals like this one is not straightforward

$$\langle f(M) \rangle = \int f(M) \times p(M|D) dM$$

When this integral cannot be computed analytically or numerically, we can use the approximation

$$\langle f(M) \rangle \approx \frac{1}{N} \sum_i f(M_i)$$

where the sum is over N models **sampled from the posterior distribution** p(M|D), which we do by means of **Markov Chain Monte Carlo (MCMC)**.

MCMC: Gibbs sampler

Suppose that my model M can be characterized by some “parameters”

$$M \equiv \{\psi_1, \ldots, \psi_p\}$$

The Gibbs sampler is an iterative process in which parameters are selected one by one and updated according to

$$\psi_j^{(t+1)} \sim p(\psi_j | \psi_1^{(t+1)}, \ldots, \psi_{j-1}^{(t+1)}, \psi_{j+1}^{(t)}, \ldots, \psi_p^{(t)}, D)$$

Unfortunately, this only works if this conditional probability can be calculated

MCMC: Metropolis-Hastings sampler

Suppose that my model M can be characterized by some “parameters”

$$M \equiv \psi$$

The MH sampler is an iterative process that proceeds as follows:

MCMC: Metropolis-Hastings sampler

Suppose that my model M can be characterized by some “parameters”

$$M \equiv \psi$$

The MH sampler is an iterative process that proceeds as follows:

- Generate a new configuration from some proposal generation distribution

$$\psi^{(t+1)} \sim q(\psi^{(t+1)}|\psi^{(t)})$$

MCMC: Metropolis-Hastings sampler

Suppose that my model M can be characterized by some “parameters”

$$M \equiv \psi$$

The MH sampler is an iterative process that proceeds as follows:

- Generate a new configuration from some proposal generation distribution

$$\psi^{(t+1)} \sim q(\psi^{(t+1)}|\psi^{(t)})$$

- Compute

$$a\left(\psi^{(t)}, \psi^{(t+1)}\right) = \min\left\{1, \frac{p(\psi^{(t+1)}|D)}{p(\psi^{(t)}|D)} \frac{q(\psi^{(t)}|\psi^{(t+1)})}{q(\psi^{(t+1)}|\psi^{(t)})}\right\}$$

MCMC: Metropolis-Hastings sampler

Suppose that my model M can be characterized by some “parameters”

$$M \equiv \psi$$

The MH sampler is an iterative process that proceeds as follows:

- Generate a new configuration from some proposal generation distribution

$$\psi^{(t+1)} \sim q(\psi^{(t+1)} | \psi^{(t)})$$

- Compute

$$a\left(\psi^{(t)}, \psi^{(t+1)}\right) = \min\left\{1, \frac{p(\psi^{(t+1)}|D)}{p(\psi^{(t)}|D)} \frac{q(\psi^{(t)}|\psi^{(t+1)})}{q(\psi^{(t+1)}|\psi^{(t)})}\right\}$$

- Accept the new configuration with probability $a\left(\psi^{(t)}, \psi^{(t+1)}\right)$

MCMC samples from the posterior

In general, MCMC gives us a sample from the posterior p(M|D), that is, a **collection of models** rather than a single best model

The ensemble allows us to do **model averaging or choosing particularly relevant models**, depending on the question we need to address

Plan for the lecture

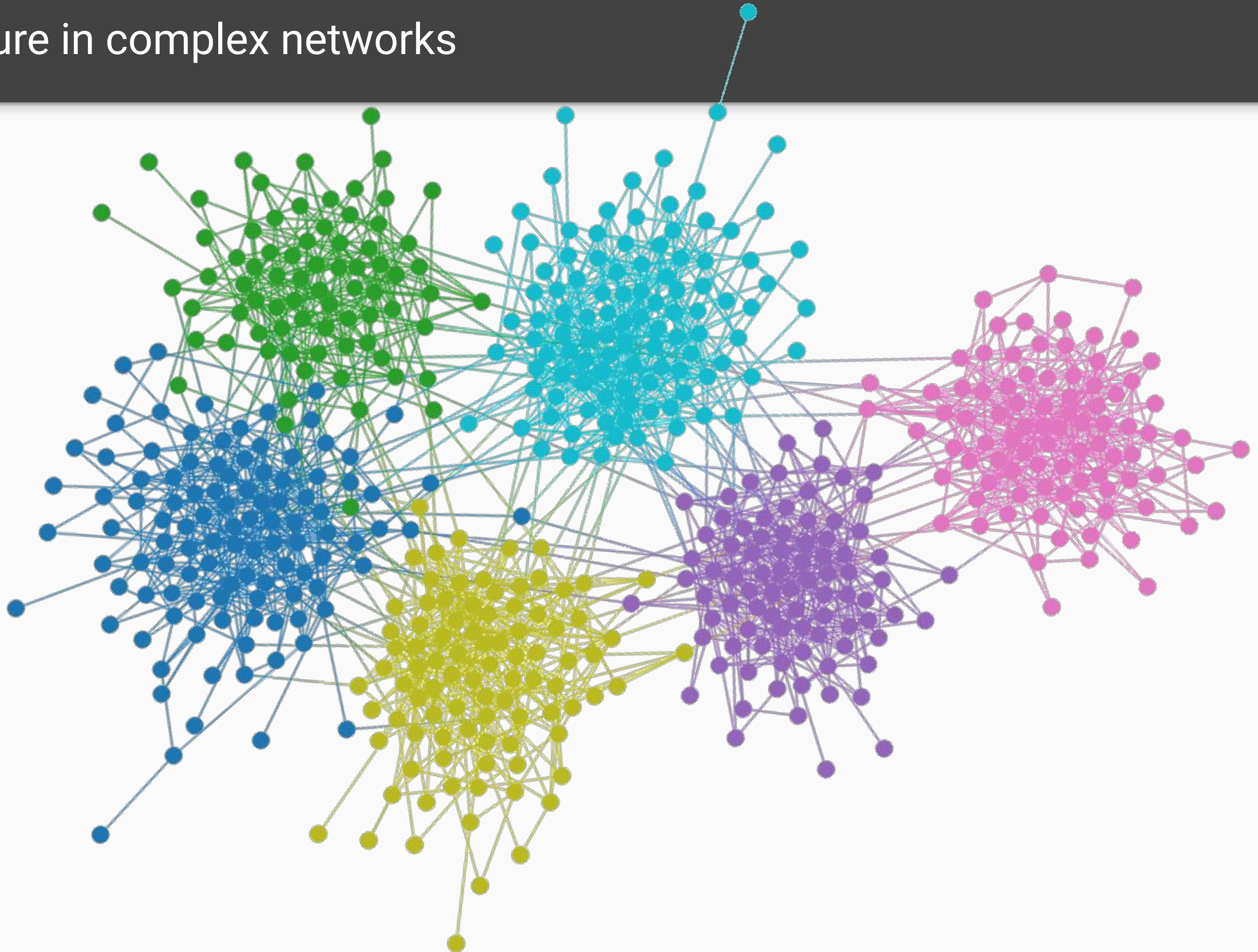
Theory of Bayesian inference:

- Bayesian interpretation of probabilities
- Bayesian model selection
- Bayesian prediction
- Markov chain Monte Carlo (MCMC)

Applications:

- Inferential community detection in complex networks
- Bayesian machine scientist

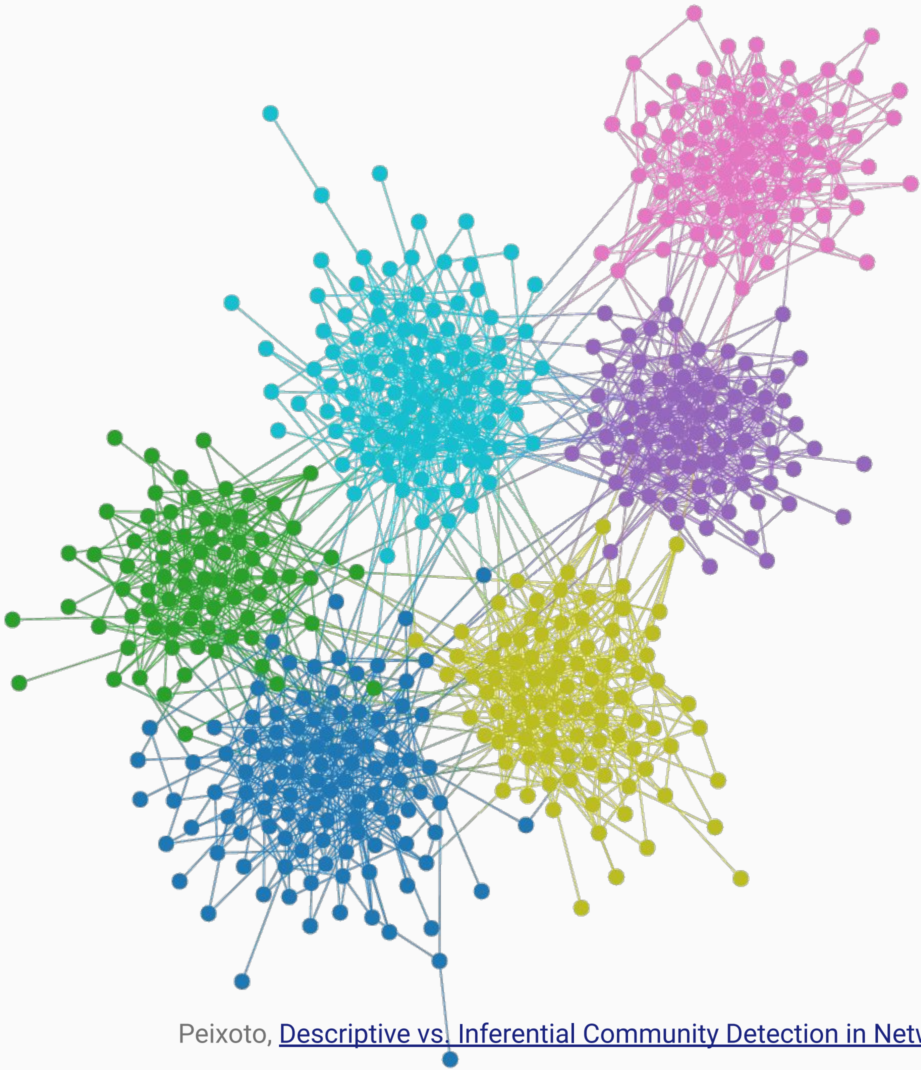
Group structure in complex networks



Community detection

We aim to divide a network—typically one that is large—into smaller groups of nodes that are ***similarly connected to others***

With such a division, we can better summarize the large-scale structure of the network by describing how these groups are connected, instead of each individual node



Peixoto, Descriptive vs. Inferential Community Detection in Networks (2023)

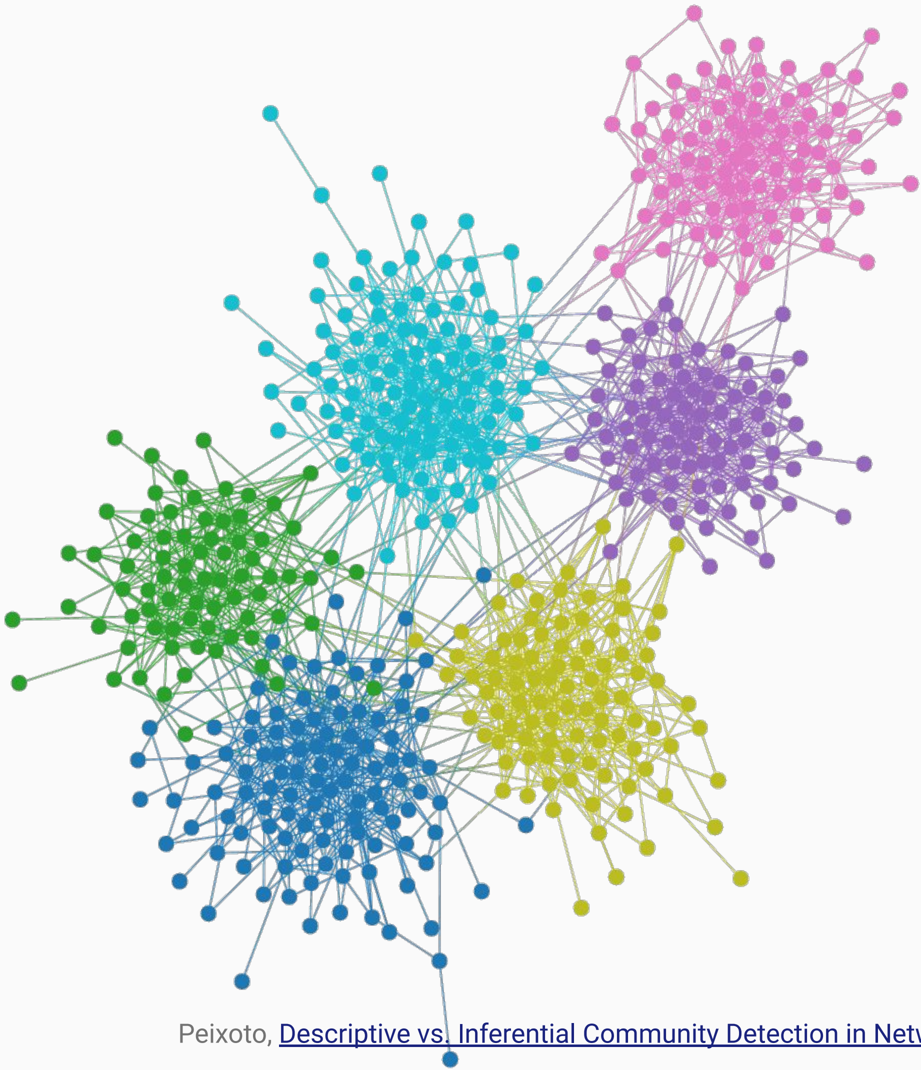
Community detection as model selection

Each partition ***M=b*** of the nodes into groups amounts to a different model of our data, that is, our observed network ***D=Aᵒ***

$$p(b|A^o) = \frac{e^{-\mathcal{L}(b,A^o)}}{Z}$$

with:

$$\mathcal{L}\left(b,\ A^o\right) =$$
$$-\log \int_\Theta d\theta\, p(A^o|b,\theta)\, p(\theta|b)\, p(b)$$



Peixoto, Descriptive vs. Inferential Community Detection in Networks (2023)

Community detection as model selection

Each partition ***M=b*** of the nodes into groups amounts to a different model of our data, that is, our observed network ***D=A°***

$$p(b|A^o) = \frac{\mathrm{e}^{-\mathcal{L}(b,A^o)}}{Z}$$

with:

$$\mathcal{L}\,(b,\ A^o) = -\log \int_\Theta d\theta\, p(A^o|b,\theta)\, p(\theta|b)\, p(b)$$

$$p(A^o|b,\theta) =$$

$$p(\theta|b) =$$

$$p(b) =$$

Community detection as model selection

Each partition ***M=b*** of the nodes into groups amounts to a different model of our data, that is, our observed network ***D=Aº***

$$p(b|A^o) = \frac{e^{-\mathcal{L}(b,A^o)}}{Z}$$

with:

$$\mathcal{L}(b,\ A^o) = -\log \int_\Theta d\theta\, p(A^o|b,\theta)\, p(\theta|b)\, p(b)$$

$$p(A^o|b,\theta) =$$

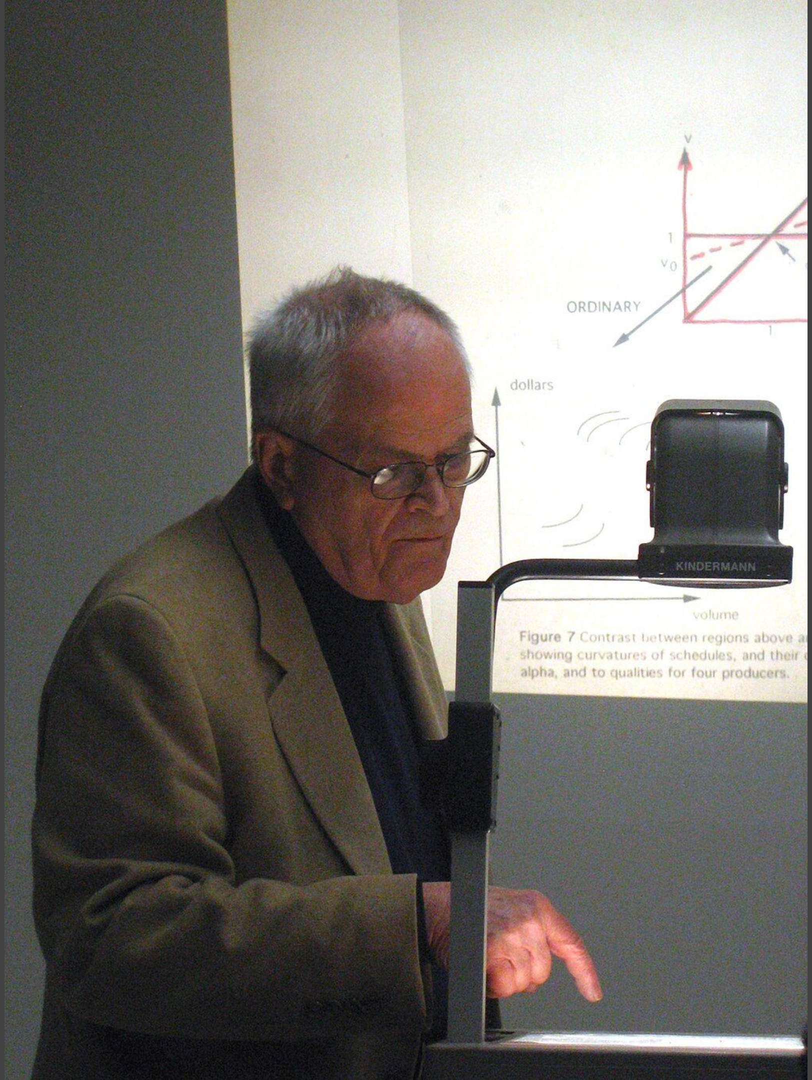
We need to specify a generative model!!!

$$p(\theta|b) =$$

$$p(b) =$$

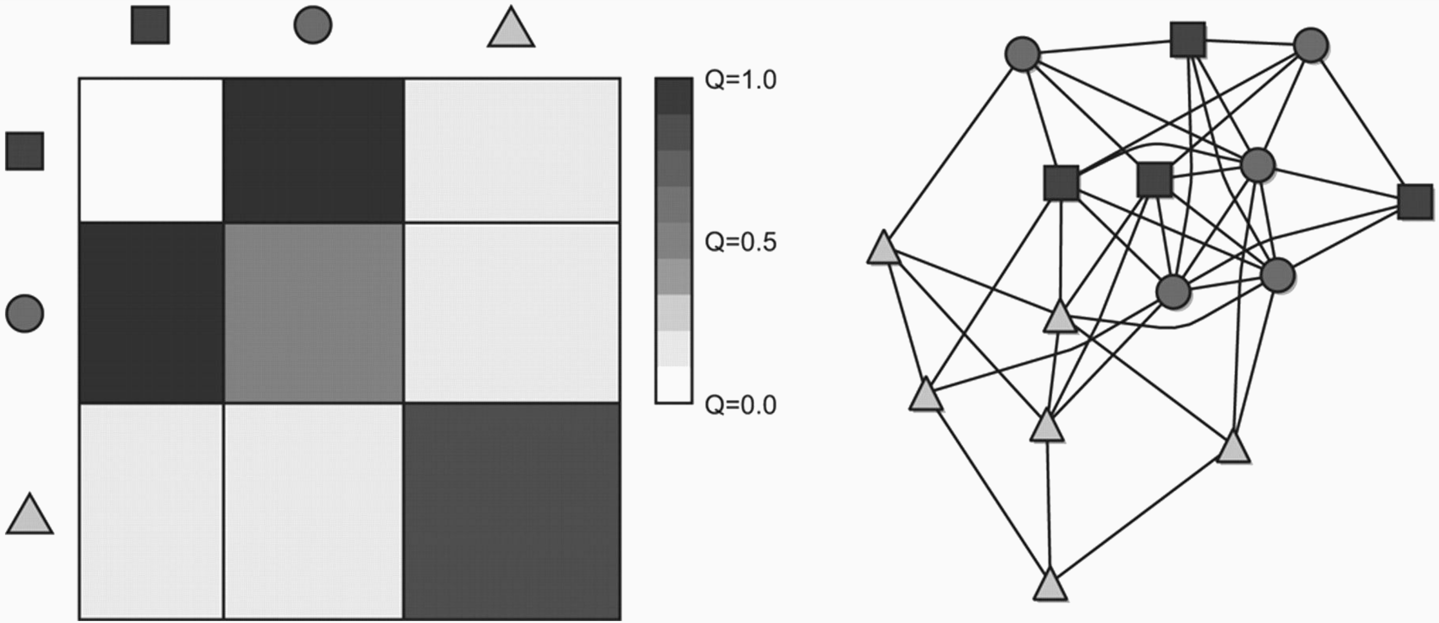
Harrison White

March 21, 1930 – May 18, 2024



The stochastic block model

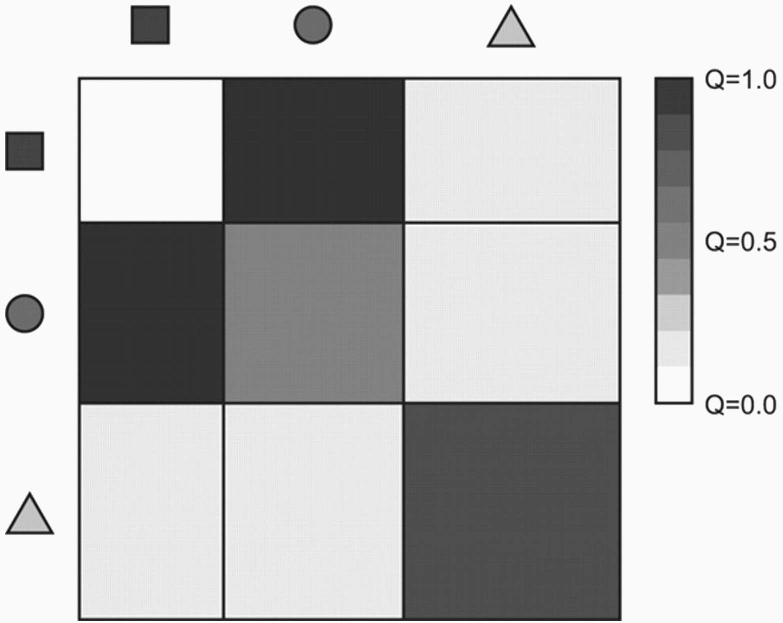
We assume that nodes belong to groups, and their interactions depend only on those groups

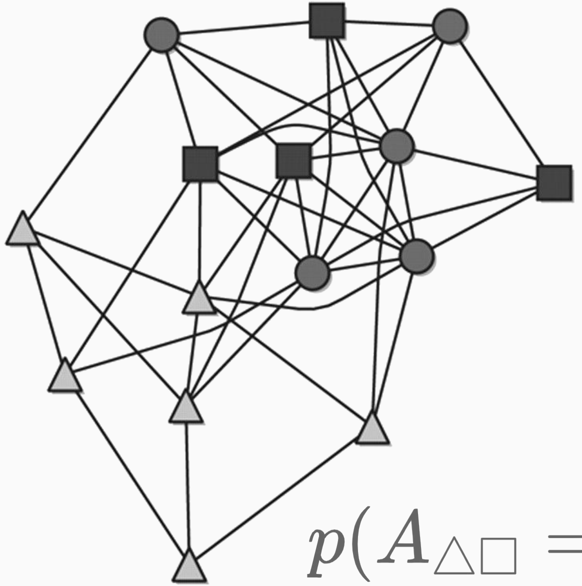


White, Boorman, Breiger, AJS (1976); Holland, Laskey, Leinhardt, Soc. Networks (1983); Nowicki, Snijders, JASA (2001)

The stochastic block model

We assume that nodes belong to groups, and their interactions depend only on those groups





$$p(A_{\triangle\square} = 1 | Q) = q_{\triangle\square}$$

White, Boorman, Breiger, AJS (1976); Holland, Laskey, Leinhardt, Soc. Networks (1983); Nowicki, Snijders, JASA (2001)

Community detection as model selection

Each partition ***M=b*** of the nodes into groups amounts to a different model of our data, that is, our observed network ***D=A^o***

$$p(b|A^o) = \frac{\mathrm{e}^{-\mathcal{L}(b,A^o)}}{Z}$$

with:

$$\mathcal{L}\left(b,\ A^o\right) = -\log \int_Q dQ\, p(A^o|b,Q)\, p(Q|b)\, p(b)$$

$$p(A^o|b,Q) =$$

$$p(Q|b) =$$

$$p(b) =$$

$$\mathcal{L}\left(b,\ A^o\right) = -\log \int_Q dQ\, p(A^o|b, Q)\, p(Q|b)\, p(b)$$

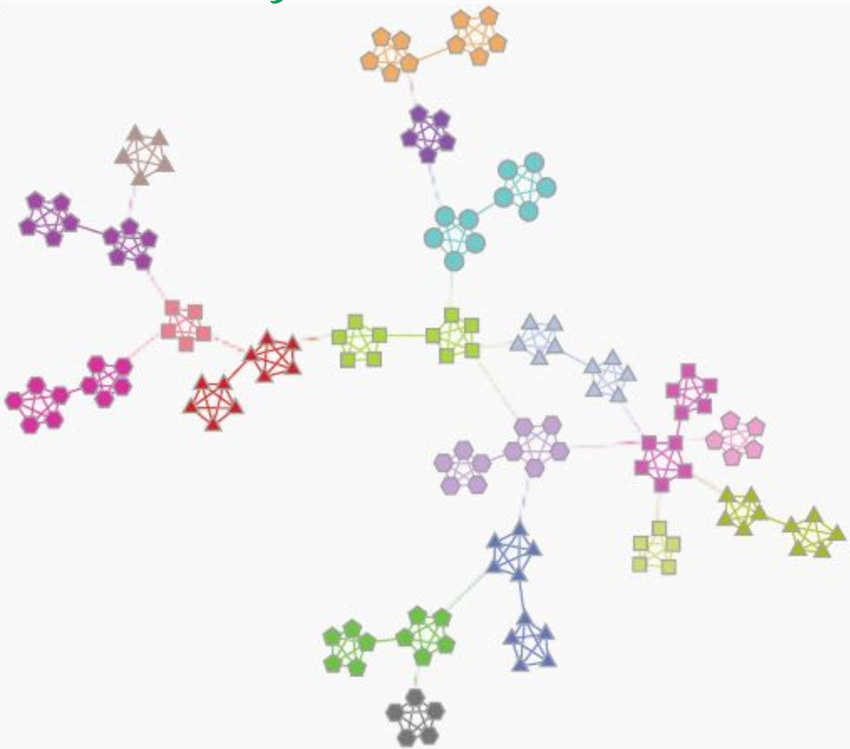
Guimera, Sales-Pardo, Proc. Natl. Acad. Sci. USA (2009)

$$\mathcal{L}(b, A^o) = -\sum_{\alpha,\beta} \log \frac{n^1_{\alpha\beta}!\ \ n^0_{\alpha\beta}!}{\left(n^1_{\alpha\beta} + n^0_{\alpha\beta} + 1\right)!} + C$$

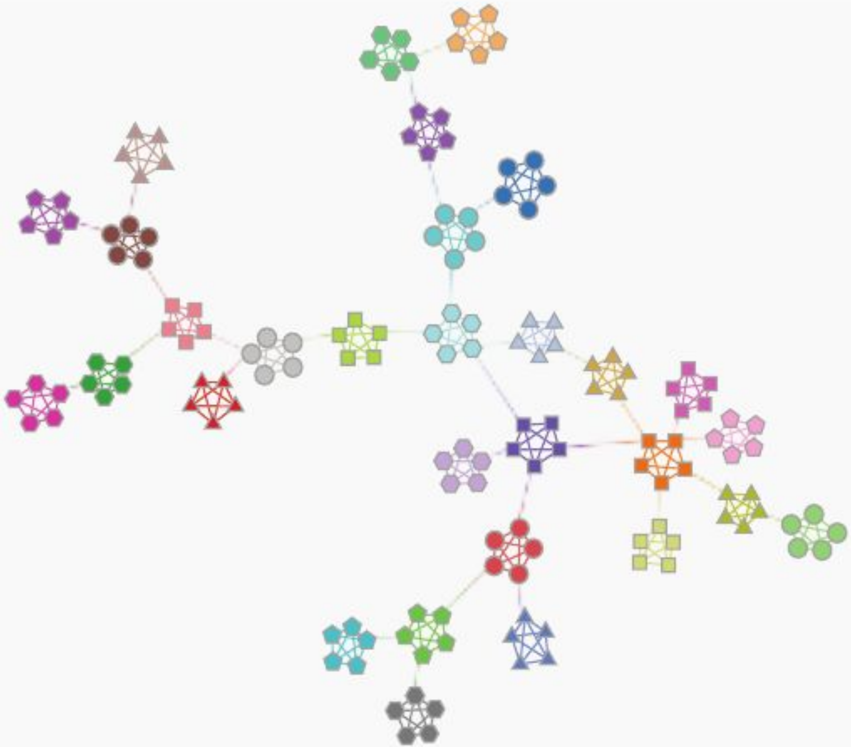
Guimera, Sales-Pardo, Proc. Natl. Acad. Sci. USA (2009)

Inferential approaches are preferable to "descriptive" approaches such as modularity maximization

Modularity maximization



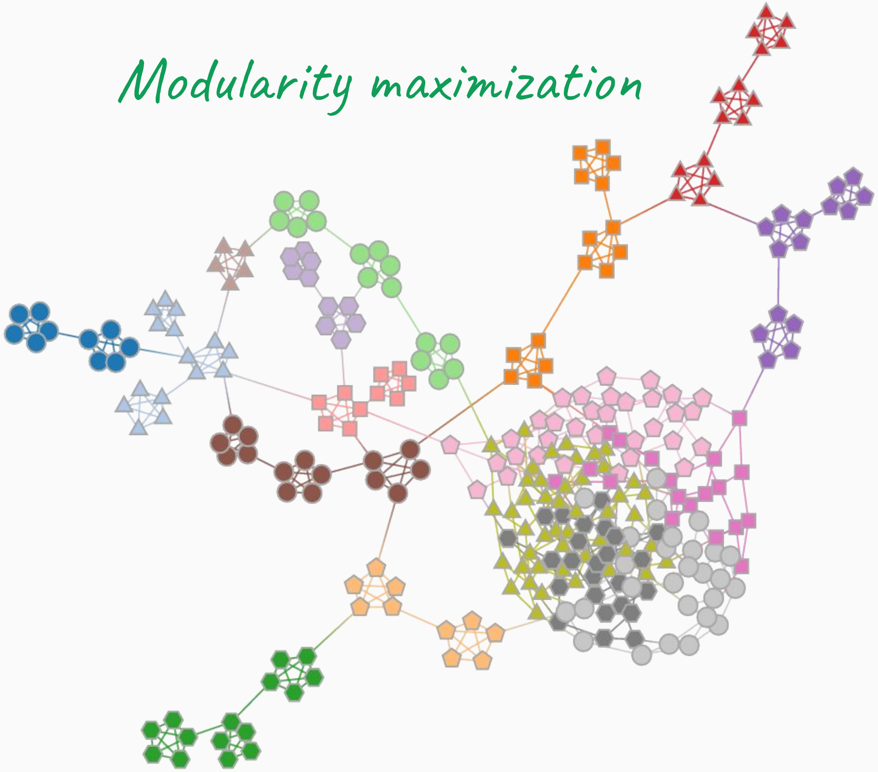
Minimum description length



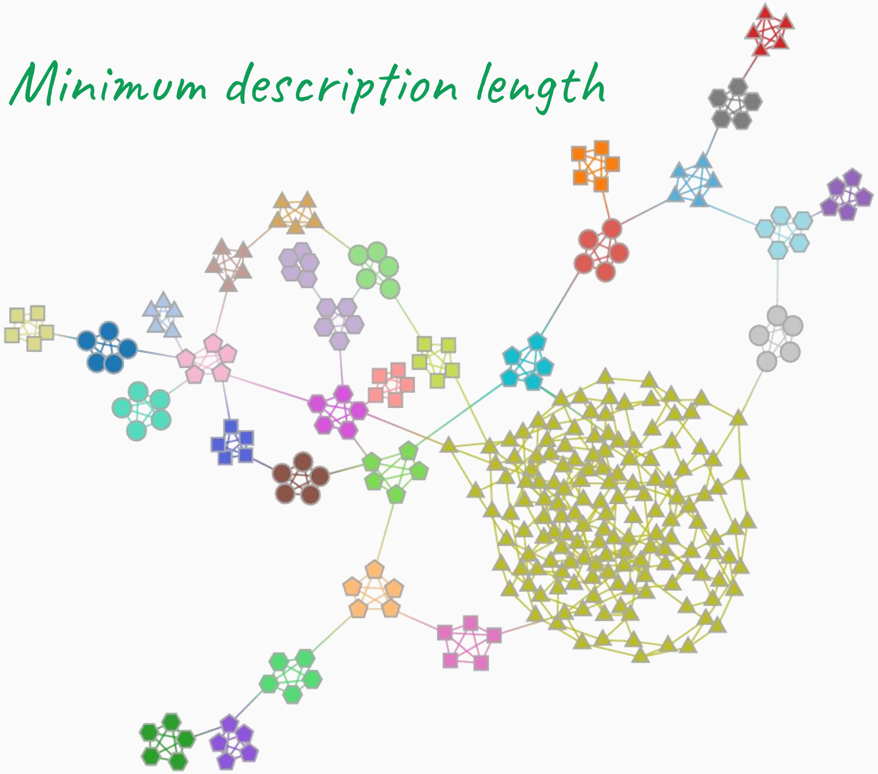
Peixoto, [Descriptive vs. Inferential Community Detection in Networks]() (2023)

Inferential approaches are preferable to "descriptive" approaches such as modularity maximization

Modularity maximization



Minimum description length



Peixoto, [Descriptive vs. Inferential Community Detection in Networks]() (2023)

Plan for the lecture

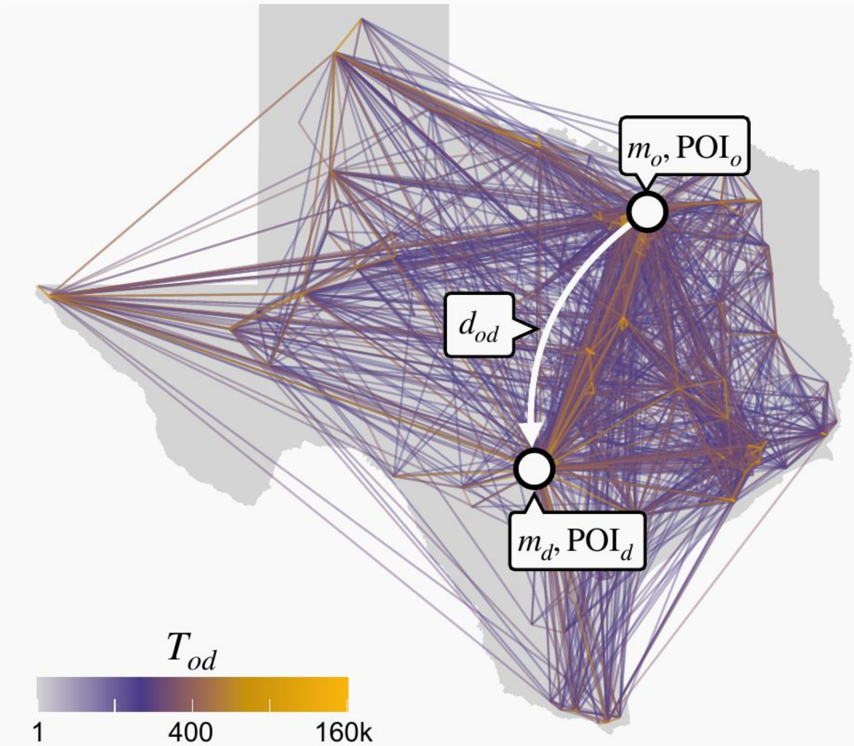
Theory of Bayesian inference:

- Bayesian interpretation of probabilities
- Bayesian model selection
- Bayesian prediction
- Markov chain Monte Carlo (MCMC)

Applications:

- Inferential community detection in complex networks
- Bayesian machine scientist

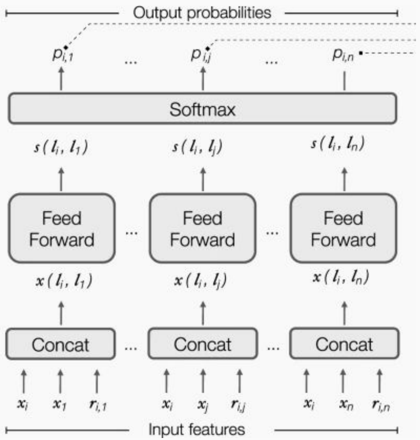
Can we find models that predict human mobility flows?



Gravity models

$$T_{od} = A\,\frac{m_o\, m_d}{d^{\alpha}}$$

“Deep gravity” models



Cabanas *et al.*, submitted (2024)

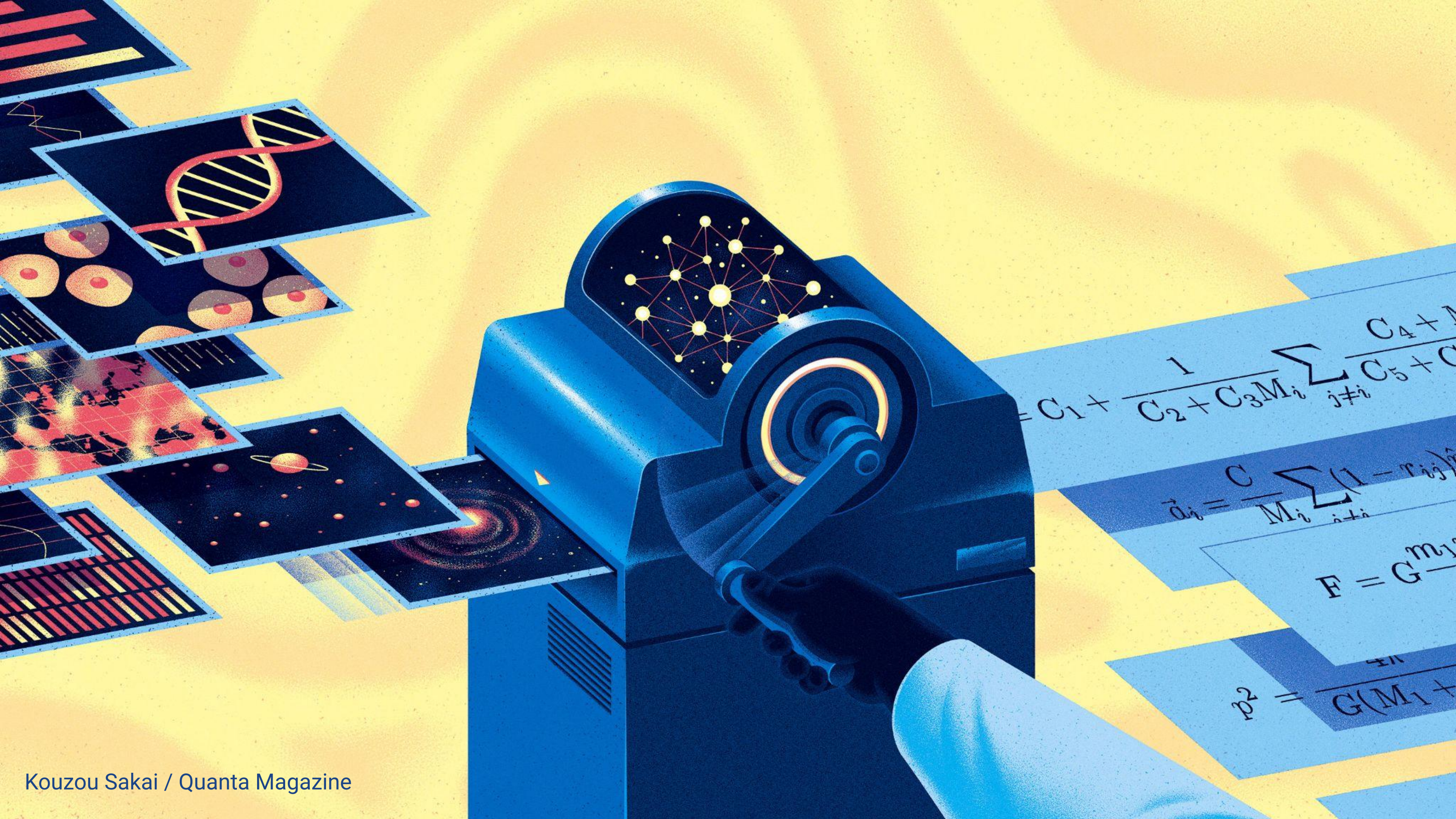
$$y=f(x,\theta)$$

Can we design a “machine scientist” that automates the task of building **closed-form mathematical models** from data?

$$y=f(x,\theta)$$

Can we design a "machine scientist" that automates the task of building **closed-form mathematical models** from data?

$$f(x) = a_0 + a_1 x \qquad f(x) = \log\left(\sin\left(\exp\left(x^{-8}\right)\right)\right)$$



Kouzou Sakai / Quanta Magazine

$$y=f(x,\theta)$$

$$p(f \mid \{x, y\})$$

This posterior over expressions/models encapsulates the full probabilistic solution to the symbolic regression problem

The most plausible model has the shortest description length (compresses the data optimally)

The posterior can be rewritten as

$$
\begin{aligned}
p(f|D) &= \frac{1}{p(D)} \int_{\Theta} d\theta \, p(D|f,\theta) \, p(\theta|f) \, p(f) \\
&= \frac{\mathrm{e}^{-\mathcal{L}(f,D)}}{p(D)}
\end{aligned}
$$

The most plausible model has the shortest description length (compresses the data optimally)

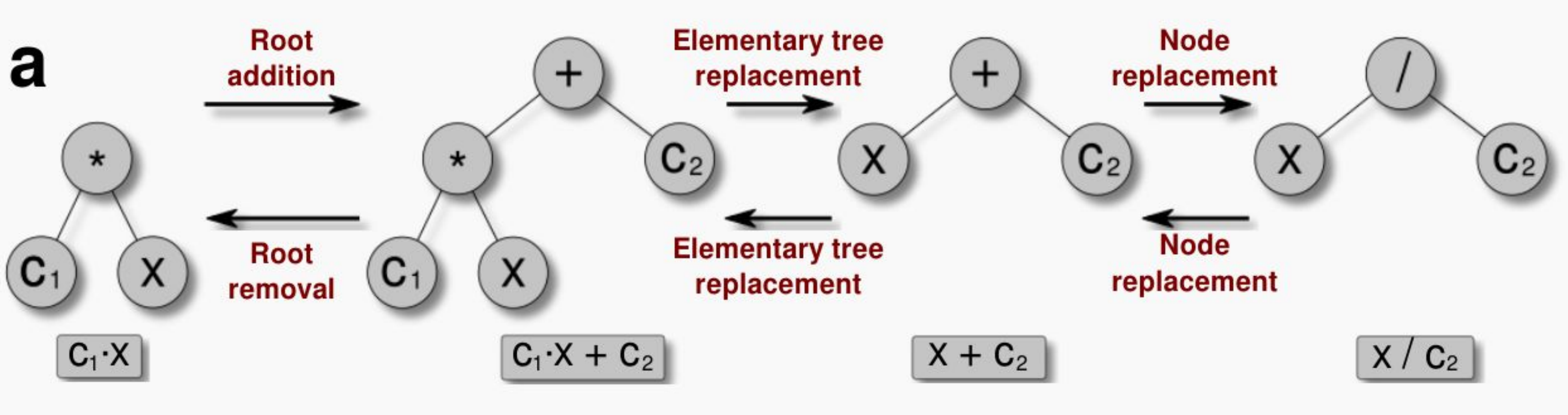
The posterior can be rewritten as

$$\begin{aligned} p(f|D) &= \frac{1}{p(D)} \int_\Theta d\theta\, p(D|f,\theta)\, p(\theta|f)\, p(f) \\ &= \frac{\mathrm{e}^{-\mathcal{L}(f,D)}}{p(D)} \end{aligned}$$

And the ***description length*** can be approximated as

$$\mathcal{L}(f,D) = \underbrace{\frac{B(f)}{2}}_{\text{BIC}} - \underbrace{\log p(f)}_{\text{prior}}$$

Exploring the space of models
A Metropolis-Hastings algorithm for sampling mathematical expressions



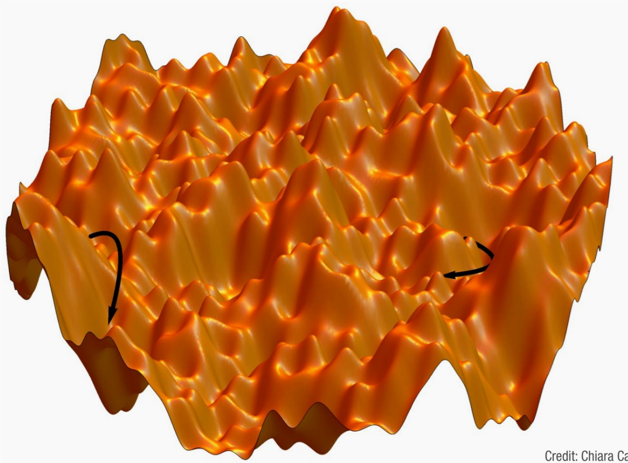
Guimera et al., *Science Advances* (2020)

All in all, we have defined our Bayesian machine scientist

It establishes the plausibility of any model by means of the posterior (i.e. description length)

$$\mathcal{L}(M, D) = \frac{B(M)}{2} - \log p(M)$$

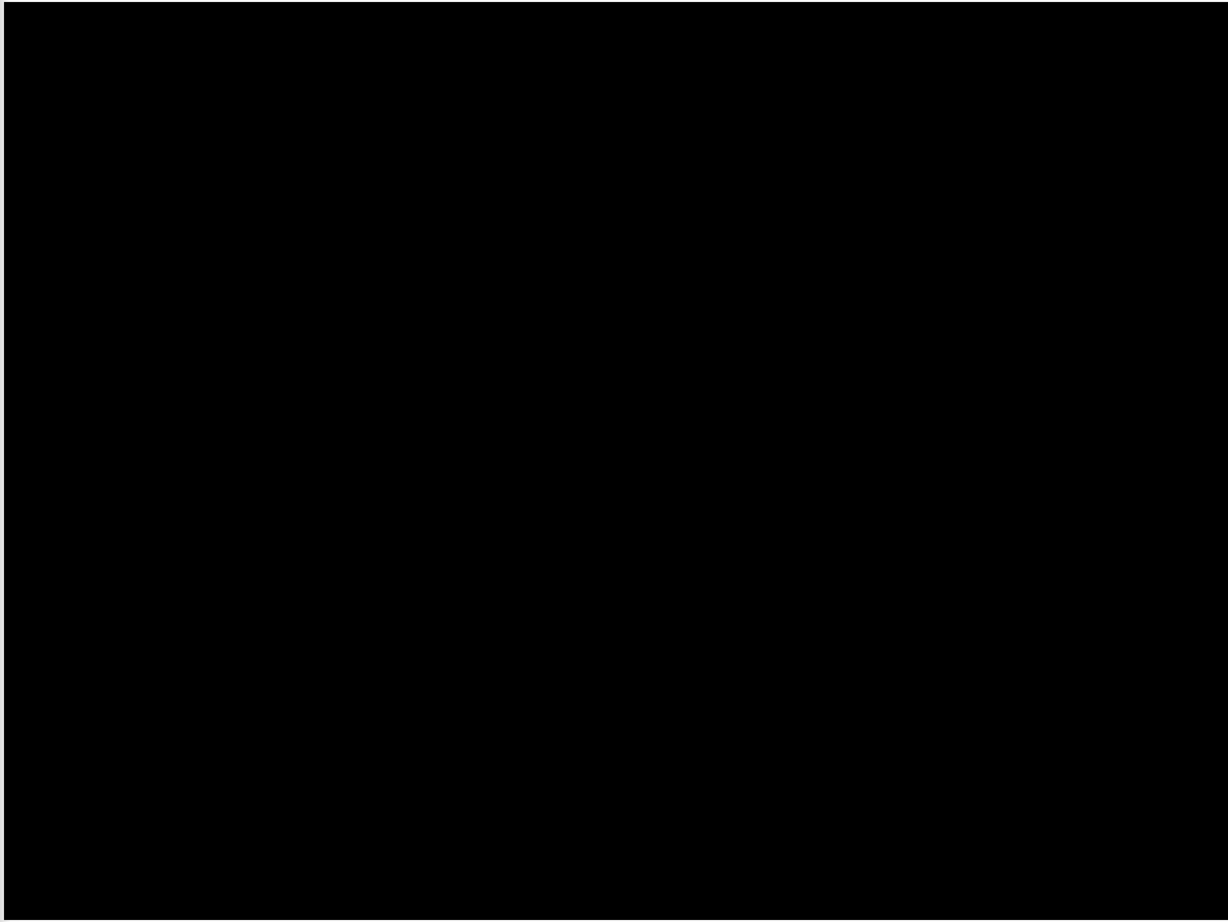
It explores the space of models and samples models from their posterior using Metropolis-Hastings



Credit: Chiara Cammarota

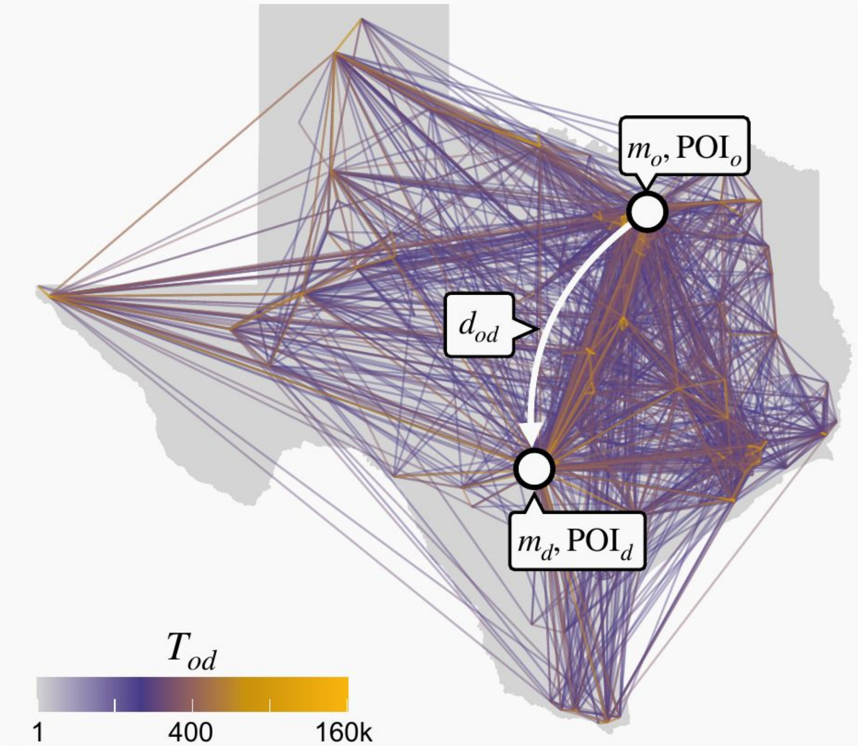
So, does it work?

We generate synthetic data and see if the machine scientist is able to recover the correct model



Guimera et al., *Science Advances* (2020)

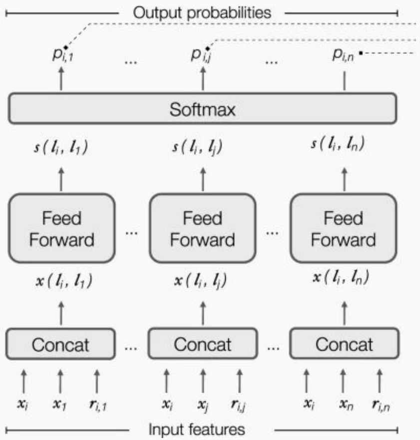
Can we find models that predict human mobility flows?



Gravity models

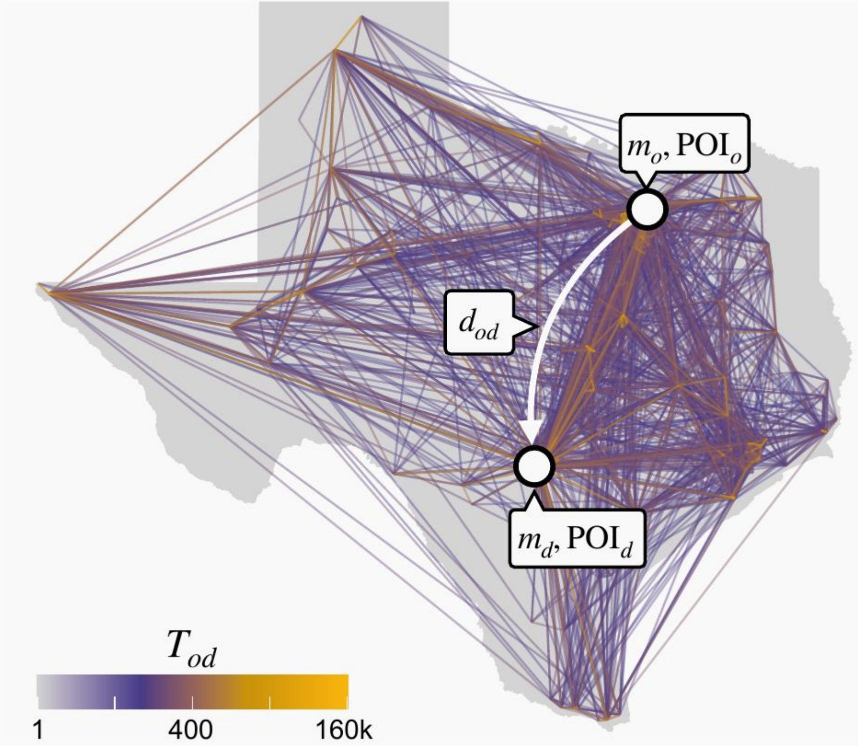
$$T_{od} = A \frac{m_o m_d}{d^\alpha}$$

"Deep gravity" models



Cabanas *et al.*, submitted (2024)

Can we find models that predict human mobility flows?



A $$\log T_{od} = A\left(1 + \frac{B((m_d+C)(m_o+D))^{\beta}}{d}\right)^{\xi}$$

B $$\log T_{od} = \log\left(A\left(\frac{B(m_d m_o + C m_d + D)}{d^{\alpha}} + 1\right)^{\gamma}\right)$$

Cabanas *et al.*, submitted (2024)

Wrapping up with a bit of wisdom

So, thanks to Cox, it was now a theorem that any set of rules for conducting inference, in which we represent degrees of plausibility by real numbers, is necessarily either equivalent to the Laplace–Jeffreys rules, or inconsistent. The reason for their pragmatic success is then pretty clear. Those who continued to oppose Bayesian methods after 1946 have been obliged to ignore not only the pragmatic success, but also the theorem.

E. T. Jaynes (1985)

Thank you

More information:

http://seeslab.info

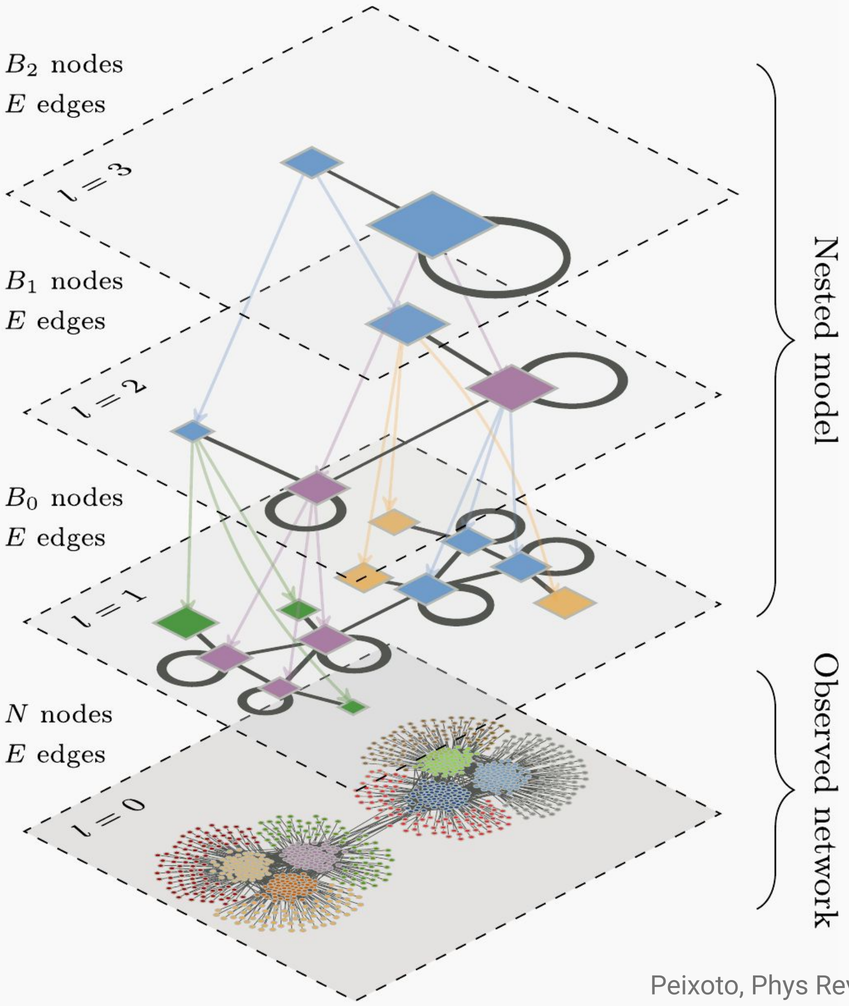
@sees_lab

Inferential approaches are not limited to the vanilla stochastic block model

There are many variations, and also non-group-based models, amenable to inferential/probabilistic treatment

Hierarchical priors (nested stochastic block model)

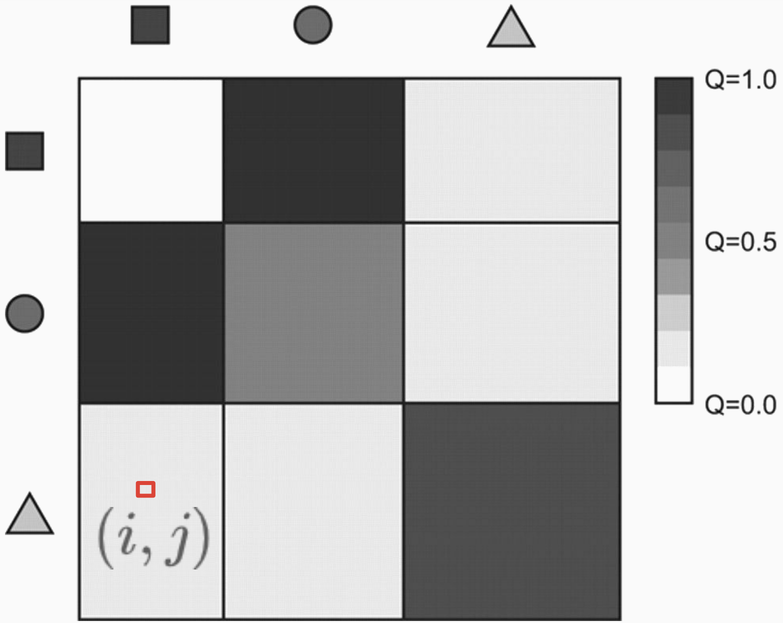
Not all partitions are equally plausible a priori



Peixoto, Phys Rev X (2014)

The degree-corrected stochastic block model

We assume that nodes belong to groups, and their interactions depend only on those groups ***and each node's overall propensity to make connections***



$$p(A_{ij} = 1 | Q, \Theta) = \theta_i \theta_j q_{\square\triangle}$$

Karrer, Newman, Phys Rev E (2011)

Stochastic block models for multilayer and temporal networks

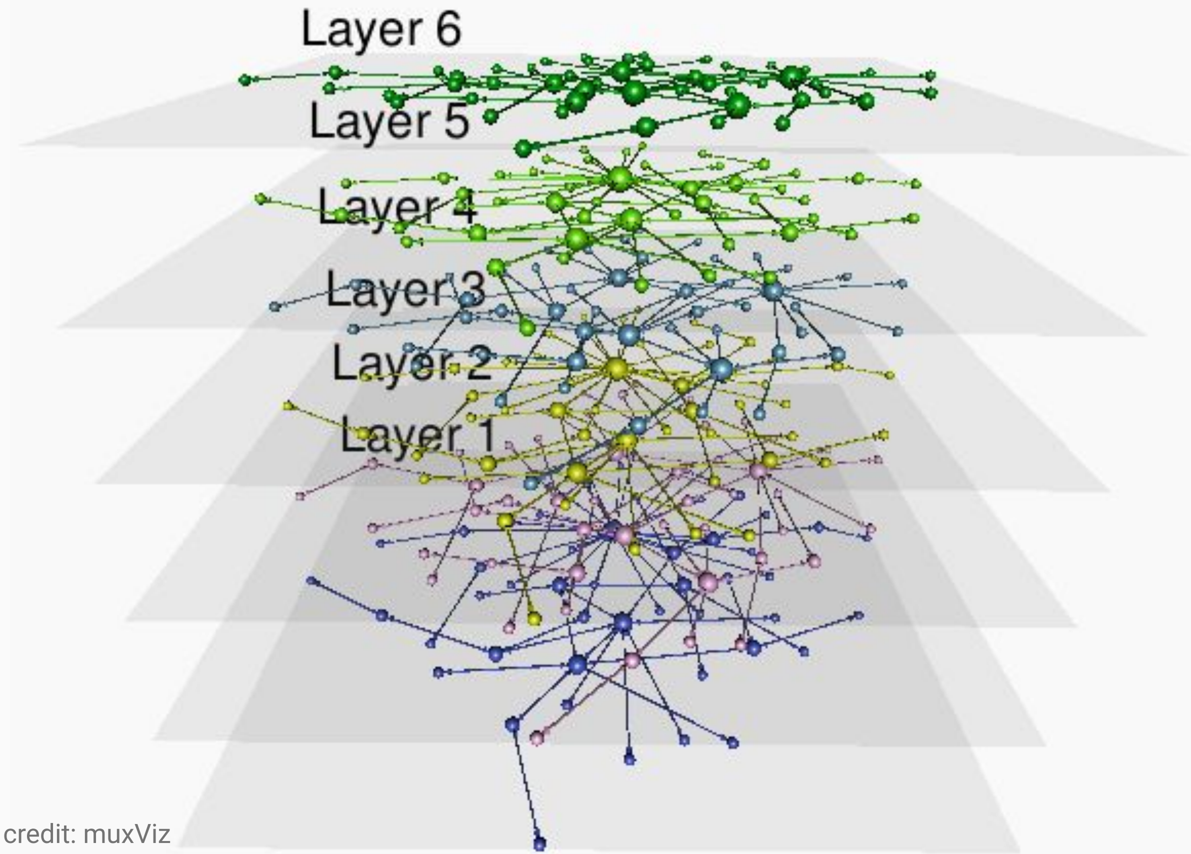


Image credit: muxViz

Vallès-Català et al., Phys Rev X (2016)
Peixoto, Rosvall, Nat Comm (2017)
Tarrés-Deulofeu et al., Phys Rev E (2019)

The mixed-membership stochastic block model

Nodes do not belong to a single group, but rather to a ***mixture of groups***

Each node *i* belongs to group *r* with probability b_{ir} such that

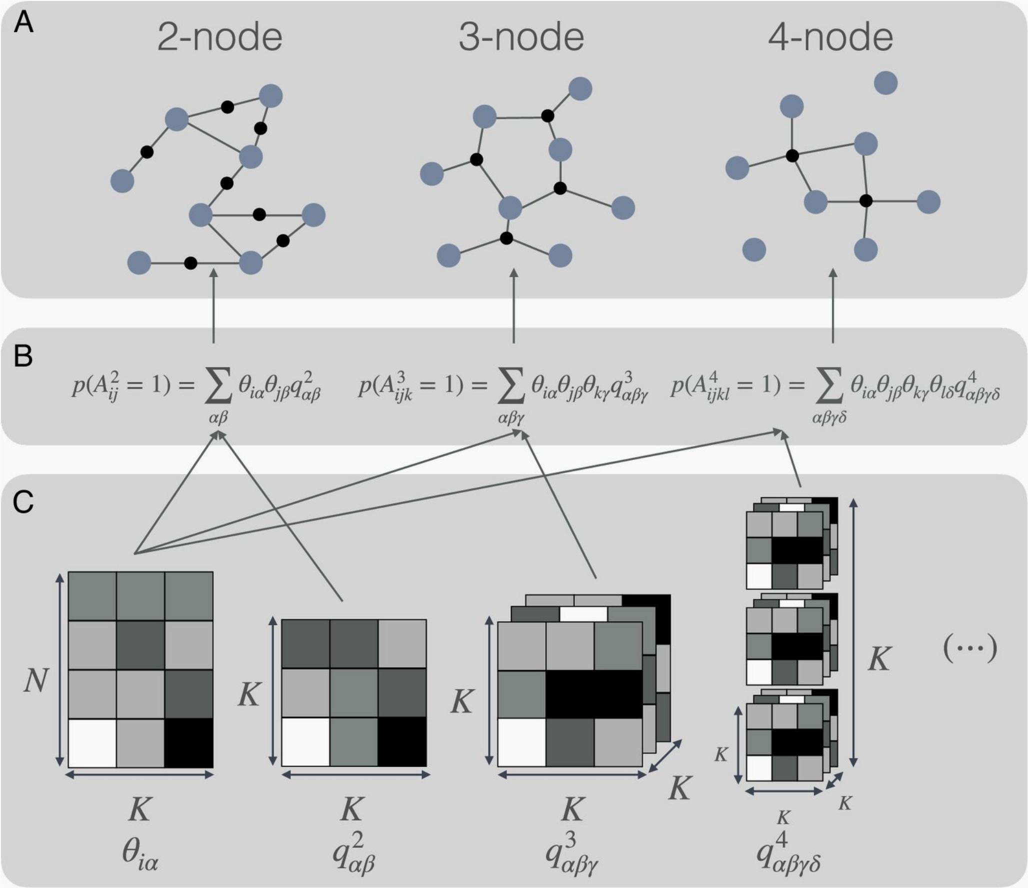
$$\sum_r b_{ir} = 1$$

Then, nodes *i* and *j* are connected with probability

$$p(A_{ij} = 1 | B, Q) = \sum_{rs} b_{ir} b_{js} q_{rs}$$

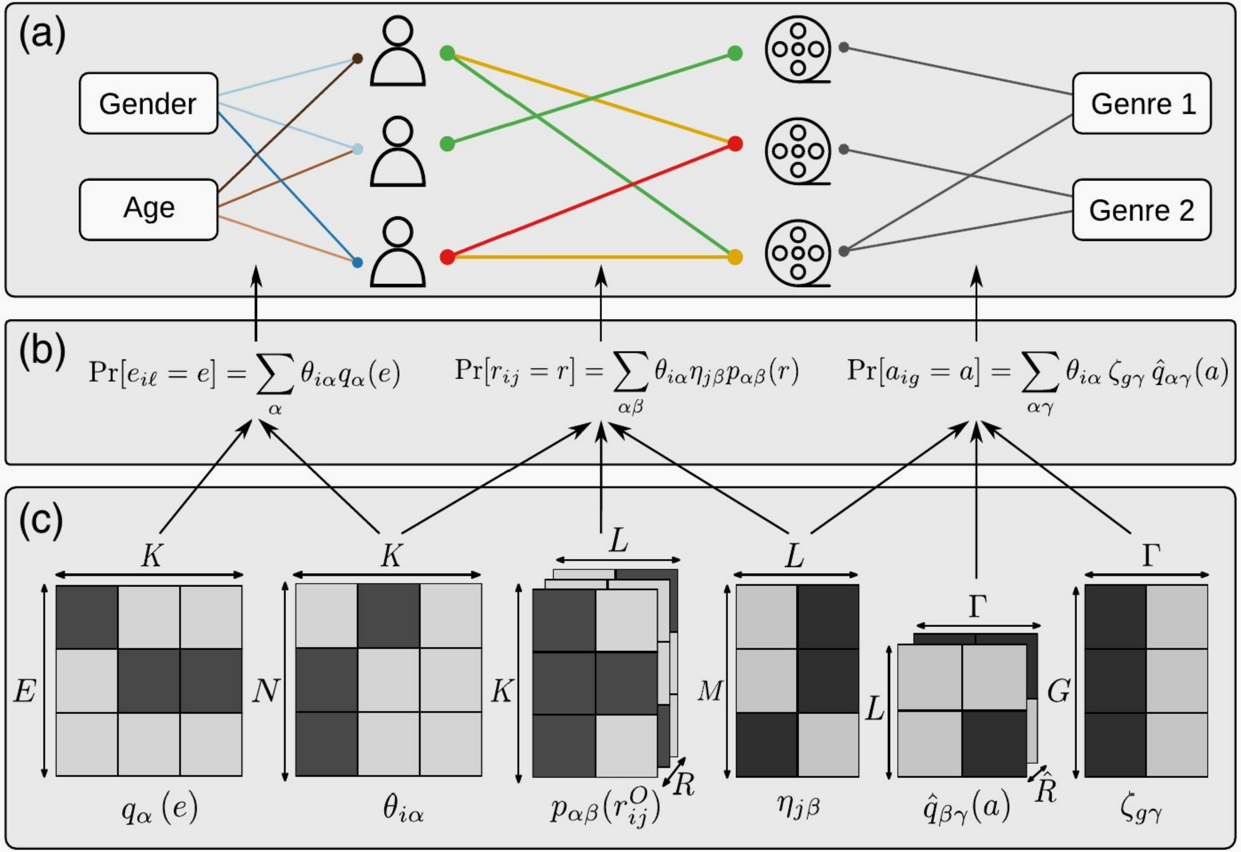
Airoldi et al., J Mach Learn Res (2008)
Godoy-Lorite, et al., Proc Natl Acad Sci USA (2016)

Mixed-membership stochastic block model for higher-order interactions



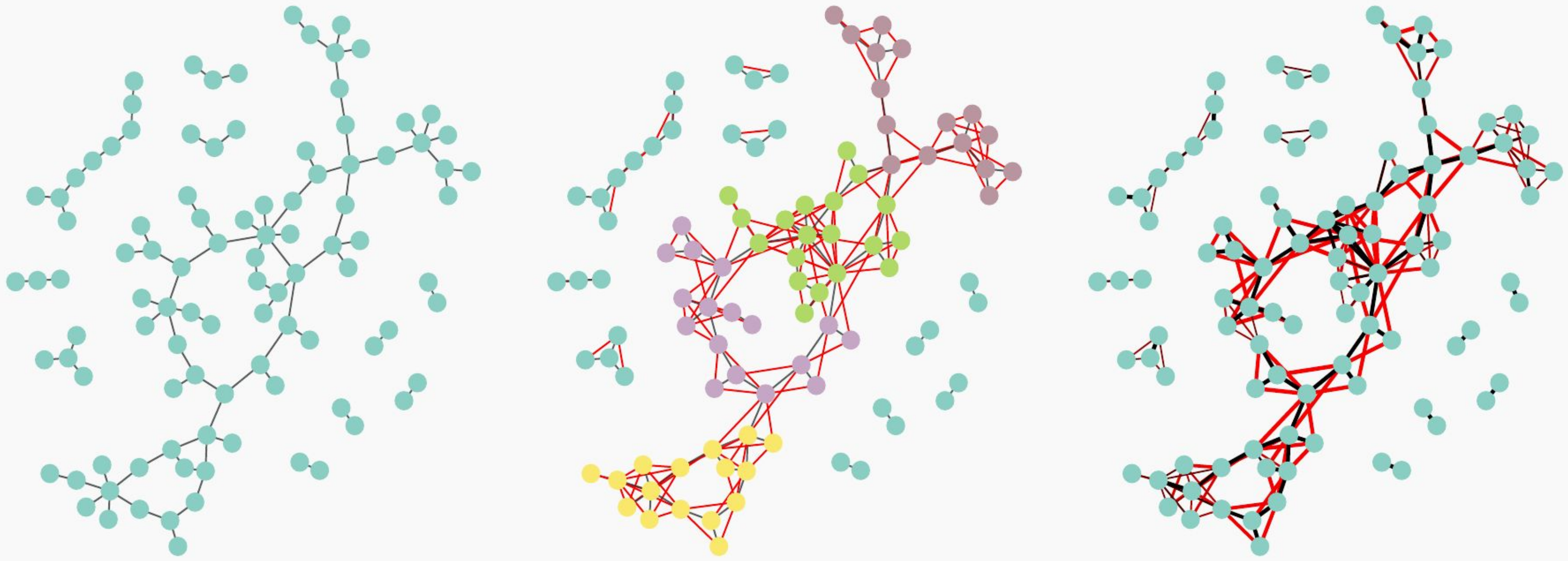
Sales-Pardo, et al., Proc Natl Acad Sci USA (2023)

Mixed-membership stochastic block model with node metadata



Fajardo-Fontiveros, et al., Phys Rev X (2022)

Generative models for non-group mechanisms (for example, triadic closure)



(a) Random seminal edges

(b) Triadic closure edges and spurious communities found with SBM ($\Sigma_{\mathrm{SBM}} = 801.7$ nats)

(c) Inference of the SBM/TC model ($\Sigma_{\mathrm{SBM/TC}} = 590.6$ nats)

Peixoto, Phys Rev X (2022)