bayesian inference: principles and practice

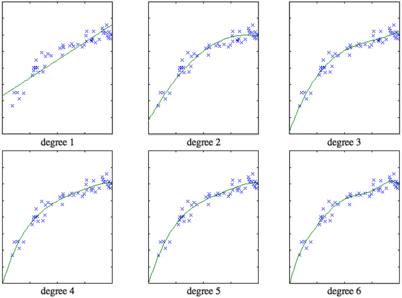
jake hofman

http://jakehofman.com

july 9, 2009

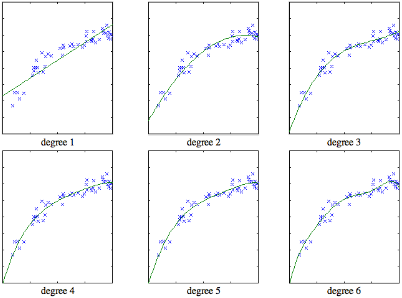
motivation

- would like models that:
  - provide predictive and explanatory power
  - are complex enough to describe observed phenomena
  - are simple enough to generalize to future observations



motivation

- would like models that:
  - provide predictive and explanatory power
  - are complex enough to describe observed phenomena
  - are simple enough to generalize to future observations



- claim: bayesian inference provides a systematic framework to infer such models from observed data

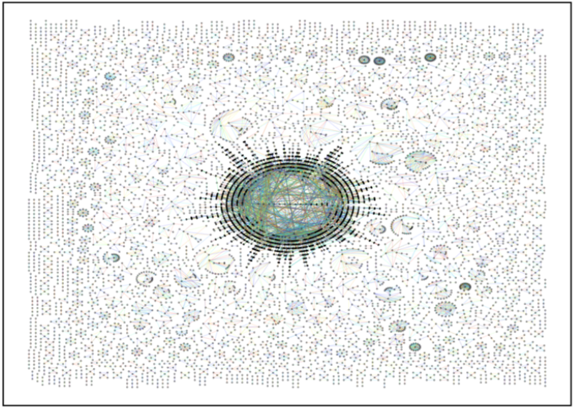
motivation

- principles behind bayesian interpretation of probability and bayesian inference are well established (bayes, laplace, etc., 18th century)

+

- recent advances in mathematical techniques and computational resources have enabled successful applications of these principles to real-world problems

motivation: a bayesian approach to network modularity



outline

1. principles (what we'd like to do)
   - background: joint, marginal, and conditional probabilities
   - bayes' theorem: inverting conditional probabilities
   - bayesian probability: unknowns as random variables
   - bayesian inference: bayesian probability + bayes' theorem
2. practice (what we're able to do)
   - monte carlo methods: representative samples
   - variational methods: bound optimization
   - references

joint, marginal, and conditional probabilities

joint distribution

$p_{XY}(X = x, Y = y)$: probability $X = x$ *and* $Y = y$

conditional distribution

$p_{X|Y}(X = x|Y = y)$: probability $X = x$ *given* $Y = y$

marginal distribution

$p_X(X)$: probability $X = x$ (*regardless of* Y)

sum and product rules

sum rule

sum out settings of irrelevant variables:

$$p(x) = \sum_{y \in \Omega_Y} p(x, y) \tag{1}$$

product rule

the joint as the product of the conditional and marginal:

$$\begin{aligned} p(x, y) &= p(x|y)\, p(y) && (2) \\ &= p(y|x)\, p(x) && (3) \end{aligned}$$

outline

1. principles (what we'd like to do)
   - background: joint, marginal, and conditional probabilities
   - bayes' theorem: inverting conditional probabilities
   - bayesian probability: unknowns as random variables
   - bayesian inference: bayesian probability + bayes' theorem
2. practice (what we're able to do)
   - monte carlo methods: representative samples
   - variational methods: bound optimization
   - references

inverting conditional probabilities

equate far right- and left-hand sides of product rule

$$p(y|x)\,p(x) = p(x,y) = p(x|y)\,p(y) \tag{4}$$

and divide:

bayes' theorem (bayes and price 1763)

the probability of Y given X from the probability of X given Y:

$$p(y|x) = \frac{p(x|y)\,p(y)}{p(x)} \tag{5}$$

where $p(x) = \sum_{y \in \Omega_Y} p(x|y)\,p(y)$ is the normalization constant

example: diagnoses a la bayes

- population 10,000
- 1% has (rare) disease
- test is 99% (relatively) effective, i.e.
    - given a patient is sick, 99% test positive
    - given a patient is healthy, 99% test negative

example: diagnoses a la bayes

- population 10,000
- 1% has (rare) disease
- test is 99% (relatively) effective, i.e.
  - given a patient is sick, 99% test positive
  - given a patient is healthy, 99% test negative
- given positive test, what is probability the patient is sick?[1]

[1] follows wiggins (2006)

example: diagnoses a la bayes

- 99 sick patients test positive, 99 healthy patients test positive

example: diagnoses a la bayes

- 99 sick patients test positive, 99 healthy patients test positive
- given positive test, 50% probability that patient is sick

example: diagnoses a la bayes

- know probability of testing positive/negative given sick/healthy
- use bayes' theorem to "invert" to probability of sick/healthy given positive/negative test

$$p\left(sick|test\ +\right) = \frac{\overbrace{p\left(test\ +|sick\right)}^{99/100}\overbrace{p\left(sick\right)}^{1/100}}{\underbrace{p\left(test\ +\right)}_{198/100^2}} = \frac{99}{198} = \frac{1}{2} \quad (6)$$

example: diagnoses a la bayes

- know probability of testing positive/negative given sick/healthy
- use bayes' theorem to "invert" to probability of sick/healthy given positive/negative test

$$p\left(sick|test\ +\right) = \frac{\overbrace{p\left(test\ +|sick\right)}^{99/100}\overbrace{p\left(sick\right)}^{1/100}}{\underbrace{p\left(test\ +\right)}_{198/100^2}} = \frac{99}{198} = \frac{1}{2} \quad (6)$$

- most "work" in calculating denominator (normalization)

outline

interpretations of probabilities
(just enough philosophy)

- frequentists: limit of relative frequency of events for large number of trials
- bayesians: measure of a state of knowledge, quantifying degrees of belief (jaynes 2003)

interpretations of probabilities
(just enough philosophy)

- frequentists: limit of relative frequency of events for large number of trials
- bayesians: measure of a state of knowledge, quantifying degrees of belief (jaynes 2003)
- key difference: bayesians permit assignment of probabilities to unknown/unobservable hypotheses (frequentists do not)

interpretations of probabilities
(just enough philosophy)

- e.g., inferring model parameters Θ from observed data $\mathcal{D}$:
  - frequentist approach: calculate parameter setting that maximizes likelihood of data (point estimate),

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, p(\mathcal{D}|\Theta) \tag{7}$$

  - bayesian approach: calculate distribution over parameter settings given data,

$$p(\Theta|\mathcal{D}) = ? \tag{8}$$

outline

1. principles (what we'd like to do)
   - background: joint, marginal, and conditional probabilities
   - bayes' theorem: inverting conditional probabilities
   - bayesian probability: unknowns as random variables
   - bayesian inference: bayesian probability + bayes' theorem
2. practice (what we're able to do)
   - monte carlo methods: representative samples
   - variational methods: bound optimization
   - references

bayesian probability + bayes' theorem

- bayesian inference:
  - treat unknown quantities as random variables
  - use bayes' theorem to systematically update prior knowledge in the presence of observed data

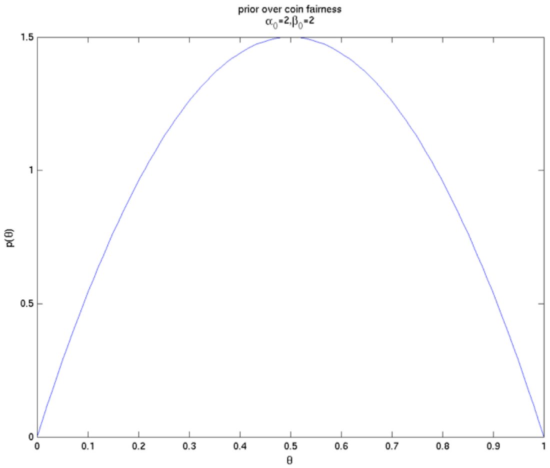
$$\overbrace{p(\Theta|\mathcal{D})}^{posterior} = \frac{\overbrace{p(\mathcal{D}|\Theta)}^{likelihood}\,\overbrace{p(\Theta)}^{prior}}{\underbrace{p(\mathcal{D})}_{evidence}} \tag{9}$$

example: coin flipping

- observe independent coin flips (bernoulli trials)
- infer distribution over coin bias

example: coin flipping

prior $p(\Theta)$ over coin bias before observing flips

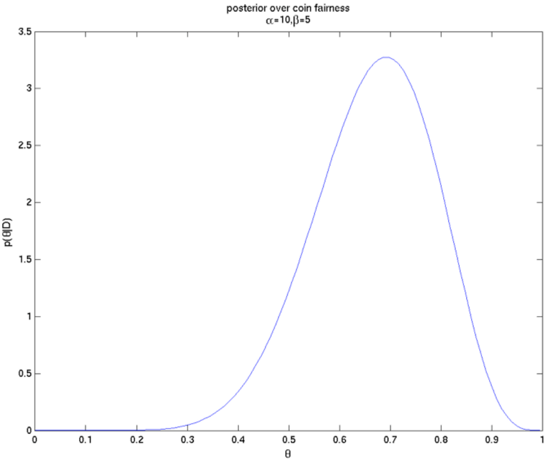


example: coin flipping

observe flips: HTHHHTTHHHH

example: coin flipping

update posterior $p(\Theta|\mathcal{D})$ using bayes' theorem

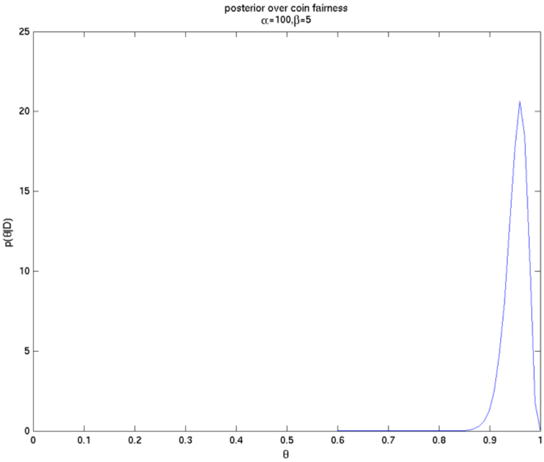


example: coin flipping

observe flips: HHHHHHHHHHHHTHHHHHHHHHH
HHHHHHHHHHHHHHHHHHHHHHH
HHHHHHTHHHHHHHHHHHHHHHH
HHHHHHHHHHHHHHHHHHHHHHH HHHHHHHHT

example: coin flipping

update posterior $p(\Theta|\mathcal{D})$ using bayes' theorem



quantities of interest

- bayesian inference maintains full posterior distributions over unknowns
- many quantities of interest require expectations under these posteriors, e.g. posterior mean and predictive distribution:

$$\bar{\Theta} = \mathbb{E}_{p(\Theta|\mathcal{D})}\left[\Theta\right] = \int d\Theta \; \Theta \; p(\Theta|\mathcal{D}) \tag{10}$$

$$p\left(x|\mathcal{D}\right) = \mathbb{E}_{p(\Theta|\mathcal{D})}\left[p\left(x|\Theta,\mathcal{D}\right)\right] = \int d\Theta \; p\left(x|\Theta,\mathcal{D}\right) p(\Theta|\mathcal{D}) \tag{11}$$

quantities of interest

- bayesian inference maintains full posterior distributions over unknowns
- many quantities of interest require expectations under these posteriors, e.g. posterior mean and predictive distribution:

$$\bar{\Theta} = \mathbb{E}_{p(\Theta|\mathcal{D})}\left[\Theta\right] = \int d\Theta \; \Theta \; p(\Theta|\mathcal{D}) \tag{10}$$

$$p\left(x|\mathcal{D}\right) = \mathbb{E}_{p(\Theta|\mathcal{D})}\left[p\left(x|\Theta,\mathcal{D}\right)\right] = \int d\Theta \; p\left(x|\Theta,\mathcal{D}\right) p(\Theta|\mathcal{D}) \tag{11}$$

- often can't compute posterior (normalization), let alone expectations with respect to it $\rightarrow$ approximation methods

outline

1. principles (what we'd like to do)
   - background: joint, marginal, and conditional probabilities
   - bayes' theorem: inverting conditional probabilities
   - bayesian probability: unknowns as random variables
   - bayesian inference: bayesian probability + bayes' theorem
2. practice (what we're able to do)
   - monte carlo methods: representative samples
   - variational methods: bound optimization
   - references

representative samples

- general approach: approximate intractable expectations via sum over representative samples[2]

$$\Phi = \mathbb{E}_{p(x)}\left[\phi(x)\right] = \int dx \underbrace{\phi(x)}_{arbitrary\ \ function} \underbrace{p(x)}_{target\ \ density} \tag{12}$$

[2]follows mackay (2003), including stolen images

representative samples

- general approach: approximate intractable expectations via sum over representative samples[2]

$$\Phi = \mathbb{E}_{p(x)}\left[\phi(x)\right] = \int dx \underbrace{\phi(x)}_{arbitrary\ \ function} \underbrace{p(x)}_{target\ \ density} \tag{12}$$

$$\Downarrow$$

$$\widehat{\Phi} = \frac{1}{R}\sum_{r=1}^{R} \phi(x^{(r)}) \tag{13}$$

[2] follows mackay (2003), including stolen images

representative samples

- general approach: approximate intractable expectations via sum over representative samples[2]

$$\Phi = \mathbb{E}_{p(x)}\left[\phi(x)\right] = \int dx \underbrace{\phi(x)}_{arbitrary\ function} \underbrace{p(x)}_{target\ density} \tag{12}$$

$$\Downarrow$$

$$\widehat{\Phi} = \frac{1}{R}\sum_{r=1}^{R}\phi(x^{(r)}) \tag{13}$$

- shifts problem to finding "good" samples

[2] follows mackay (2003), including stolen images

representative samples

- further complication: in general we can only evaluate the target density to within a multiplicative (normalization) constant, i.e.

$$p(x) = \frac{p^*(x)}{Z} \tag{14}$$

and $p^*(x^{(r)})$ can be evaluated with Z unknown

sampling methods

- monte carlo methods
  - uniform sampling
  - importance sampling
  - rejection sampling
  - . . .
- markov chain monte carlo (mcmc) methods
  - metropolis-hastings
  - gibbs sampling
  - . . .

uniform sampling

- sample uniformly from state space of all x values
- evaluate non-normalized density $p^*(x^{(r)})$ at each $x^{(r)}$
- approximate normalization constant as

$$Z_R = \sum_{r=1}^{R} p^*(x^{(r)}) \tag{15}$$

- estimate expectation as

$$\widehat{\Phi} = \sum_{r=1}^{R} \phi(x^{(r)}) \frac{p^*(x^{(r)})}{Z_R} \tag{16}$$

uniform sampling

- sample uniformly from state space of all x values
- evaluate non-normalized density $p^*(x^{(r)})$ at each $x^{(r)}$
- approximate normalization constant as

$$Z_R = \sum_{r=1}^{R} p^*(x^{(r)}) \tag{15}$$

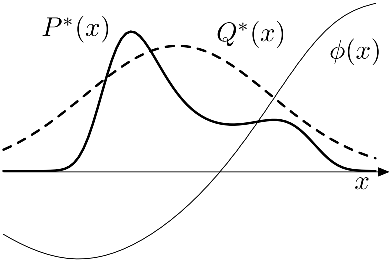
- estimate expectation as

$$\widehat{\Phi} = \sum_{r=1}^{R} \phi(x^{(r)}) \frac{p^*(x^{(r)})}{Z_R} \tag{16}$$

- requires prohibitively large number of samples in high dimensions with concentrated density

importance sampling

- modify uniform sampling by introducing a sampler density $q(x) = \frac{q^*(x)}{Z_Q}$
- choose $q(x)$ simple enough that $q^*(x)$ can be sampled from, with hope that $q^*(x)$ is a reasonable approximation to $p^*(x)$



importance sampling

- adjust estimator by weighting "importance" of each sample

$$\widehat{\Phi} = \frac{\sum_{r=1}^{R} w_r \phi(x^{(r)})}{\sum_R w_r} \tag{17}$$

where

$$w_r = \frac{p^*(x^{(r)})}{q^*(x^{(r)})} \tag{18}$$

importance sampling

- adjust estimator by weighting "importance" of each sample

$$\hat{\Phi} = \frac{\sum_{r=1}^{R} w_r \phi(x^{(r)})}{\sum_R w_r} \tag{17}$$

where

$$w_r = \frac{p^*(x^{(r)})}{q^*(x^{(r)})} \tag{18}$$

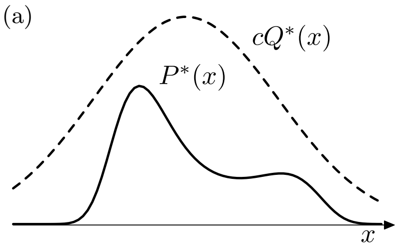
- difficult to choose "good" $q^*(x)$ as well as estimate reliability of estimator

rejection sampling

- similar to importance sampling, but proposal density strictly bounds target density, i.e.

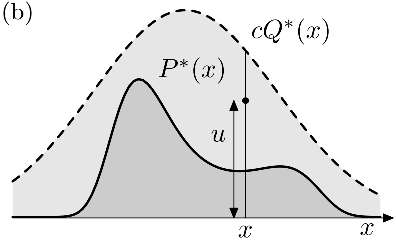
$$cq^*(x) > p^*(x), \tag{19}$$

for some known value c and all x



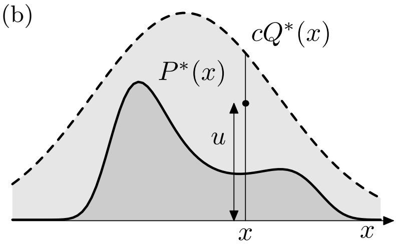
rejection sampling

- generate sample x from $q^*(x)$
- generate uniformly random number u from $[0, cq^*(x)]$
- add x to set $\{x^{(r)}\}$ if $p^*(x) > u$
- estimate expectation as $\widehat{\Phi} = \frac{1}{R}\sum_{r=1}^{R} \phi(x^{(r)})$



rejection sampling

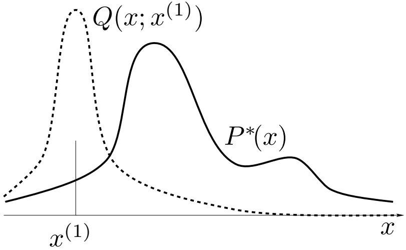
- generate sample x from $q^*(x)$
- generate uniformly random number u from $[0, cq^*(x)]$
- add x to set $\{x^{(r)}\}$ if $p^*(x) > u$
- estimate expectation as $\widehat{\Phi} = \frac{1}{R}\sum_{r=1}^{R} \phi(x^{(r)})$



- c often prohibitively large for poor choice of $q^*(x)$ or high dimensions

metropolis-hastings

- use a local proposal density $q(x'; x^{(t)})$, depending on current state $x^{(t)}$



- construct markov chain through state space, converging to target density
- note: proposal density needn't closely approximate target density

metropolis-hastings

- at time t, generate tenative state x' from $q(x'; x^{(t)})$
- evaluate

$$a = \frac{p^*(x^{(t)})}{p^*(x')} \frac{q(x^{(t)}; x')}{q(x'; x^{(t)})} \tag{20}$$

- if $a \geq 1$, accept the new state; else accept the new state with probability a
- if new state is rejected, set $x^{(t+1)} = x^{(t)}$

metropolis-hastings

- at time t, generate tenative state x' from $q(x'; x^{(t)})$
- evaluate

$$a = \frac{p^*(x^{(t)})}{p^*(x')} \frac{q(x^{(t)}; x')}{q(x'; x^{(t)})} \tag{20}$$

- if $a \geq 1$, accept the new state; else accept the new state with probability a
- if new state is rejected, set $x^{(t+1)} = x^{(t)}$
- effective for high dimensional problems, but difficult to assess "convergence" of markov chain[3]

[3] see neal (1993)

gibbs sampling

- metropolis method where proposal density is chosen as conditional distribution, i.e.

$$q(x_i'; x^{(t)}) = p\left(x_i | \{x_j^{(t)}\}_{j\neq i}\right) \tag{21}$$

- useful when joint density factorizes, as in sparse graphical model[4]

[4] see wainwright & jordan 2008

gibbs sampling

- metropolis method where proposal density is chosen as conditional distribution, i.e.

$$q(x_i'; x^{(t)}) = p\left(x_i | \{x_j^{(t)}\}_{j \neq i}\right) \tag{21}$$

- useful when joint density factorizes, as in sparse graphical model[4]
- similar difficulties to metropolis, but no concerns about adjustable parameters

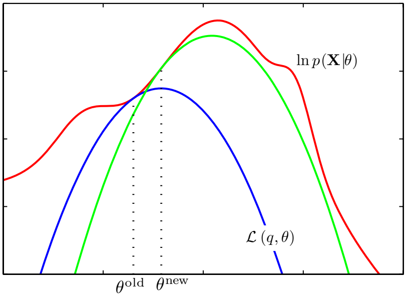
[4] see wainwright & jordan 2008

outline

1. principles (what we'd like to do)
   - background: joint, marginal, and conditional probabilities
   - bayes' theorem: inverting conditional probabilities
   - bayesian probability: unknowns as random variables
   - bayesian inference: bayesian probability + bayes' theorem
2. practice (what we're able to do)
   - monte carlo methods: representative samples
   - variational methods: bound optimization
   - references

bound optimization

- general approach: replace integration with optimization
- construct auxiliary function upper-bounded by log-evidence, maximize auxiliary function



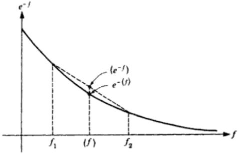
[5]

[5] image from bishop (2006)

variational bayes

- bound log of expected value by expected value of log using jensen's inequality[6]:

$$
\begin{aligned}
-\ln p(\mathcal{D}) &= -\ln \int d\Theta \; p(\mathcal{D}|\Theta)p(\Theta) \\
&= -\ln \int d\Theta \; \frac{p(\mathcal{D}|\Theta)p(\Theta)}{\color{red}{q(\Theta)}} \color{red}{q(\Theta)} \\
&\leq -\int d\Theta \; \ln \left[\frac{p(\mathcal{D}|\Theta)p(\Theta)}{q(\Theta)} \right] q(\Theta)
\end{aligned}
$$



- for sufficiently simple (i.e. factorized) approximating distribution $q(\Theta)$, right-hand side can be easily evaluated and optimized

[6] image from feynman (1972)

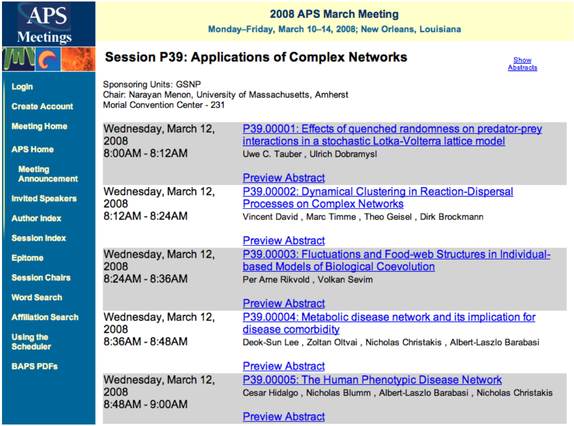
variational bayes

- iterative coordinate ascent algorithm provides controlled analytic approxmations to posterior and evidence
- approximate posterior $q(\Theta)$ minimizes kullback-leibler distance to true posterior
- resulting deterministic algorithm is often fast and scalable

variational bayes

- iterative coordinate ascent algorithm provides controlled analytic approxmations to posterior and evidence
- approximate posterior $q(\Theta)$ minimizes kullback-leibler distance to true posterior
- resulting deterministic algorithm is often fast and scalable
- complexity of approximation often limited (to, e.g., mean-field theory, assuming weak interaction between unknowns)
- iterative algorithm requires restarts, no guarantees on quality of approximation

example: a bayesian approach to network modularity



APS Meetings

2008 APS March Meeting

Monday–Friday, March 10–14, 2008; New Orleans, Louisiana

Login
Create Account
Meeting Home
APS Home
Meeting Announcement
Invited Speakers
Author Index
Session Index
Epitome
Session Chairs
Word Search
Affiliation Search
Using the Scheduler
BAPS PDFs

Session P39: Applications of Complex Networks

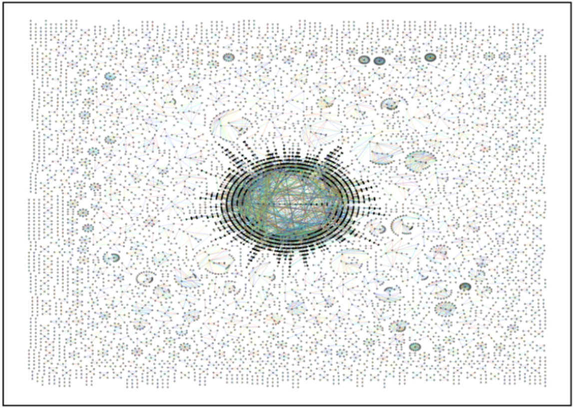
Show Abstracts

Sponsoring Units: GSNP
Chair: Narayan Menon, University of Massachusetts, Amherst
Morial Convention Center - 231

Wednesday, March 12, 2008 8:00AM - 8:12AM	P39.00001: Effects of quenched randomness on predator-prey interactions in a stochastic Lotka-Volterra lattice model Uwe C. Tauber , Ulrich Dobramysl Preview Abstract
Wednesday, March 12, 2008 8:12AM - 8:24AM	P39.00002: Dynamical Clustering in Reaction-Dispersal Processes on Complex Networks Vincent David , Marc Timme , Theo Geisel , Dirk Brockmann Preview Abstract
Wednesday, March 12, 2008 8:24AM - 8:36AM	P39.00003: Fluctuations and Food-web Structures in Individual-based Models of Biological Coevolution Per Arne Rikvold , Volkan Sevim Preview Abstract
Wednesday, March 12, 2008 8:36AM - 8:48AM	P39.00004: Metabolic disease network and its implication for disease comorbidity Deok-Sun Lee , Zoltan Oltvai , Nicholas Christakis , Albert-Laszlo Barabasi Preview Abstract
Wednesday, March 12, 2008 8:48AM - 9:00AM	P39.00005: The Human Phenotypic Disease Network Cesar Hidalgo , Nicholas Blumm , Albert-Laszlo Barabasi , Nicholas Christakis Preview Abstract

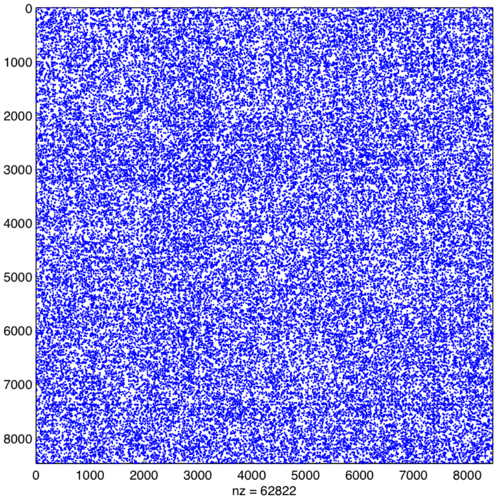
example: a bayesian approach to network modularity

nodes: authors, edges: co-authored papers

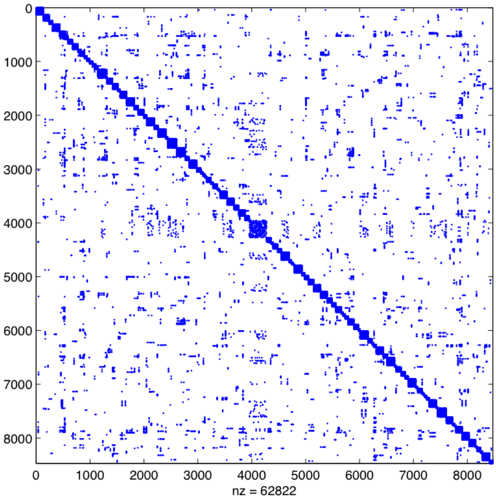


can we infer (community) structure in the giant component?

example: a bayesian approach to network modularity

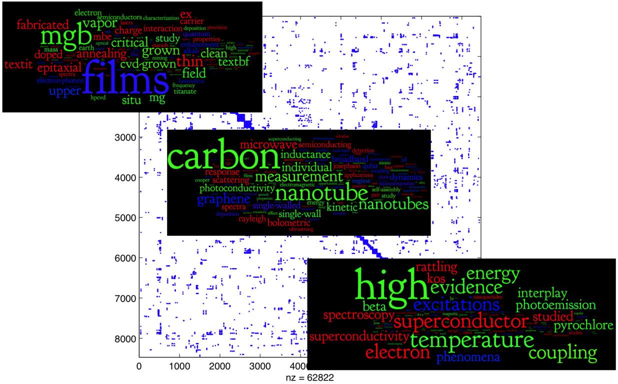


example: a bayesian approach to network modularity



example: a bayesian approach to network modularity

inferred topological communities correspond to sub-disciplines



outline

1. principles (what we'd like to do)
   - background: joint, marginal, and conditional probabilities
   - bayes' theorem: inverting conditional probabilities
   - bayesian probability: unknowns as random variables
   - bayesian inference: bayesian probability + bayes' theorem
2. practice (what we're able to do)
   - monte carlo methods: representative samples
   - variational methods: bound optimization
   - references

- “information theory, inference, and learning algorithms”, mackay (2003)
- “pattern recognition and machine learning”, bishop (2006)
- “bayesian data analysis”, gelman, et. al. (2003)
- “probabilistic inference using markov chain monte carlo methods”, neal (1993)
- “graphical models, exponential families, and variational inference”, wainwright & jordan (2006)
- “probability theory: the logic of science”, jaynes (2003)
- “what is bayes’ theorem ...”, wiggins (2006)
- bayesian inference view on cran
- variational-bayes.org
- variational bayesian inference for network modularity