Message Passing Algorithms: A Success Looking for Theoreticians

Andrea Montanari

Stanford University

June 5, 2010

What is this talk about? A couple of examples

$A \in \mathbb{F}_2^{m \times n}$, $y \in \mathbb{F}_2^n$

$$\begin{cases} \text{minimize} & d(x, y)\,, \\ \text{subject to} & Ax = 0\,. \end{cases}$$

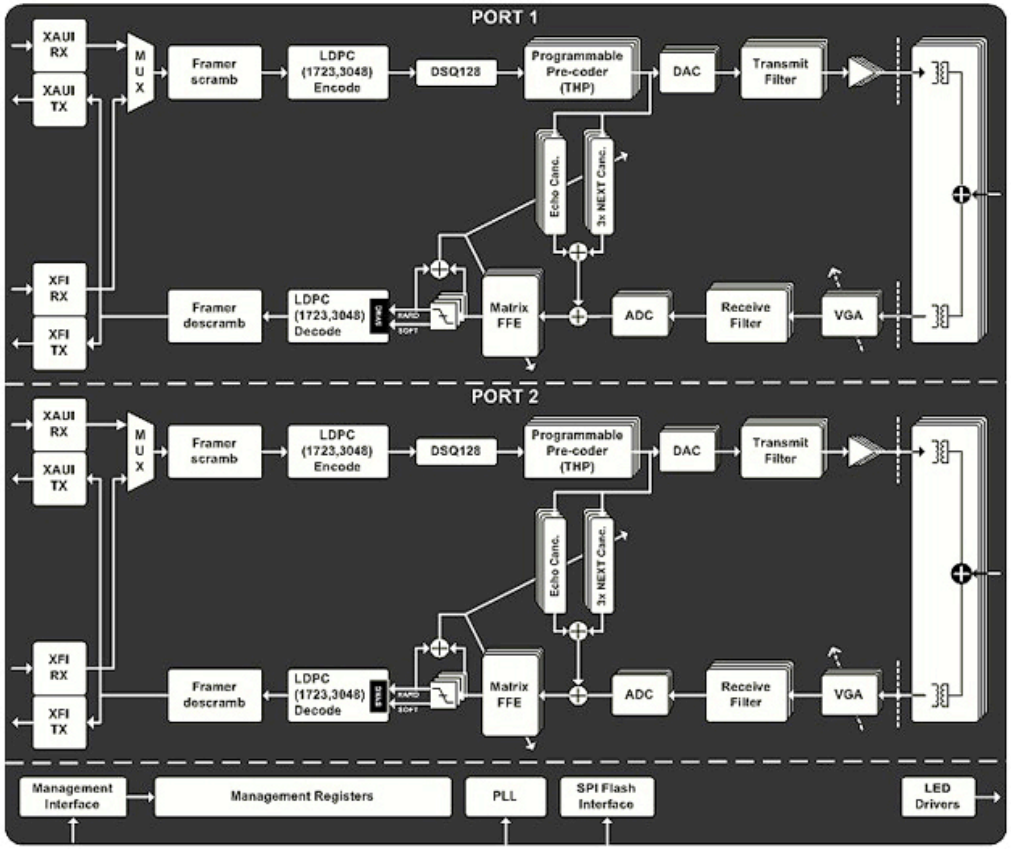
What is this talk about? A couple of examples

$A \in \mathbb{F}_2^{m \times n}$, $y \in \mathbb{F}_2^n$

$$\begin{cases} \text{minimize} & d(x, y)\,, \\ \text{subject to} & Ax = 0\,. \end{cases}$$

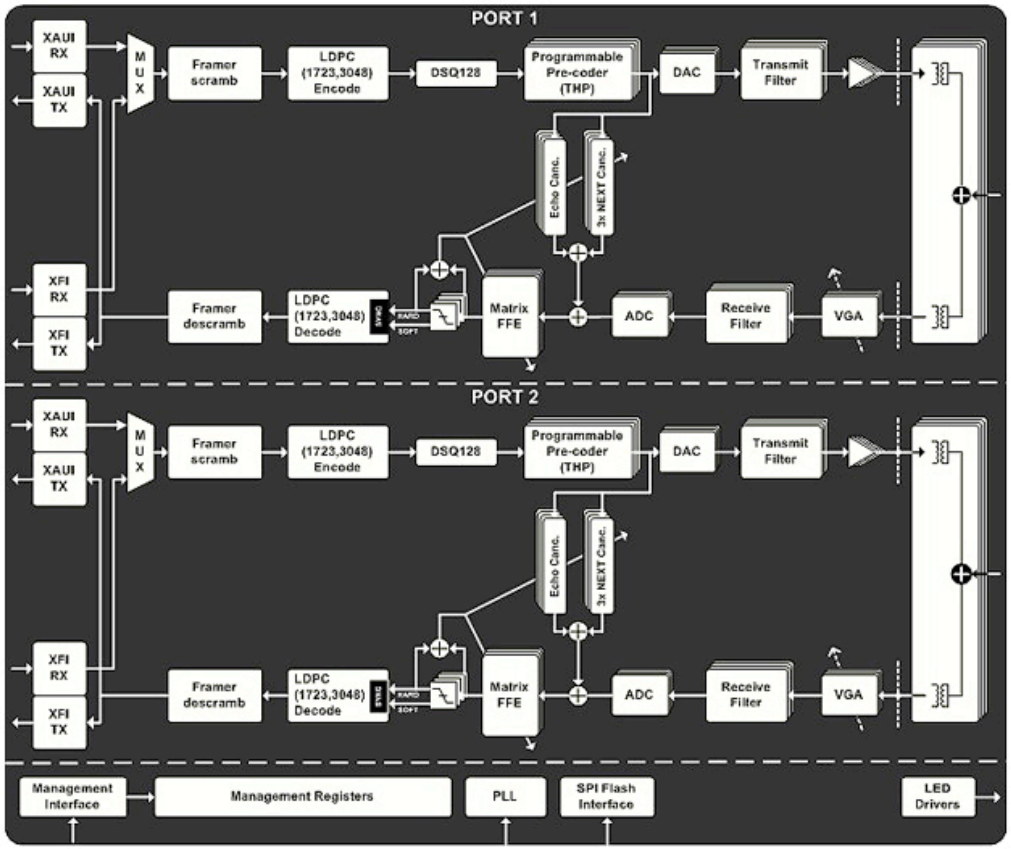
You should not hope for easy solutions...

You should not hope for easy solutions...



One of the little boxes solves it for $n = 2048$. In 10^{-6} secs.

You should not hope for easy solutions...



Uses a message passing algorithm + $A \approx$ random sparse

Do not worry!

No more hardware diagrams in this talk!!!

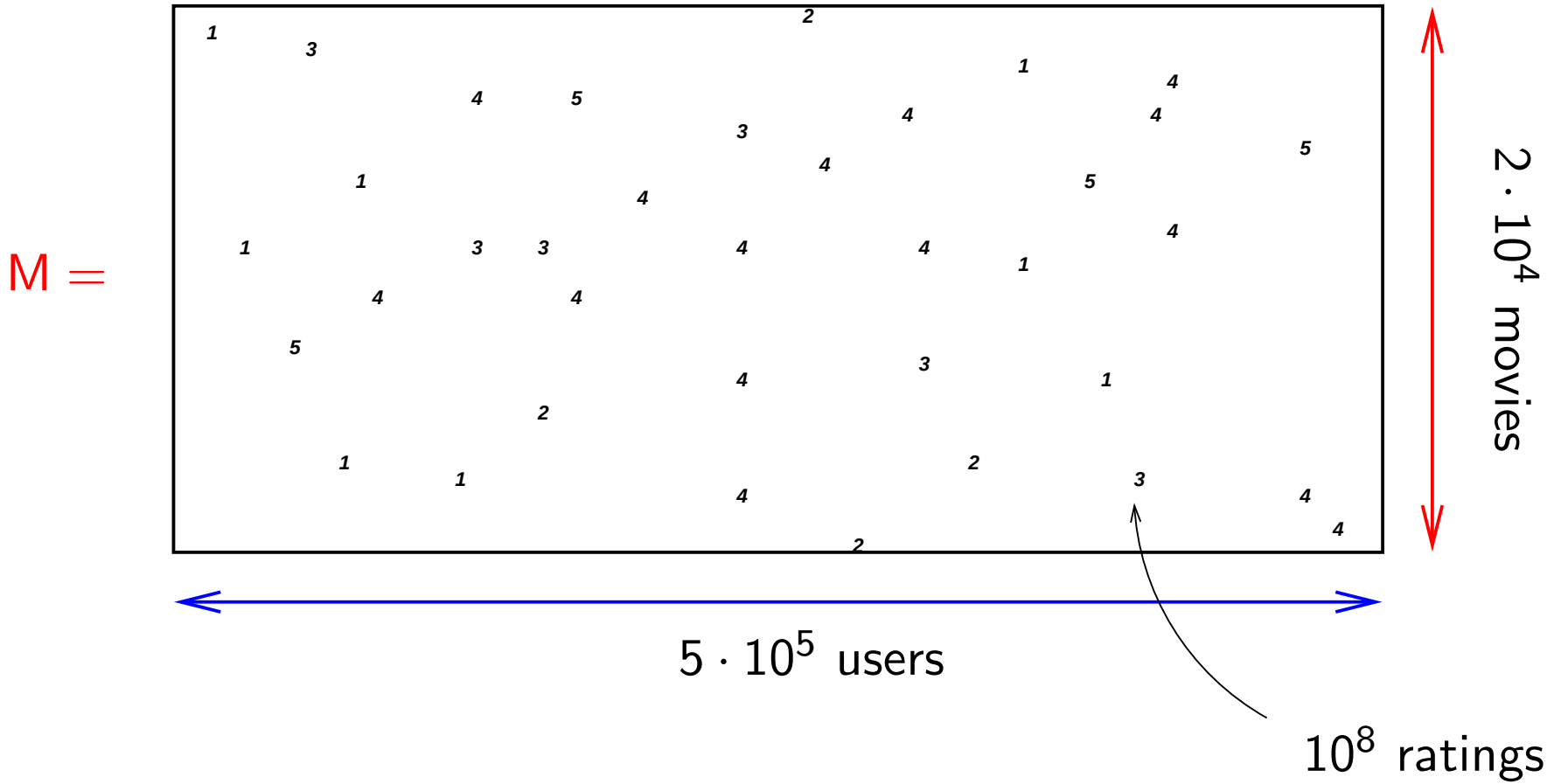
Do not worry!

No more hardware diagrams in this talk!!!

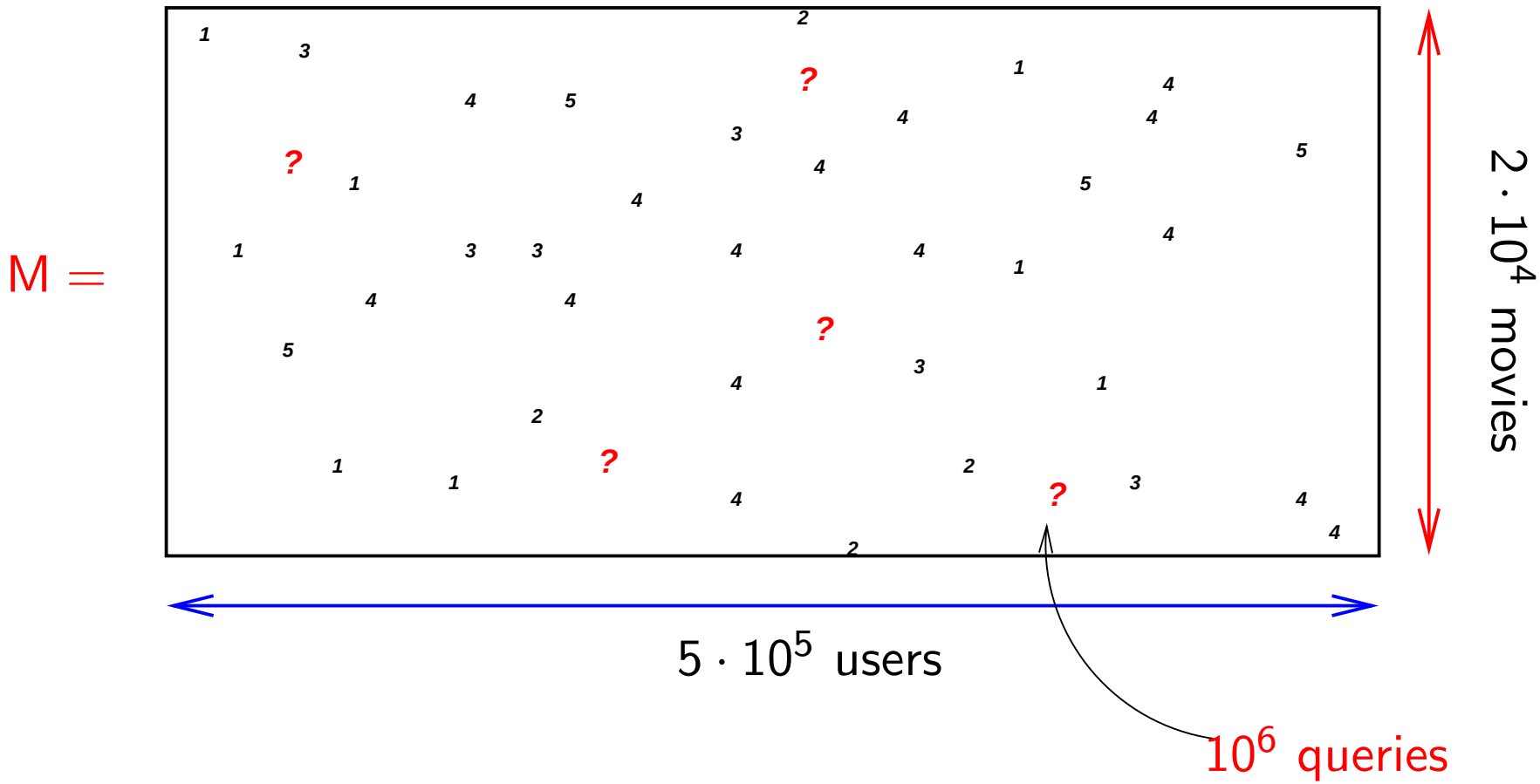
Second example: Learning low-rank matrices

Second example: Learning low-rank matrices

The Netflix dataset



Learning low-rank matrices: The Netflix dataset



A prize awarded for:

$$\text{RMSE} < 0.8563 \qquad ;-)$$

A popular cost function

$$\mathcal{C}(X,Y) \equiv \frac{1}{2} \sum_{(i,j)\in \text{Revealed}} \left| M_{ij} - (XY^T)_{ij} \right|^2 + \frac{\lambda}{2} \|X\|_F^2 + \frac{\lambda}{2} \|Y\|_F^2 \, .$$

$$X \in \mathbb{R}^{n\times r}, \; Y \in \mathbb{R}^{m\times r}$$

[Srebro, Rennie, Jaakkola 2005]

- Non convex.
- Large (!) scale.

A popular cost function

$$\mathcal{C}(X,Y) \equiv \frac{1}{2} \sum_{(i,j)\in \text{Revealed}} \left| M_{ij} - (XY^T)_{ij} \right|^2 + \frac{\lambda}{2} \|X\|_F^2 + \frac{\lambda}{2} \|Y\|_F^2 \, .$$

$$X \in \mathbb{R}^{n\times r}, \; Y \in \mathbb{R}^{m\times r}$$

[Srebro, Rennie, Jaakkola 2005]

- Non convex.
- Large (!) scale.

A popular cost function

$$\mathcal{C}(X,Y) \equiv \frac{1}{2} \sum_{(i,j)\in \text{Revealed}} \left| M_{ij} - (XY^T)_{ij} \right|^2 + \frac{\lambda}{2} \|X\|_F^2 + \frac{\lambda}{2} \|Y\|_F^2 \, .$$

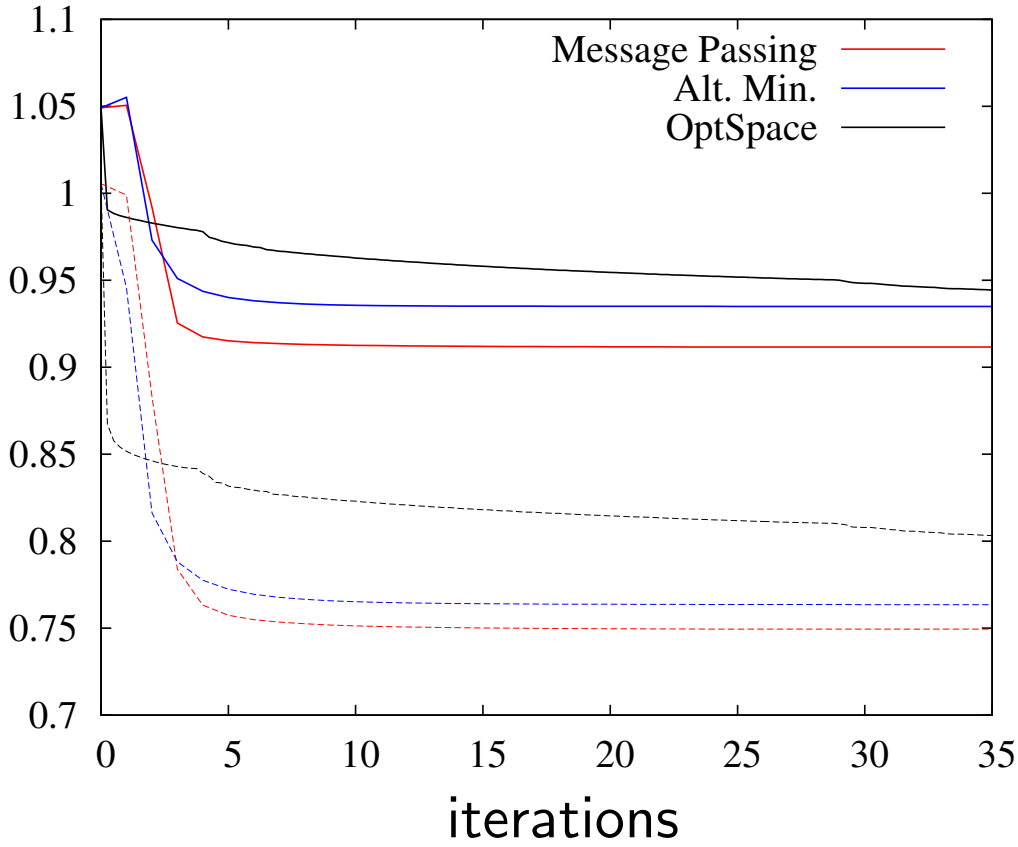
$$X \in \mathbb{R}^{n\times r}, \; Y \in \mathbb{R}^{m\times r}$$

[Srebro, Rennie, Jaakkola 2005]

- Non convex.
- Large (!) scale.

Three algorithms

RMSE



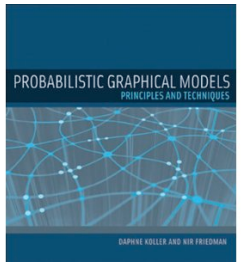
OptSpace $\sim$ Gradient descent [Keshavan-Montanari-Oh 2009]

Alternating Least Squares [Koren-Bell 2008]

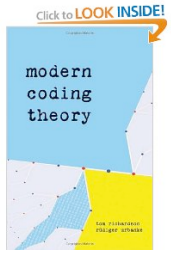
Message Passing [Keshavan-Montanari 2010]

Examples everywhere!

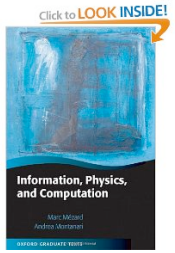
- Machine learning and AI.



- Coding and communications.



- Large random structures, statistical mechanics,. . .



- . . .

Outline

k-SAT: A (Very) Simple Algorithm

k-SAT

$x = (x_1, \dots, x_n) \in \{0,1\}^n$

Instance ($k = 3$):

$$(x_1 \vee x_4 \vee \overline{x}_6) \wedge (x_7 \vee x_8 \vee \overline{x}_{10}) \wedge \cdots \wedge (\overline{x}_{12} \vee \overline{x}_{17} \vee x_{19})$$

Broder-Frieze-Upfal (1993)

PURE LITERAL	
1:	**Repeat :**
2:	Find x_i pure literal;
3:	Fix x_i;

x_i is a pure literal if it never appears negated
or if only appears negated

Broder-Frieze-Upfal (1993)

Random k-SAT:
Uniformly random formula with n variables and $n\alpha$ clauses.

Analysis:

I. Markov chain in reduced state space.

II. ODE method.

Broder-Frieze-Upfal (1993)

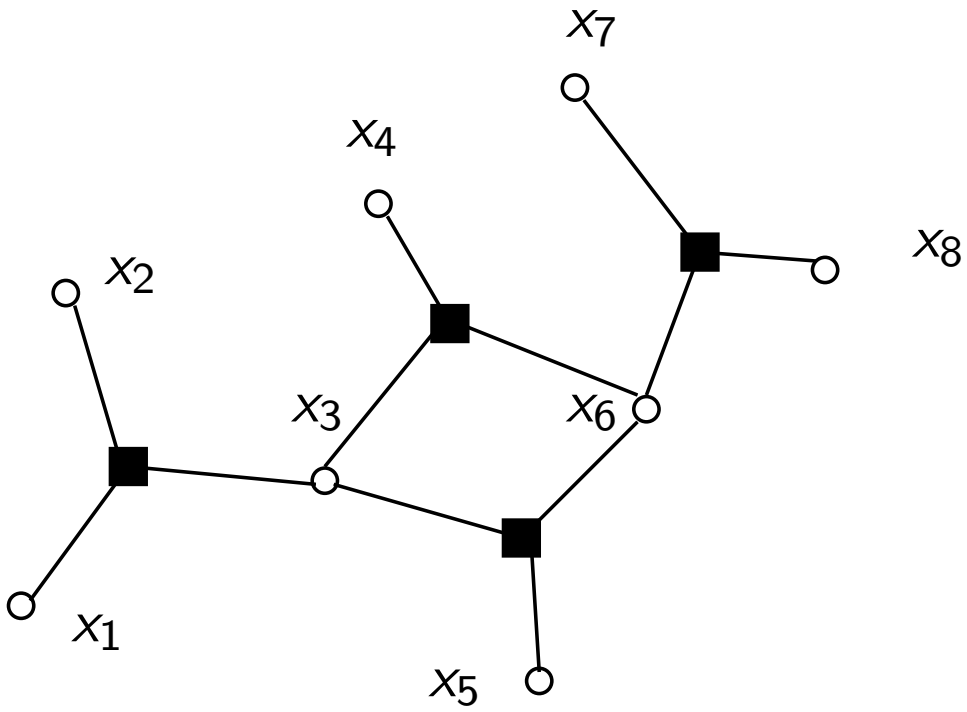
Random k-SAT:
Uniformly random formula with n variables and $n\alpha$ clauses.

Analysis:

I. Markov chain in reduced state space.

II. ODE method.

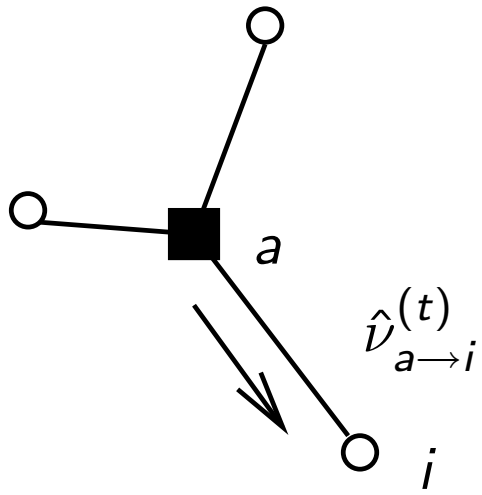
Message passing: 1. Factor graph

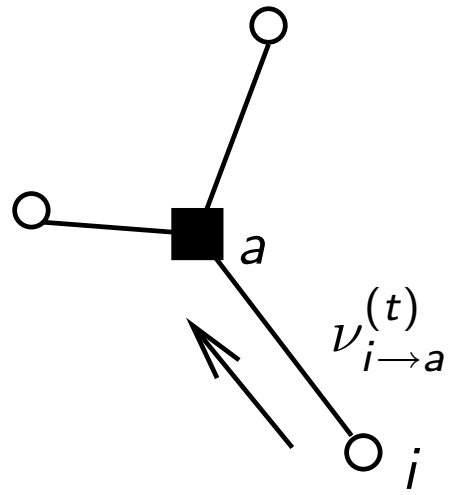


$$(x_1 \vee x_2 \vee \overline{x}_3) \wedge (x_3 \vee \overline{x}_4 \vee \overline{x}_6) \wedge (\overline{x}_3 \vee x_5 \vee \overline{x}_6) \wedge (x_6 \vee x_7 \vee \overline{x}_8)$$

[Labeled bipartite graph]

Message passing: 2. Messages

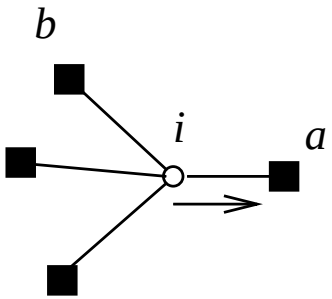




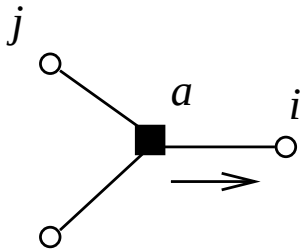
$$\nu^{(t)}_{i\to a}, \hat{\nu}^{(t)}_{a\to i} \in \{\text{free}, \text{cons}\}\,.$$

Message passing: 3. Update rules

$s_{jb} \to$ label on edge (j, b)



$$\nu_{i\to a}^{(t+1)} = \begin{cases} \text{cons} & \text{if } s_{ib} \neq s_{ia},\ \hat{\nu}_{b\to i}^{(t)} = \text{cons for some } b \in \partial i \setminus a, \\ \text{free} & \text{otherwise.} \end{cases}$$

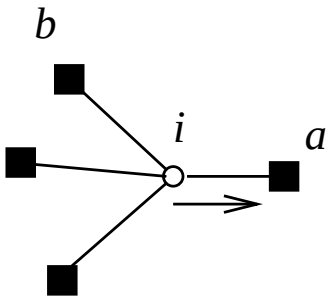


$$\hat{\nu}_{a\to i}^{(t)} = \begin{cases} \text{free} & \text{if } \nu_{j\to a}^{(t)} = \text{free for some } j \in \partial a \setminus i, \\ \text{cons} & \text{otherwise.} \end{cases}$$

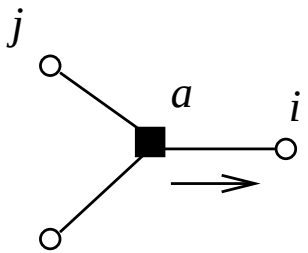
Equivalent to Pure Literal!

Message passing: 3. Update rules

$s_{jb} \to$ label on edge (j, b)



$$\nu_{i\to a}^{(t+1)} = \begin{cases} \text{cons} & \text{if } s_{ib} \neq s_{ia},\ \hat{\nu}_{b\to i}^{(t)} = \text{cons for some } b \in \partial i \setminus a, \\ \text{free} & \text{otherwise.} \end{cases}$$

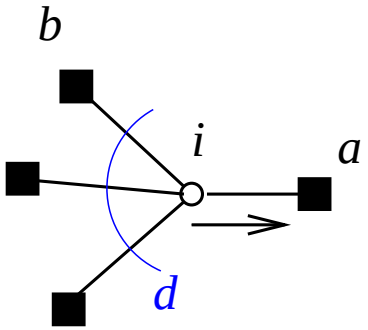


$$\hat{\nu}_{a\to i}^{(t)} = \begin{cases} \text{free} & \text{if } \nu_{j\to a}^{(t)} = \text{free for some } j \in \partial a \setminus i, \\ \text{cons} & \text{otherwise.} \end{cases}$$

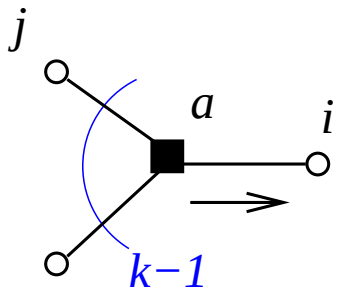
Equivalent to Pure Literal!

Message passing: 4. Analysis (density evolution)

$\phi_t = \mathbb{P}\{\nu^{(t)}_{j\to a} = \text{cons}\}, \qquad \widehat{\phi}_t = \mathbb{P}\{\hat{\nu}^{(t)}_{a\to i} = \text{cons}\},$



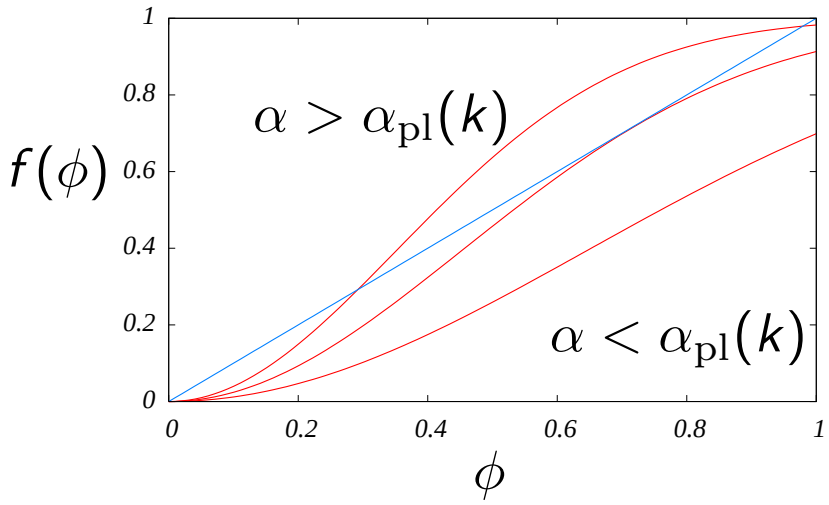
$$\phi_{t+1} = 1 - \mathbb{E}\Big\{\Big[1 - \frac{1}{2}\widehat{\phi}_t\Big]^d\Big\} = 1 - e^{-\frac{k\alpha}{2}\widehat{\phi}_t}\,,$$



$$\widehat{\phi}_t = \phi_t^{k-1}\,,$$

Message passing: 4. Analysis

$$\phi_{t+1} = 1 - \exp\{-k\alpha\phi_t^{k-1}/2\} \equiv f(\phi_t)$$



$$\alpha_{\rm pl}(k) = \sup\left\{\, \alpha \,:\, 1 - e^{-k\alpha x^{k-1}/2} \le x \;\; \forall x \in [0,1]\right\}$$

Theorem (Broder-Frieze-Upfal 93, Molloy 04)

PURE LITERAL *finds a solution whp if* $\alpha < \alpha_{\rm pl}$ *and fails whp if* $\alpha > \alpha_{\rm pl}$.

This is a proof because

$\mathsf{B}_I(t) \equiv$ 'ball' of radius t around uniform $I \in [n]$

$\mathsf{T} \equiv$ some random rooted tree

$\mathsf{T}(t) \equiv$ its first t generations

Definition (Benjamini-Schramm 1996, Aldous-Steele 2003)

The sequences of (factor) graphs $G_n = (V_n = [n], E_n)$ converges locally to T if, for any t, $\mathsf{B}_I(t)$ converges in distribution to $\mathsf{T}(t)$.

This is a proof because

$\mathsf{B}_I(t) \equiv$ 'ball' of radius t around uniform $I \in [n]$

$\mathsf{T} \equiv$ some random rooted tree

$\mathsf{T}(t) \equiv$ its first t generations

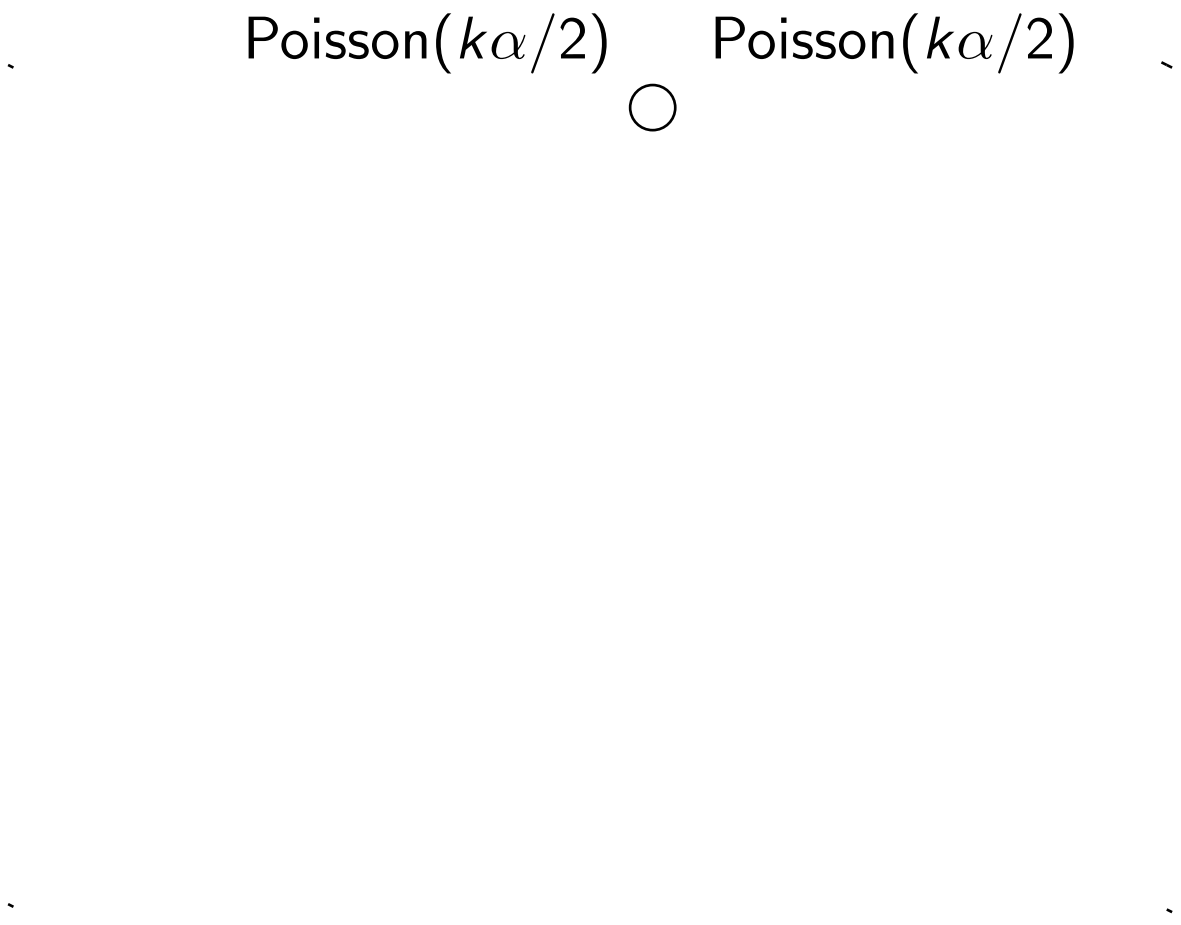
Definition (Benjamini-Schramm 1996, Aldous-Steele 2003)

The sequences of (factor) graphs $G_n = (V_n = [n], E_n)$ converges locally to T if, for any t, $\mathsf{B}_I(t)$ converges in distribution to $\mathsf{T}(t)$.

Example

Lemma

Random k-SAT instances $\overset{\text{locally}}{\longrightarrow}$ *Poisson Galton-Watson trees.*

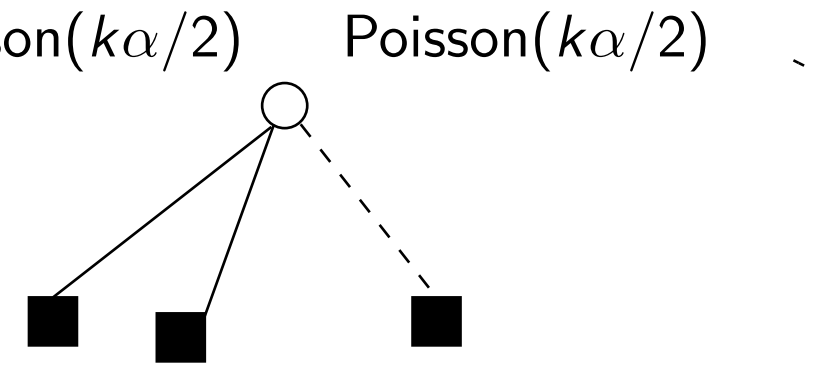


[In fact a little bit more is needed...]

Example

Lemma

Random k-SAT instances $\overset{\text{locally}}{\longrightarrow}$ *Poisson Galton-Watson trees.*

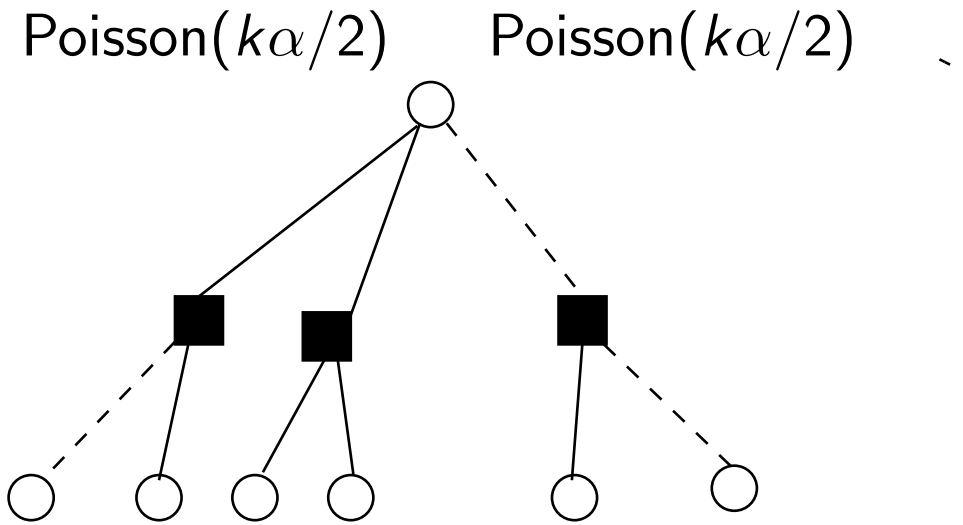


[In fact a little bit more is needed. . .]

Example

Lemma

Random k-SAT instances $\overset{\text{locally}}{\longrightarrow}$ *Poisson Galton-Watson trees.*

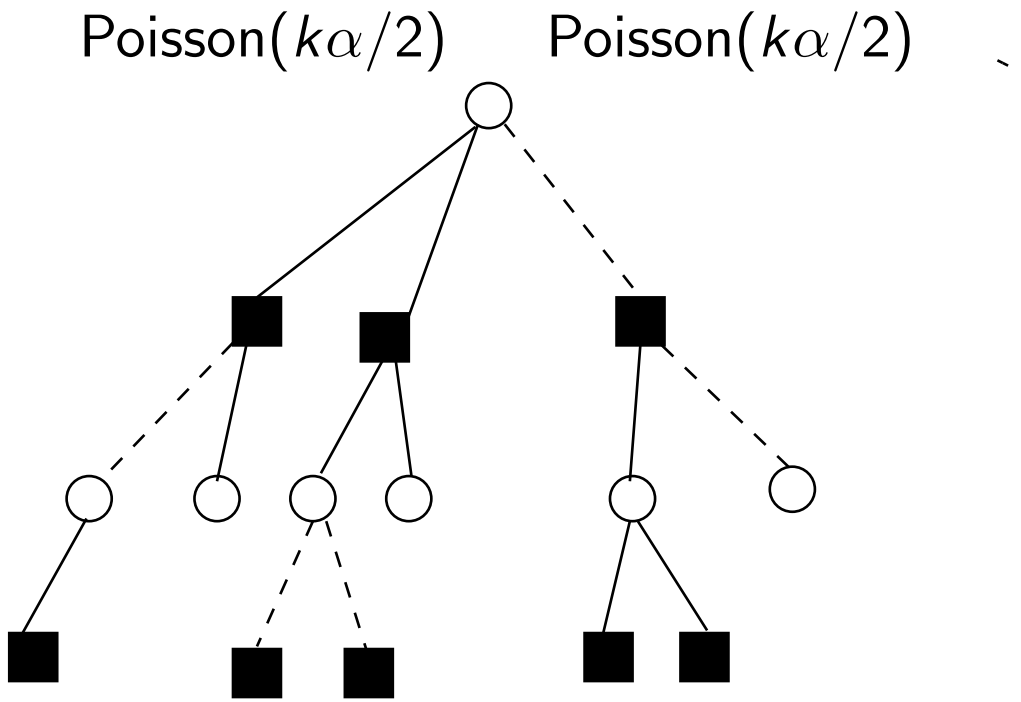


[In fact a little bit more is needed. . .]

Example

Lemma

Random k-SAT instances $\overset{\text{locally}}{\longrightarrow}$ *Poisson Galton-Watson trees.*

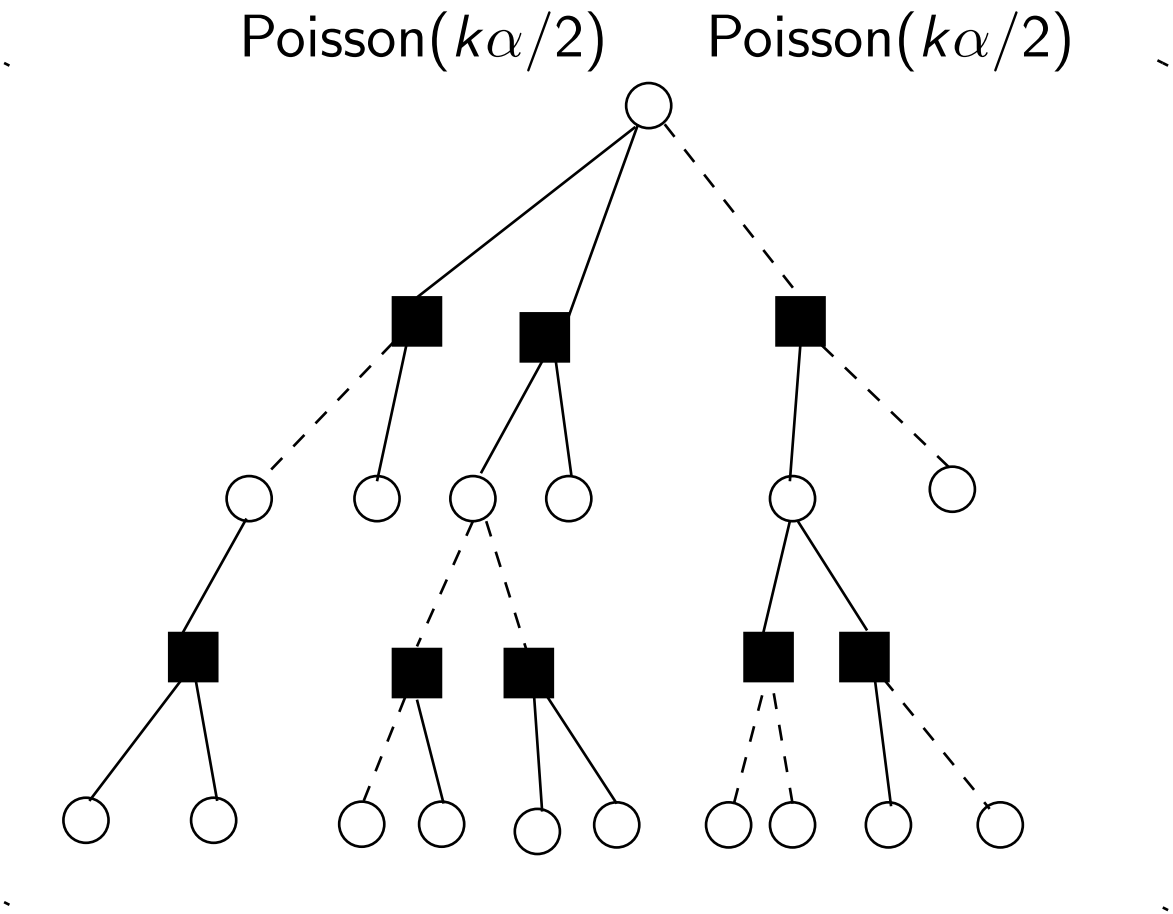


[In fact a little bit more is needed. . .]

Example

Lemma

Random k-SAT instances $\overset{\text{locally}}{\longrightarrow}$ *Poisson Galton-Watson trees.*

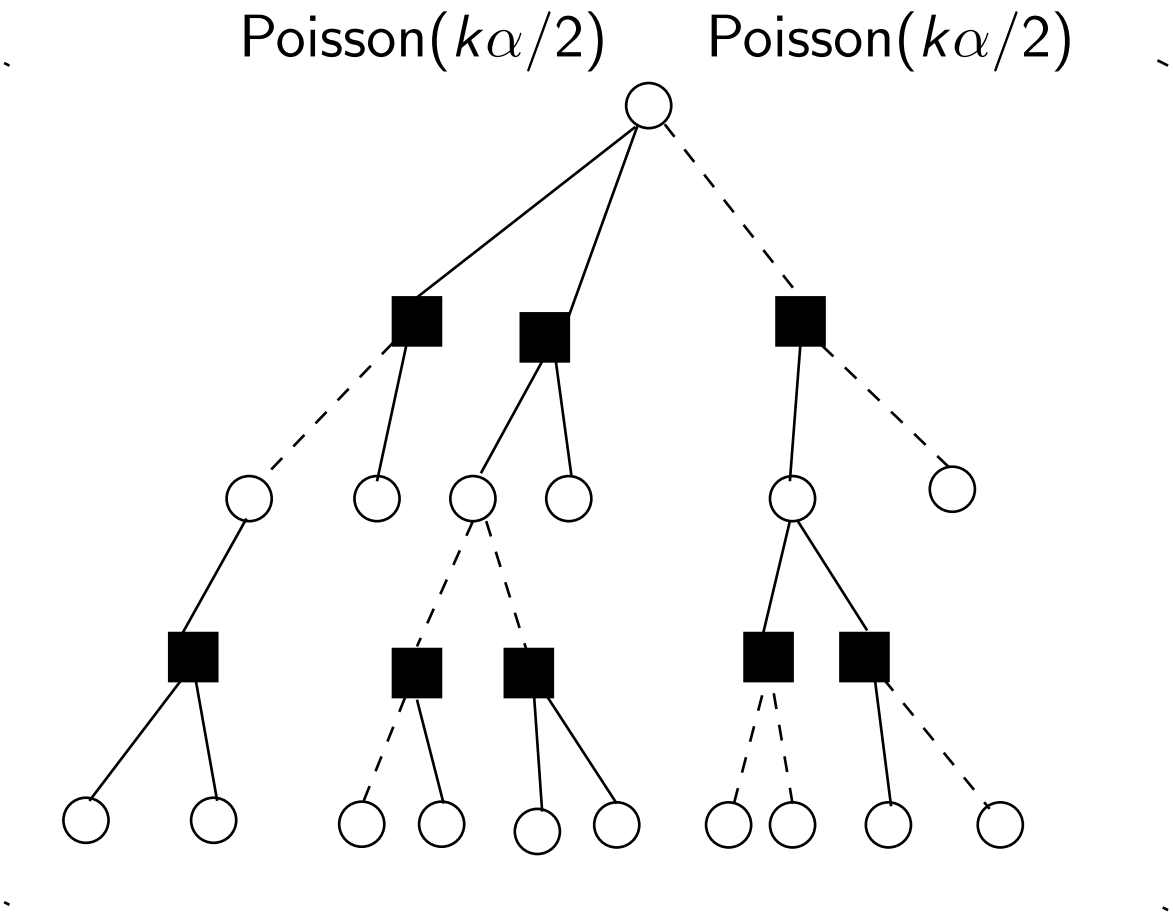


[In fact a little bit more is needed. . .]

Example

Lemma

Random k-SAT instances $\overset{\text{locally}}{\longrightarrow}$ *Poisson Galton-Watson trees.*

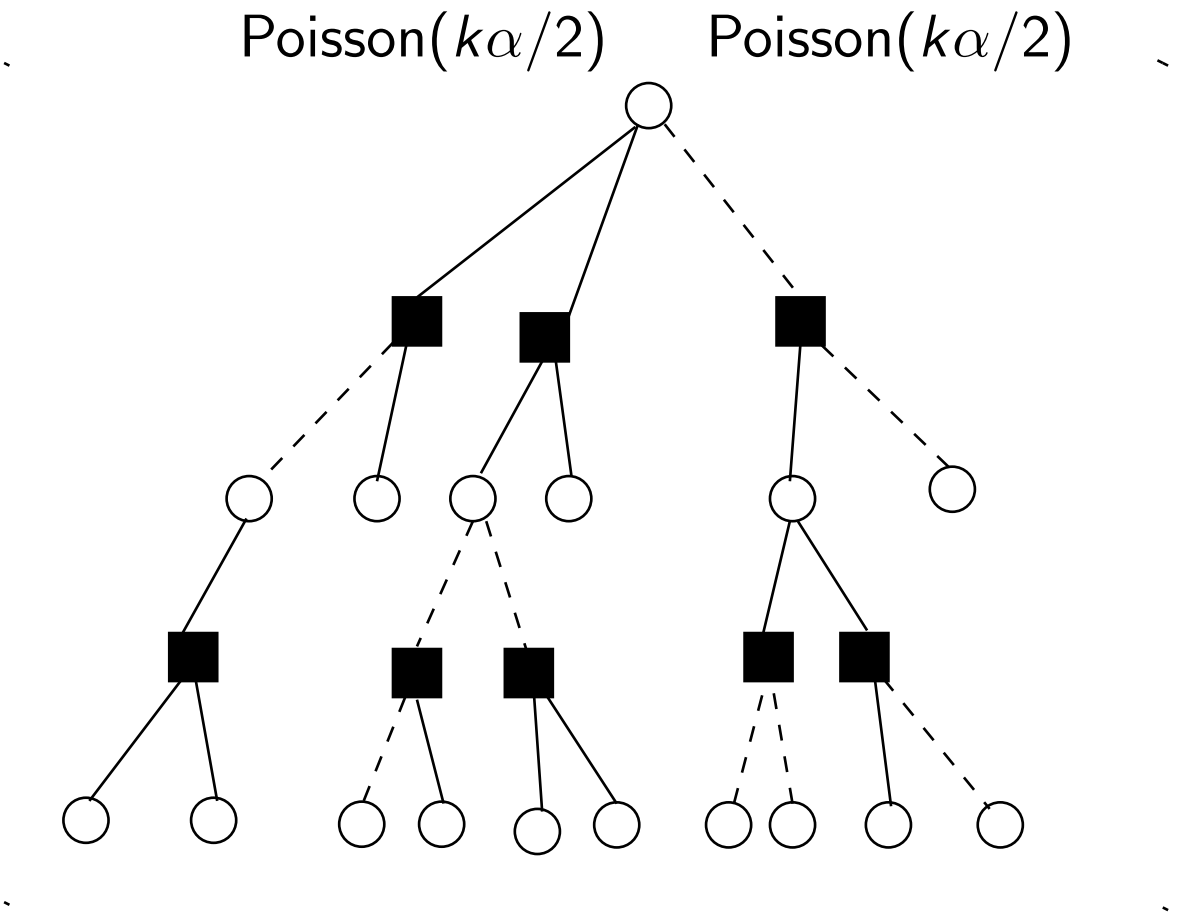


[In fact a little bit more is needed...]

Example

Lemma

Random k-SAT instances $\overset{\text{locally}}{\longrightarrow}$ *Poisson Galton-Watson trees.*



Analysis generalizes to other ensembles

A parenthesis: Generalizations are useful

LDPC codes [Gallager 1966, Luby et al. 2001]

$$\text{Code} = \left\{ x \in \{0,1\}^n \ : \ Ax = 0 \mod 2 \right\}$$

$A =$ adjacency matrix of sparse (pseudo)random graph

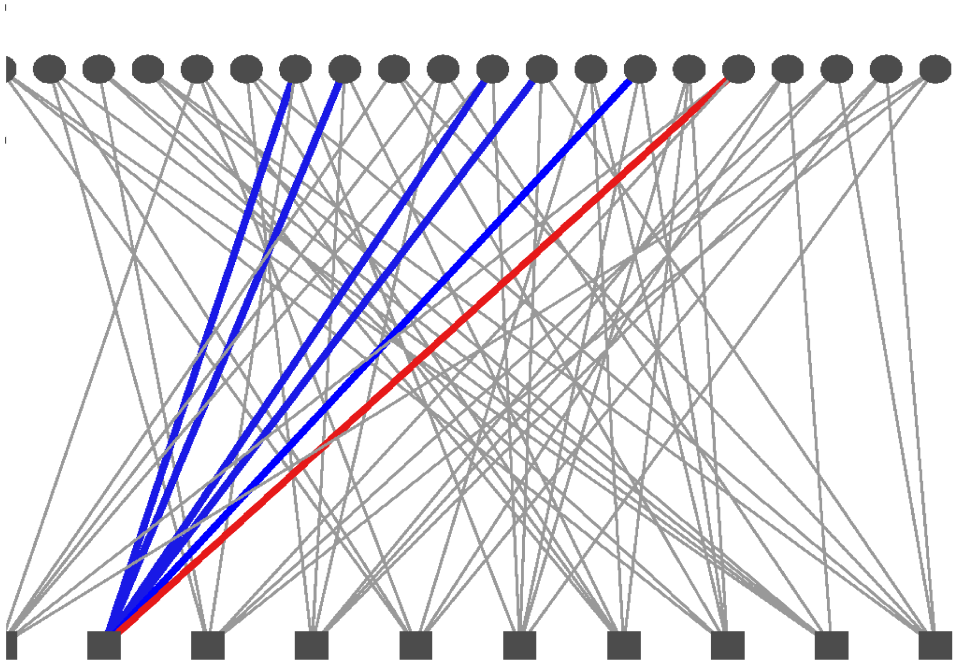
A parenthesis: Generalizations are useful

LDPC codes [Gallager 1966, Luby et al. 2001]

$$\text{Code} = \left\{ x \in \{0,1\}^n \ : \ Ax = 0 \mod 2 \right\}$$

$A =$ adjacency matrix of sparse (pseudo)random graph

A parenthesis: Generalizations are useful

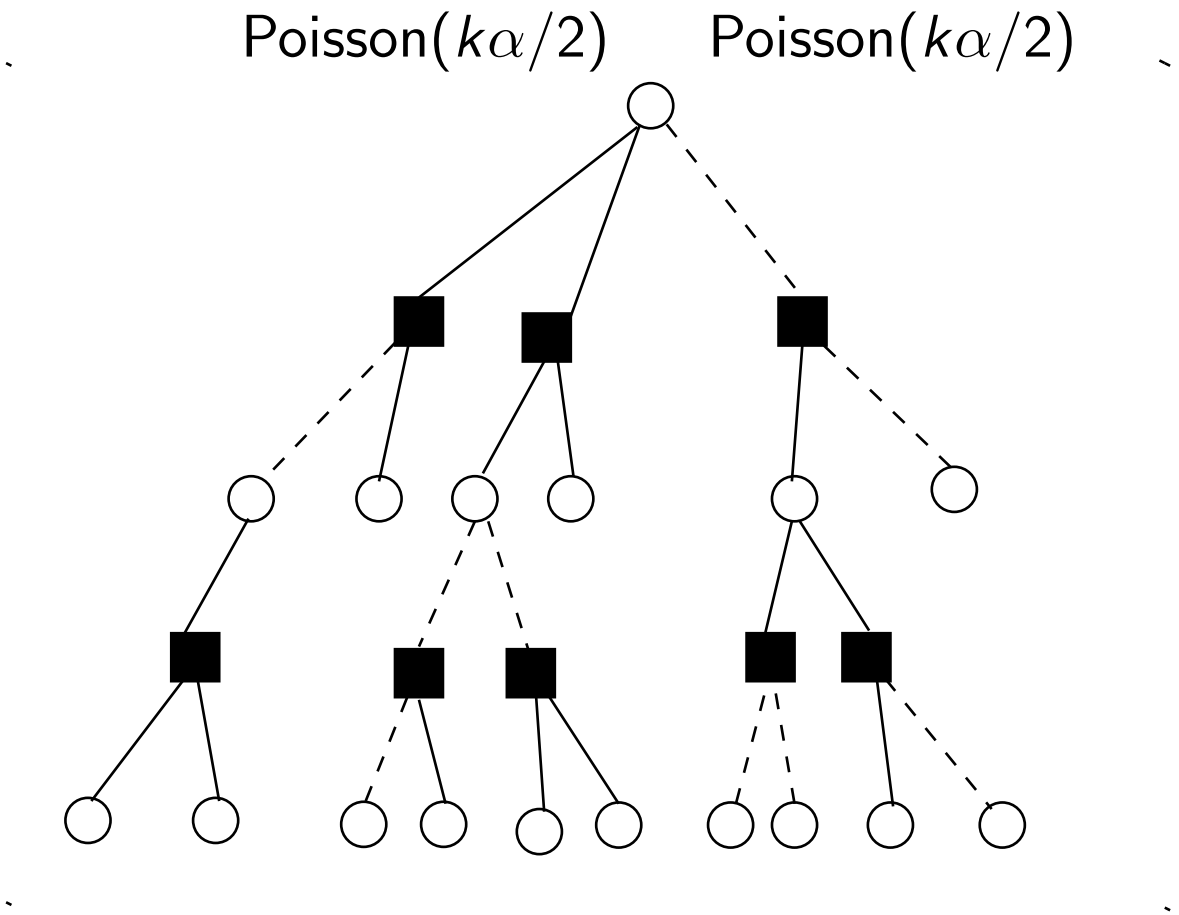


Luby et al. 2001:

$\rightarrow$ Random graph w degree distributions (λ, ρ)

$\rightarrow$ Optimization over (λ, ρ)

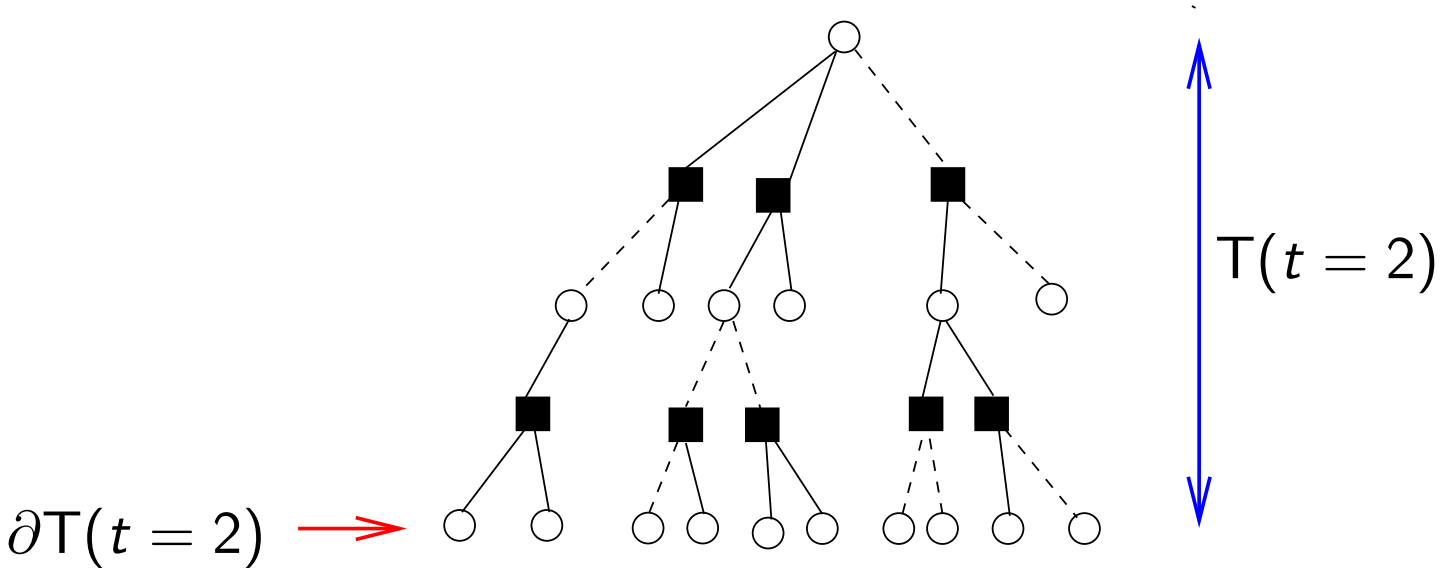
Relation with Gibbs measures



$$\mu^{\mathsf{T}}(x) = \text{ uniform measure over solutions of } \mathsf{T}(\infty)$$

Relation with Gibbs measures

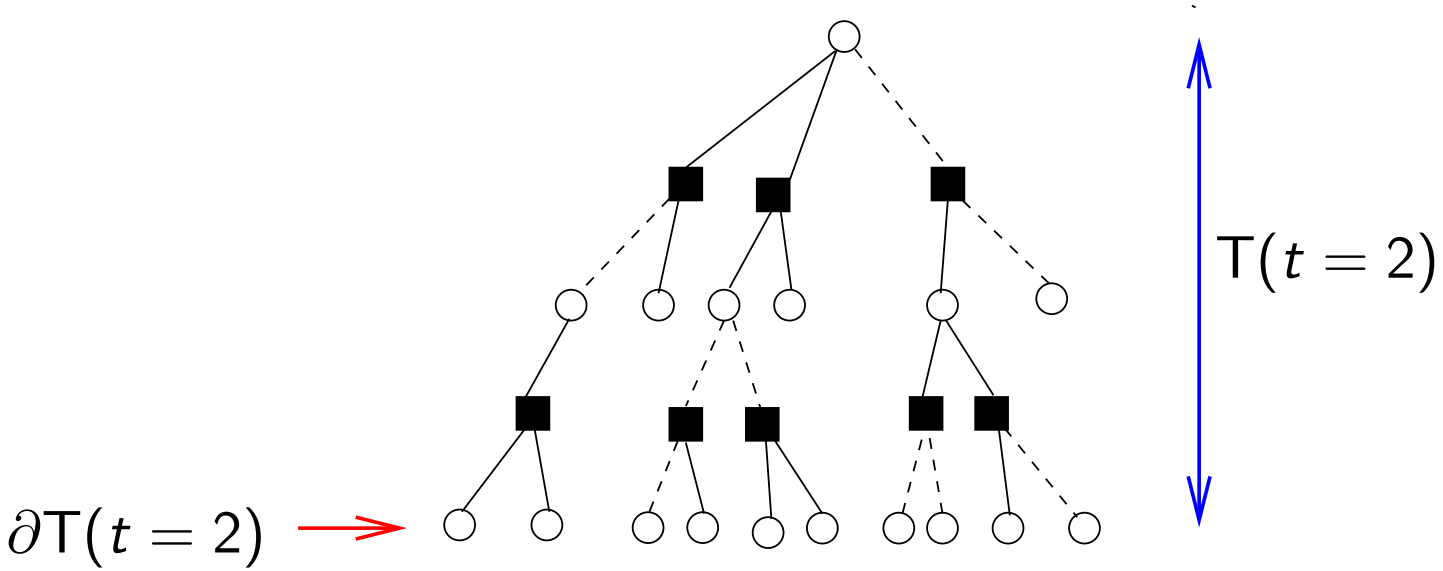
What does it mean uniform ???



Definition (DLR)

μ^{T} is 'uniform' (Gibbs) if $\mu^{\mathsf{T}}(x_{\mathsf{T}(t)}|x_{\partial \mathsf{T}(t)})$ is uniform for all t.

Relation with Gibbs measures



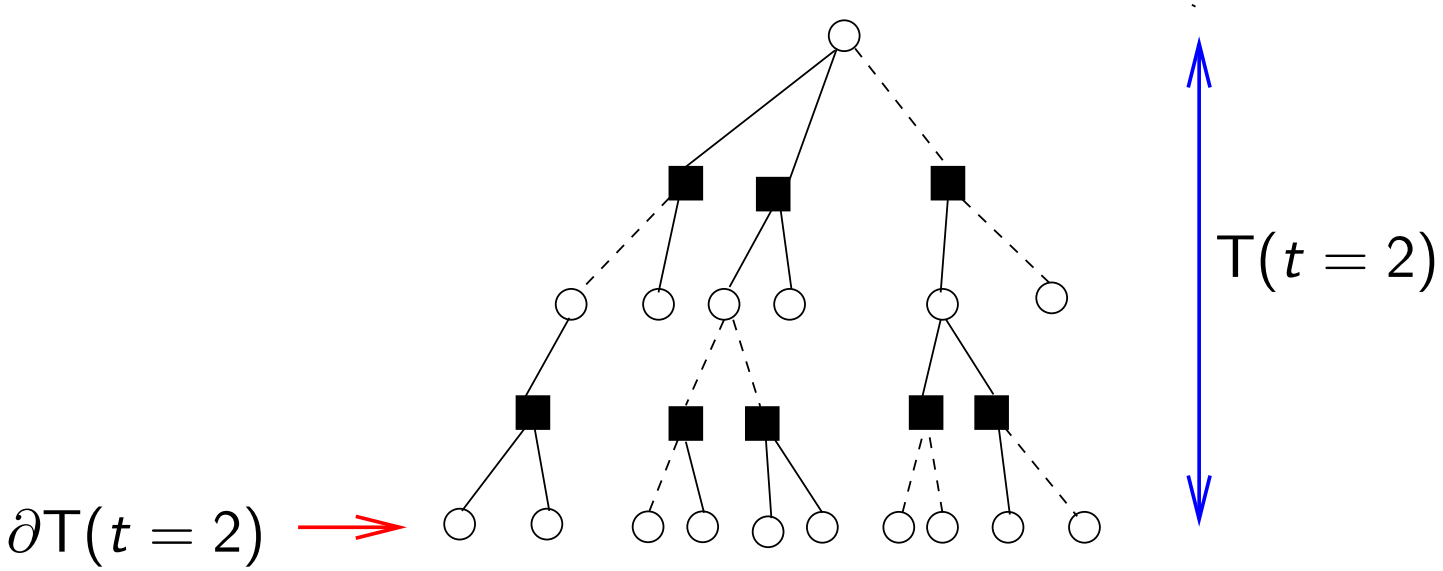
Lemma (Montanari-Shah 2010)

μ^{T} is unique if and only if $\alpha < \alpha_{\rm u}(k)$ with

$$\alpha_{\rm u}(k) = \frac{2\log k}{k}\left\{1 + o_k(1)\right\}.$$

In particular $\alpha_{\rm pl}(k) = \alpha_{\rm u}(k) + O(k^{-2})$.

More concretely



$$\sup_{x,x'} \left| \mu^{\mathsf{T}}(x_\emptyset = 1 | x_{\partial \mathsf{T}(t)}) - \mu^{\mathsf{T}}(x_\emptyset = 1 | x_{\partial \mathsf{T}(t)}) \right| \le A\, e^{-b\, t}\,.$$

Relation with MCMC

To avoid pathologies:

$$\mu^{n,\beta}(x) \equiv \frac{1}{Z_{n,\beta}} \exp\Big\{ -\beta \big|\{\text{clauses violated by } x\}\big| \Big\}$$

Conjecture

Heath bath/Glauber/Gibbs sampler has $\tau_{\rm mix} = O(n^C)$ *whp for* $\alpha < \alpha_{\rm u}(k)$.

[Mossel-Sly 201?]

Relation with MCMC

To avoid pathologies:

$$\mu^{n,\beta}(x) \equiv \frac{1}{Z_{n,\beta}} \exp\Big\{ -\beta \big|\{\text{clauses violated by } x\}\big| \Big\}$$

Conjecture

Heath bath/Glauber/Gibbs sampler has $\tau_{\rm mix} = O(n^C)$ *whp for* $\alpha < \alpha_{\rm u}(k)$.

[Mossel-Sly 201?]

Relation with MCMC

To avoid pathologies:

$$\mu^{n,\beta}(x) \equiv \frac{1}{Z_{n,\beta}} \exp\Big\{ -\beta \big|\{\text{clauses violated by } x\}\big| \Big\}$$

Conjecture

Heath bath/Glauber/Gibbs sampler has $\tau_{\rm mix} = O(n^C)$ whp for $\alpha < \alpha_{\rm u}(k)$.

[Mossel-Sly 201?]

Relation with MCMC

To avoid pathologies:

$$\mu^{n,\beta}(x) \equiv \frac{1}{Z_{n,\beta}} \exp\Big\{ -\beta \big|\{\text{clauses violated by } x\}\big| \Big\}$$

Conjecture

Heath bath/Glauber/Gibbs sampler has $\tau_{\rm mix} = O(n^C)$ whp for $\alpha < \alpha_{\rm u}(k)$.

[Mossel-Sly 201?]

Relation with MCMC

To avoid pathologies:

$$\mu^{n,\beta}(x) \equiv \frac{1}{Z_{n,\beta}} \exp\Big\{ -\beta \big|\{\text{clauses violated by } x\}\big| \Big\}$$

Conjecture

Heath bath/Glauber/Gibbs sampler has $\tau_{\rm mix} = O(n^C)$ *whp for* $\alpha < \alpha_{\rm u}(k)$.

[Mossel-Sly 201?]

Taking it seriously

How would you do it in your dreams?

I would use

MARGINAL ($i \in [n]$)
1: $\cdots$;
2: $\cdots$;
3: $\cdots$;
4: Return $\mu(x_i = 1)$.

$\mu(x_i = 1) =$ fraction of solutions in which $x_i = 1$

How would you do it in your dreams?

I would use

MARGINAL ($i \in [n]$)
1: $\cdots$;
2: $\cdots$;
3: $\cdots$;
4: Return $\mu(x_i = 1)$.

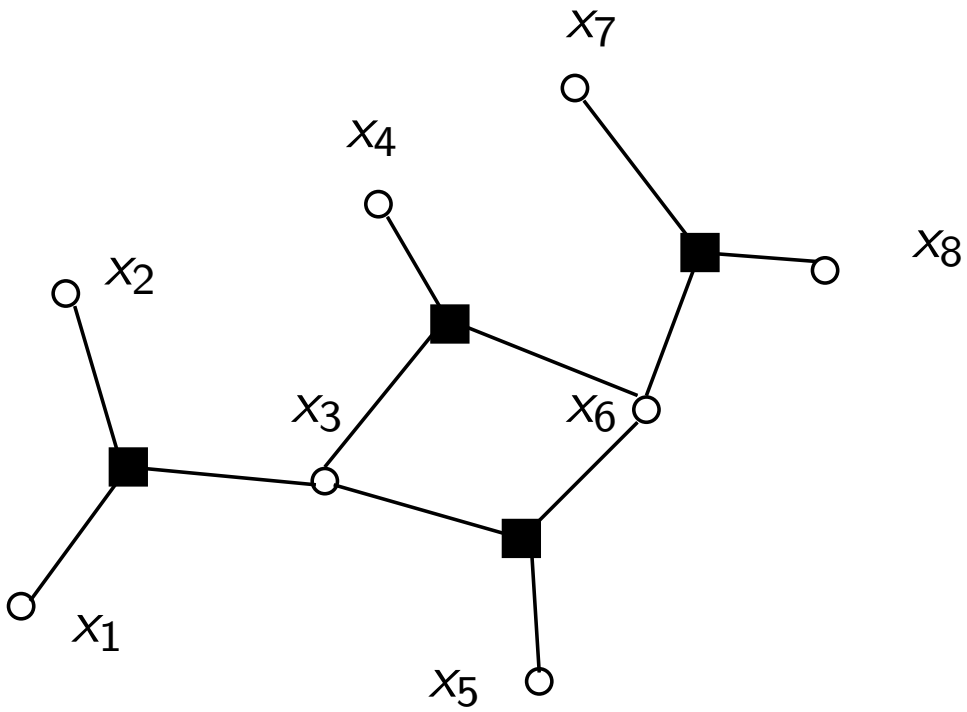
$\mu(x_i = 1) =$ fraction of solutions in which $x_i = 1$

How would you do it in your dreams?

SOLVER	
1:	**Repeat :**
2:	Choose x_i;
3:	$\mu(x_i = 1) = \text{MARGINAL}(i)$;
4 :	Fix $x_i = 1$ with probability $\mu(x_i = 1)$;
5 :	$x_i = 0$ otherwise;

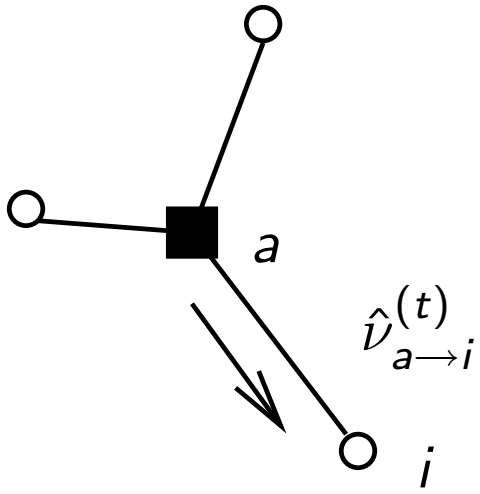
Samples a solution uniformly.

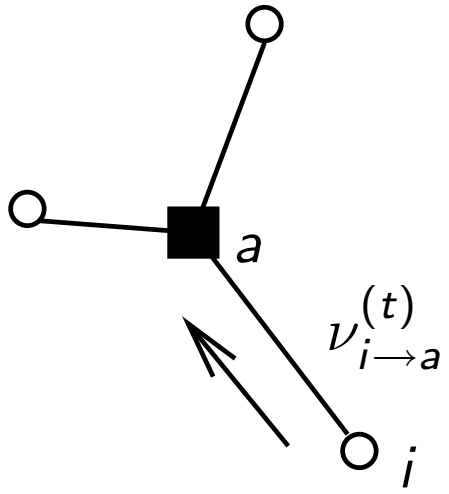
Message passing implementation of MARGINAL(i)



$$(x_1 \vee x_2 \vee \overline{x}_3) \wedge (x_3 \vee \overline{x}_4 \vee \overline{x}_6) \wedge (\overline{x}_3 \vee x_5 \vee \overline{x}_6) \wedge (x_6 \vee x_7 \vee \overline{x}_8)$$

Message passing implementation of MARGINAL(i)

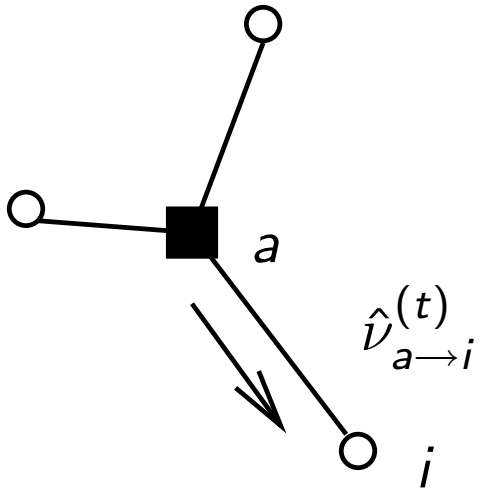


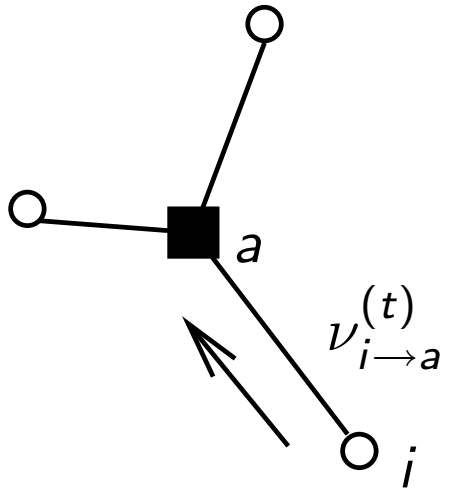


$\nu^{(t)}_{i\to a}, \hat{\nu}^{(t)}_{a\to i} \in \mathfrak{M}(\{0,1\})$ (simplex of prob measures on $\{0,1\}$).

$$\nu^{(t)}_{i\to a} = (\nu^{(t)}_{i\to a}(0), \nu^{(t)}_{i\to a}(1))$$

Message passing implementation of MARGINAL(i)



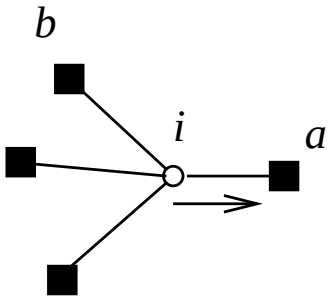


$\nu^{(t)}_{i\to a}, \hat{\nu}^{(t)}_{a\to i} \in \mathfrak{M}(\{0,1\})$ (simplex of prob measures on $\{0,1\}$).

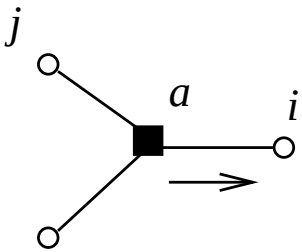
$$\nu^{(t)}_{i\to a} = (\nu^{(t)}_{i\to a}(0), \nu^{(t)}_{i\to a}(1))$$

Message passing implementation of MARGINAL(i)

$s_{jb} \to$ label on edge (j, b)



$$\nu_{i\to a}^{(t+1)}(x_i) \cong \prod_{b\in\partial i\setminus a} \hat{\nu}_{b\to i}^{(t)}(x_i)$$

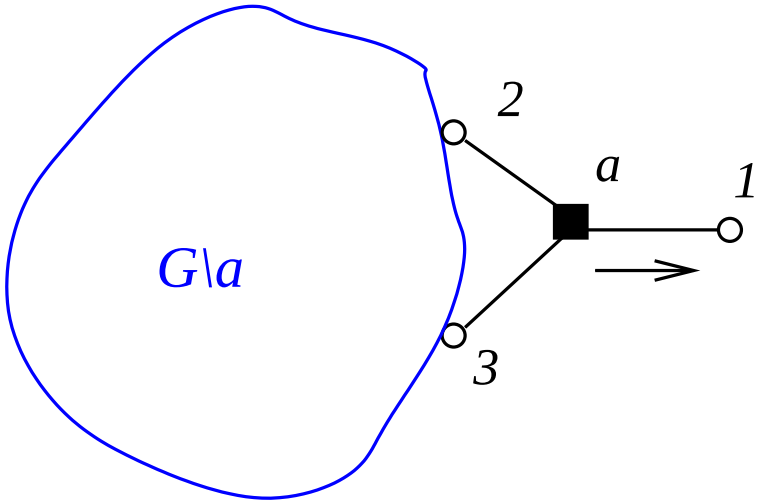


$$\hat{\nu}_{a\to i}^{(t)}(x_i) \cong \begin{cases} 1 - \prod_{j\in\partial a\setminus i} \nu_{j\to a}(x_j = s_{ja}) & \text{if } x_i = s_{ia}, \\ 1 & \text{otherwise.} \end{cases}$$

[Belief propagation]

What are these messy equations?

Clause a: $(x_1 \vee x_2 \vee x_3)$

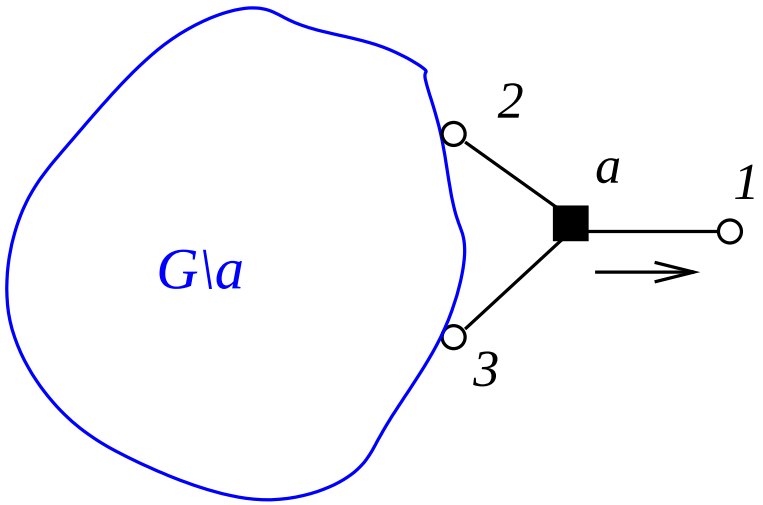


$$N_G(x_A = \cdot) \equiv \text{number of solutions such that } x_A = \cdot\,,$$

$$\mu^G(x_i = 1) = \frac{N_G(x_i = 1)}{N_G(x_i = 0) + N_G(x_i = 1)}\,.$$

What are these messy equations?

Clause a: $(x_1 \vee x_2 \vee x_3)$

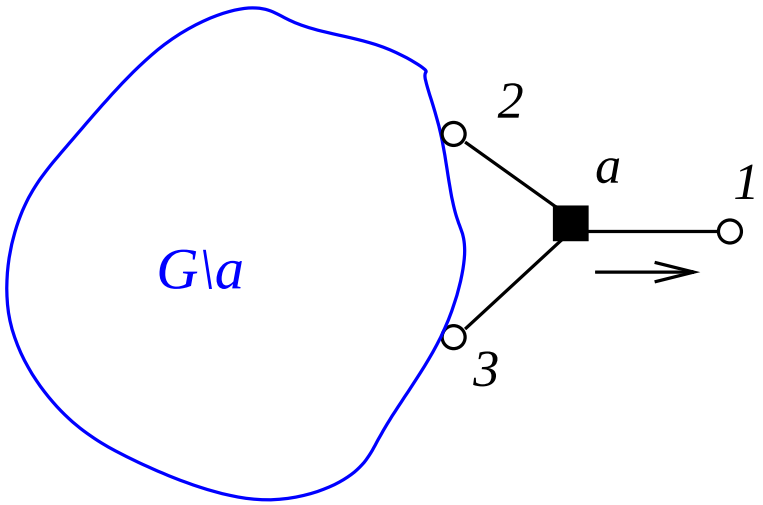


$$
\begin{aligned}
N_G(x_1 = 1) &= N_{G\setminus a}(x_2 = 0, x_3 = 0) + N_{G\setminus a}(x_2 = 0, x_3 = 1) \\
&\quad + N_{G\setminus a}(x_2 = 1, x_3 = 0) + N_{G\setminus a}(x_2 = 1, x_3 = 1)
\end{aligned}
$$

$$
\begin{aligned}
N_G(x_1 = 0) &= N_{G\setminus a}(x_2 = 0, x_3 = 1) + N_{G\setminus a}(x_2 = 1, x_3 = 0) \\
&\quad + N_{G\setminus a}(x_2 = 1, x_3 = 1) \\
&= N_G(x_1 = 1) - N_{G\setminus a}(x_2 = 0, x_3 = 0)
\end{aligned}
$$

What are these messy equations?

Clause a: $(x_1 \vee x_2 \vee x_3)$

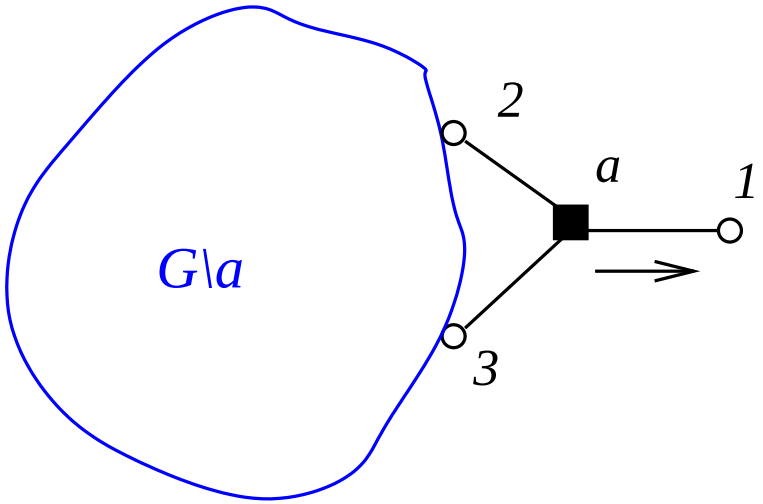


$$
\begin{aligned}
N_G(x_1 = 1) &= N_{G\setminus a}(x_2 = 0, x_3 = 0) + N_{G\setminus a}(x_2 = 0, x_3 = 1) \\
&\quad + N_{G\setminus a}(x_2 = 1, x_3 = 0) + N_{G\setminus a}(x_2 = 1, x_3 = 1)
\end{aligned}
$$

$$
\begin{aligned}
N_G(x_1 = 0) &= N_{G\setminus a}(x_2 = 0, x_3 = 1) + N_{G\setminus a}(x_2 = 1, x_3 = 0) \\
&\quad + N_{G\setminus a}(x_2 = 1, x_3 = 1) \\
&= N_G(x_1 = 1) - N_{G\setminus a}(x_2 = 0, x_3 = 0)
\end{aligned}
$$

What are these messy equations?

Clause a: $(x_1 \vee x_2 \vee x_3)$

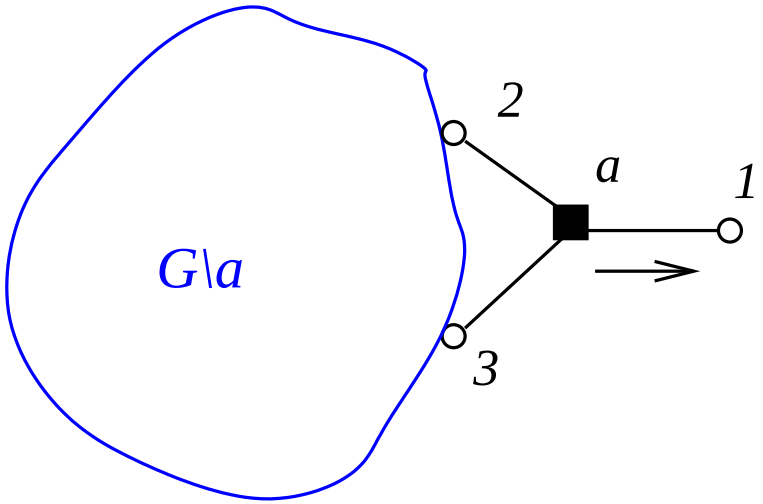


$$\frac{N_G(x_1 = 0)}{N_G(x_1 = 1)} = \frac{\mu^G(x_1 = 0)}{\mu^G(x_1 = 1)} = 1 - \mu^{G\setminus a}(x_2 = 0, x_3 = 0)$$

$$\approx 1 - \mu^{G\setminus a}(x_2 = 0)\mu^{G\setminus a}(x_3 = 0)\,.$$

What are these messy equations?

Clause a: $(x_1 \vee x_2 \vee x_3)$



$$\frac{N_G(x_1=0)}{N_G(x_1=1)} = \frac{\mu^G(x_1=0)}{\mu^G(x_1=1)} = 1-\mu^{G\setminus a}(x_2=0,x_3=0)$$

$$\approx 1-\mu^{G\setminus a}(x_2=0)\mu^{G\setminus a}(x_3=0)\,.$$

BP-guided decimation

BP-GUIDED DECIMATION

1: **Repeat :**
2: Choose x_i;
3: Compute $\mu(x_i = 1)$ using BP;
4 : Fix $x_i = 1$ with probability $\mu(x_i = 1)$;
5 : $x_i = 0$ otherwise;

Is it worth the effort???

Proposition

BP computes the correct marginals for if $\alpha < \alpha_{\rm u}(k)$*, where*

$$\alpha_{\rm u}(k) = \frac{2\log k}{k}\{1 + o_k(1)\} = \alpha_{\rm pl}(k)\{1 + o_k(1)\} \qquad :-($$

Proof.

A general argument, wait 2 minutes. □

Is it worth the effort???

Proposition

BP computes the correct marginals for if $\alpha < \alpha_{\rm u}(k)$, *where*

$$\alpha_{\rm u}(k) = \frac{2\log k}{k}\{1 + o_k(1)\} = \alpha_{\rm pl}(k)\{1 + o_k(1)\} \qquad :-($$

Proof.

A general argument, wait 2 minutes. □

Is it worth the effort???

Proposition

BP computes the correct marginals for if $\alpha < \alpha_{\rm u}(k)$, *where*

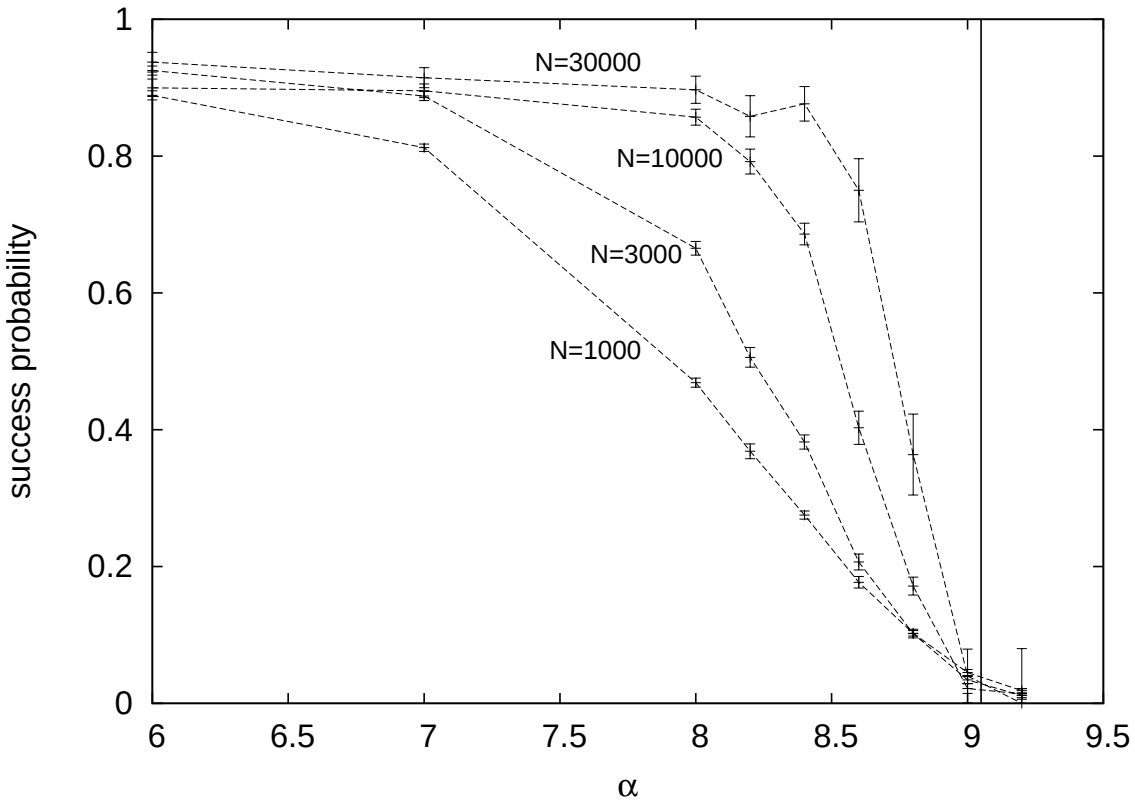
$$\alpha_{\rm u}(k) = \frac{2\log k}{k}\{1 + o_k(1)\} = \alpha_{\rm pl}(k)\{1 + o_k(1)\} \qquad :-($$

Proof.

A general argument, wait 2 minutes. □

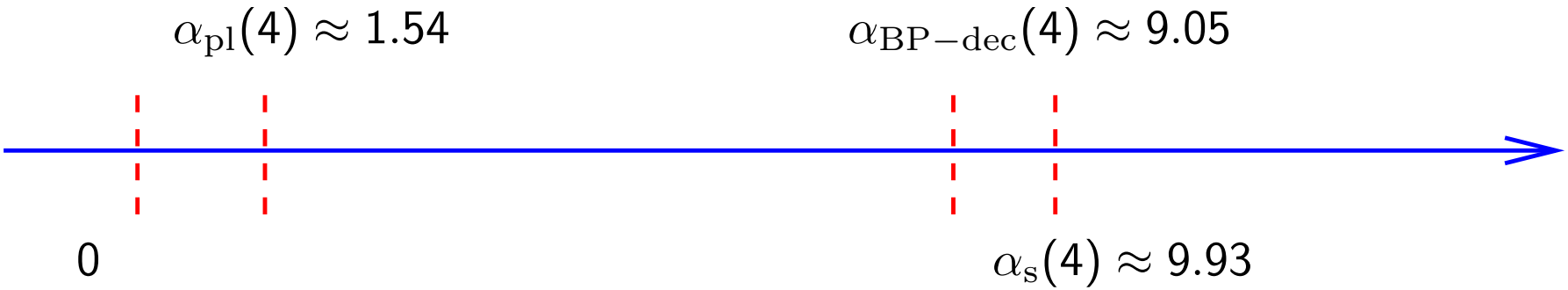
Is it worth the effort???

Is it worth the effort???



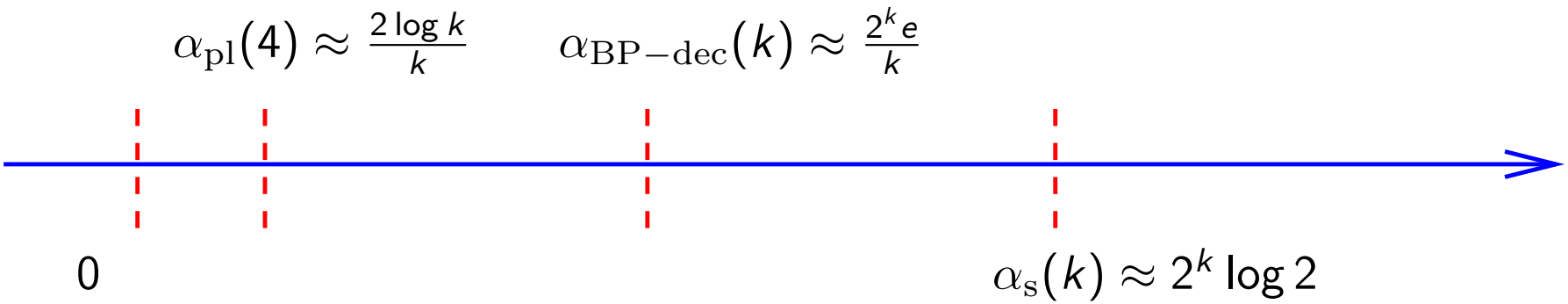
[Montanari-Ricci-Semerjian 2007]

The context: 4-SAT



[Mézard-Parisi-Zecchina 2002 (Conj)]

The context: k-SAT

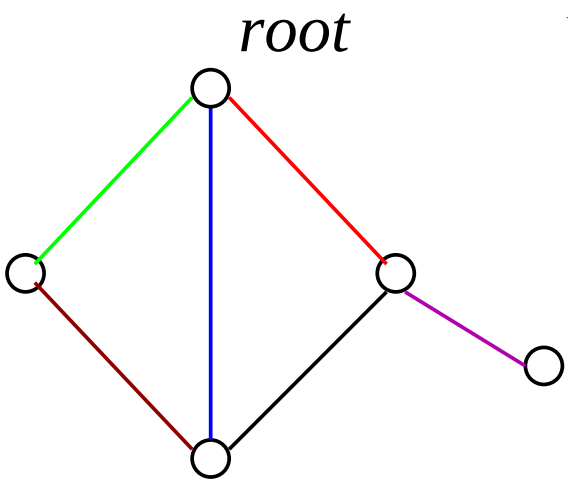


[Achlioptas-Peres 2004]

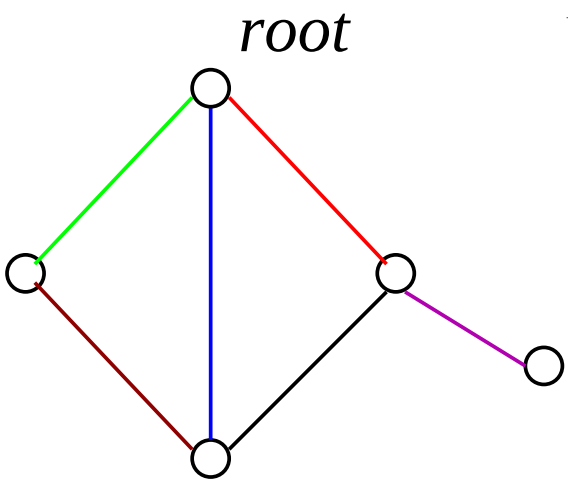
I owed you a general argument

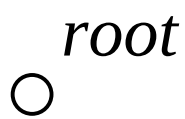
Of trees and loops

Computation tree

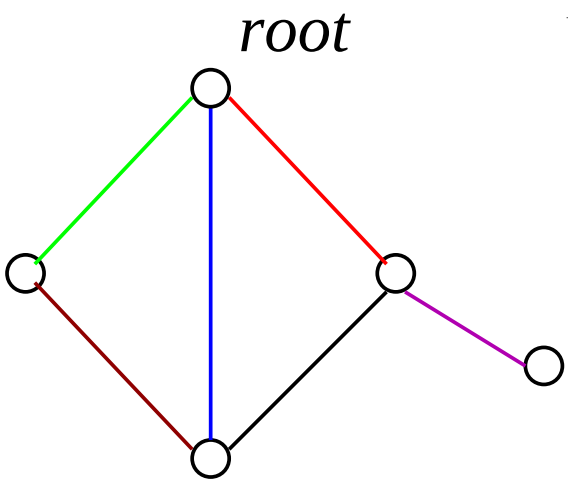


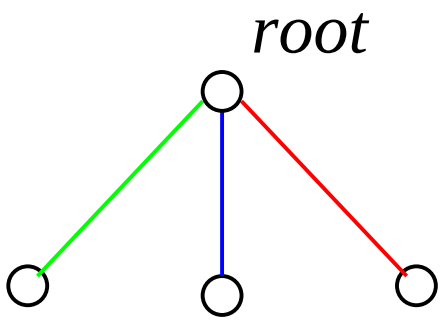
Computation tree, $T_{G,i}(0)$



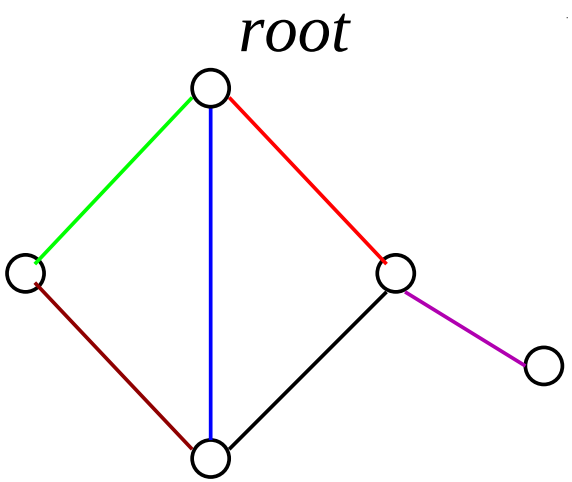


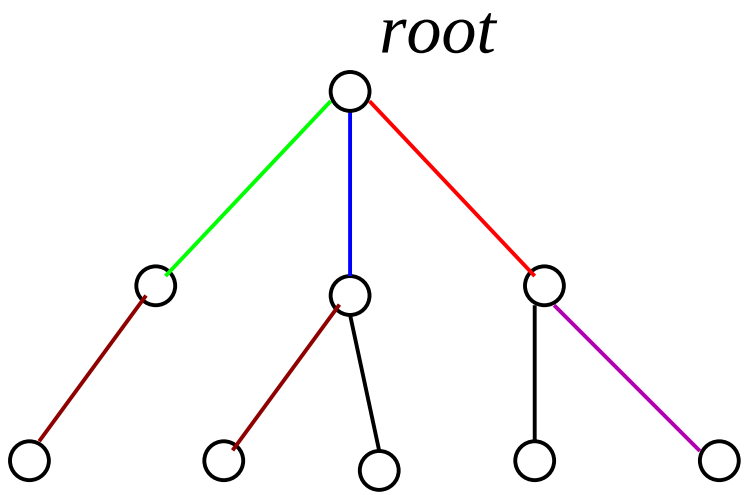
Computation tree, $T_{G,i}(1)$



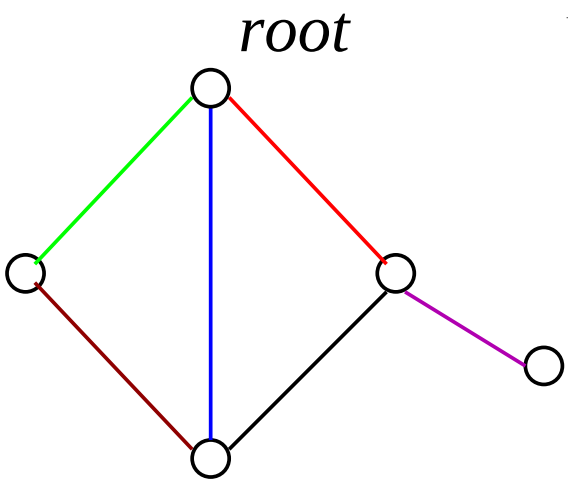


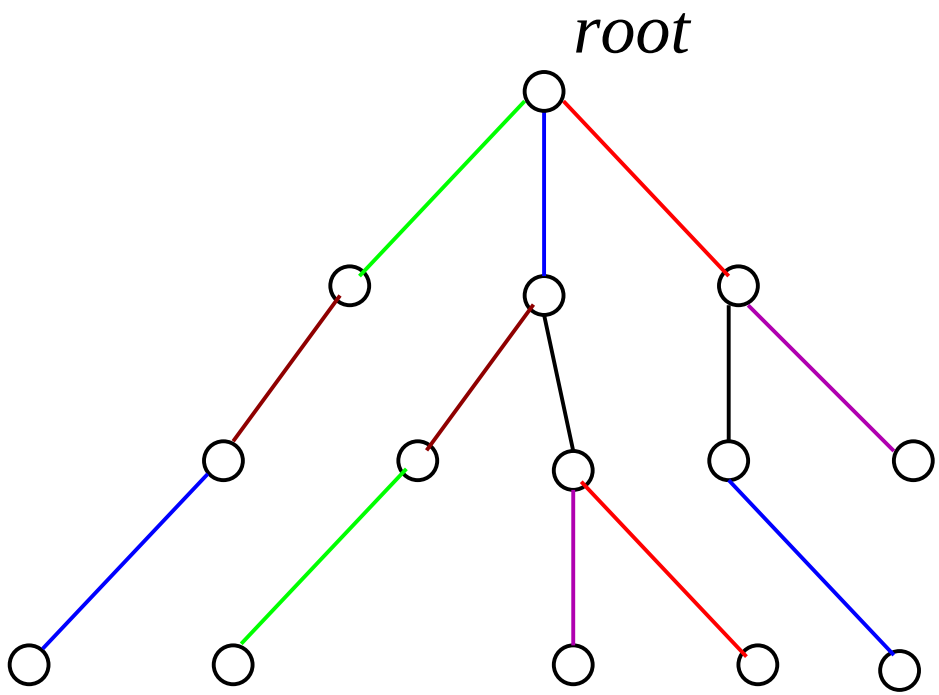
Computation tree, $T_{G,i}(2)$



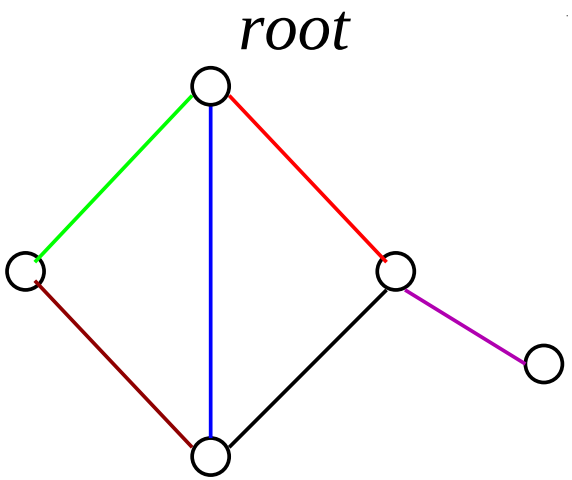


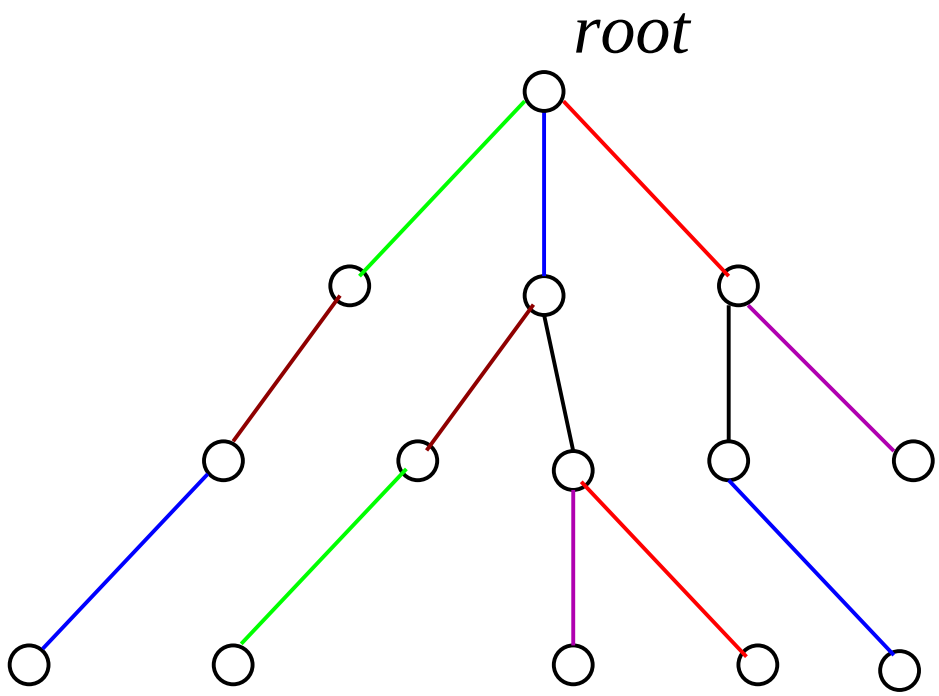
Computation tree, $T_{G,i}(3)$





So what?



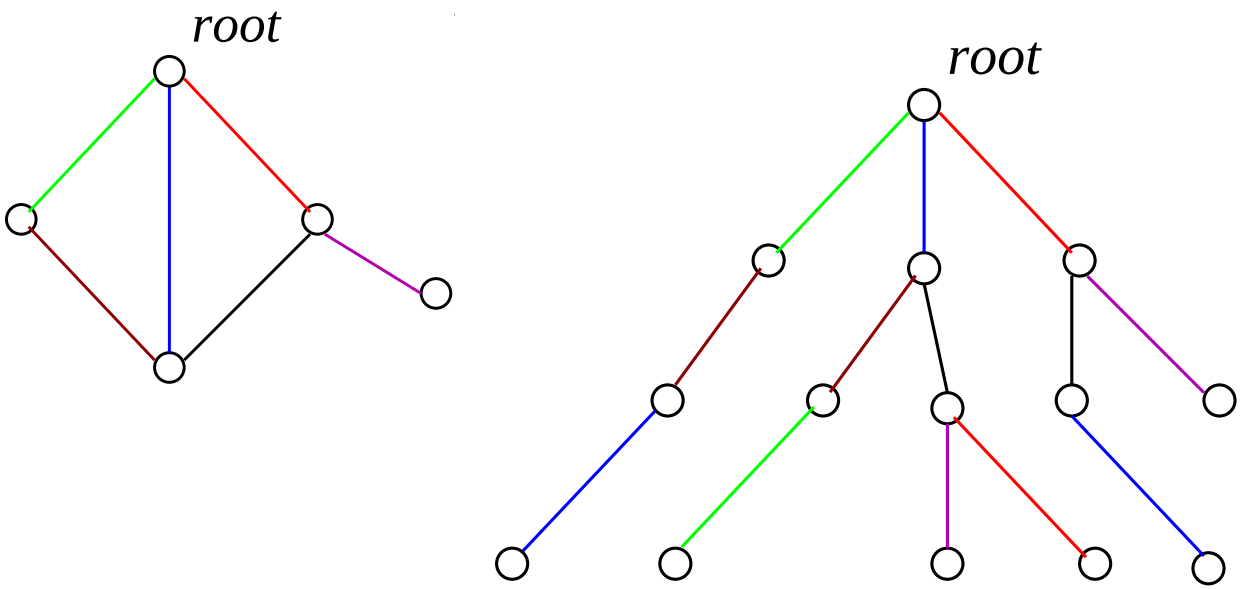


Remark

After t iterations $BP(i,t)$ outputs

$$\mu^{\mathsf{T}_{G,i}(t)}(x_i)$$

And now for the general argument



Theorem (Tatikonda-Jordan 2002)

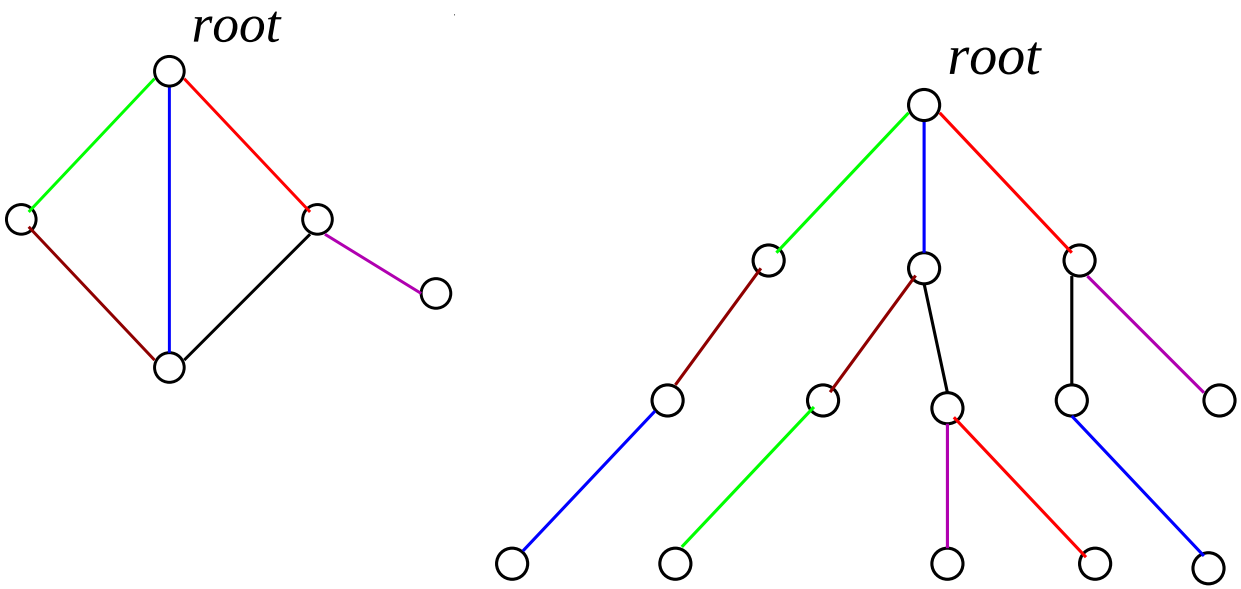
1. *If the Gibbs measure on* $\mathsf{T}_{G,i}(\infty)$ *is unique then BP converges.*

2. *Further, if* G *has large girth, then* $\mu^{\mathrm{BP}(t)}(x_i) = \mu^{\mathsf{T}_{G,i}(t)}(x_i) \approx \mu^{G}(x_i)$.

Proof.

By definition. □

And now for the general argument



Theorem (Tatikonda-Jordan 2002)

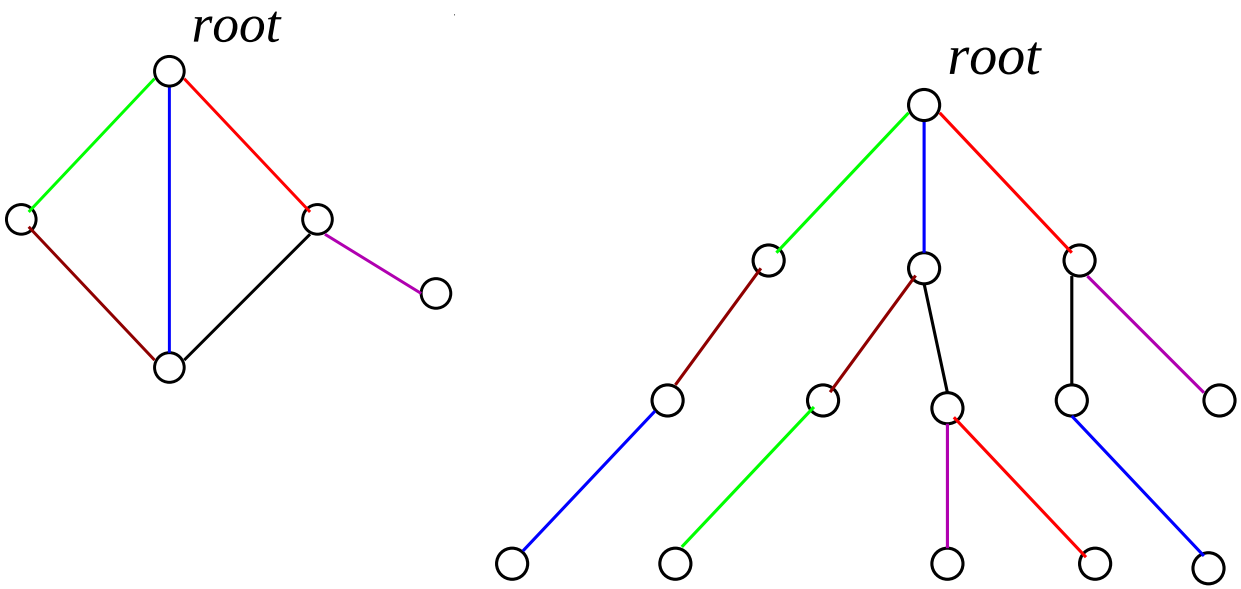
1. *If the Gibbs measure on* $\mathsf{T}_{G,i}(\infty)$ *is unique then BP converges.*

2. *Further, if* G *has large girth, then* $\mu^{\mathrm{BP(t)}}(x_i) = \mu^{\mathsf{T}_{G,i}(t)}(x_i) \approx \mu^{G}(x_i)$.

Proof.

By definition. □

And now for the general argument



Theorem (Tatikonda-Jordan 2002)

1. *If the Gibbs measure on* $\mathsf{T}_{G,i}(\infty)$ *is unique then BP converges.*

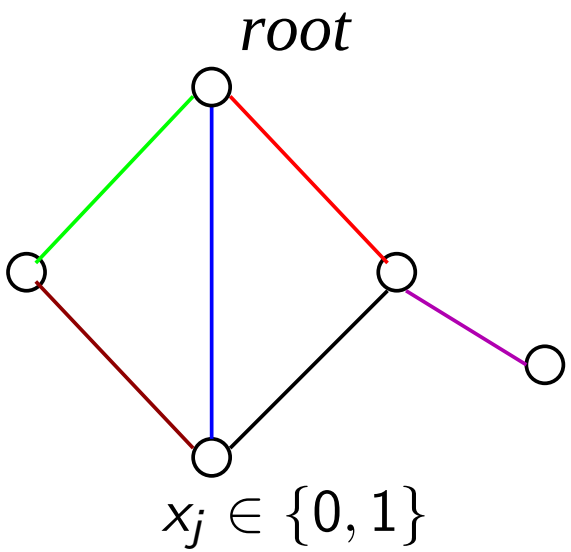
2. *Further, if* G *has large girth, then* $\mu^{\mathrm{BP(t)}}(x_i) = \mu^{\mathsf{T}_{G,i}(t)}(x_i) \approx \mu^{G}(x_i)$.

Proof.

By definition. □

What about graphs with many short loops?

Example: Independent sets



$$\mu^G(x) = \frac{1}{Z} \prod_{(i,j)\in E} \mathbb{I}(x_i \neq x_j)\ \lambda^{|x|}$$

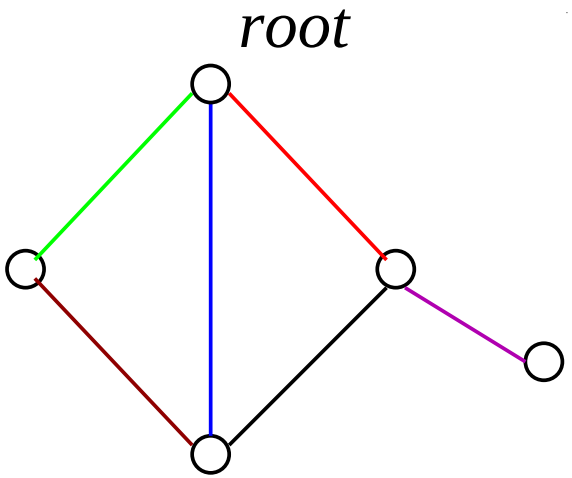
Weitz's Self Avoiding Walk tree

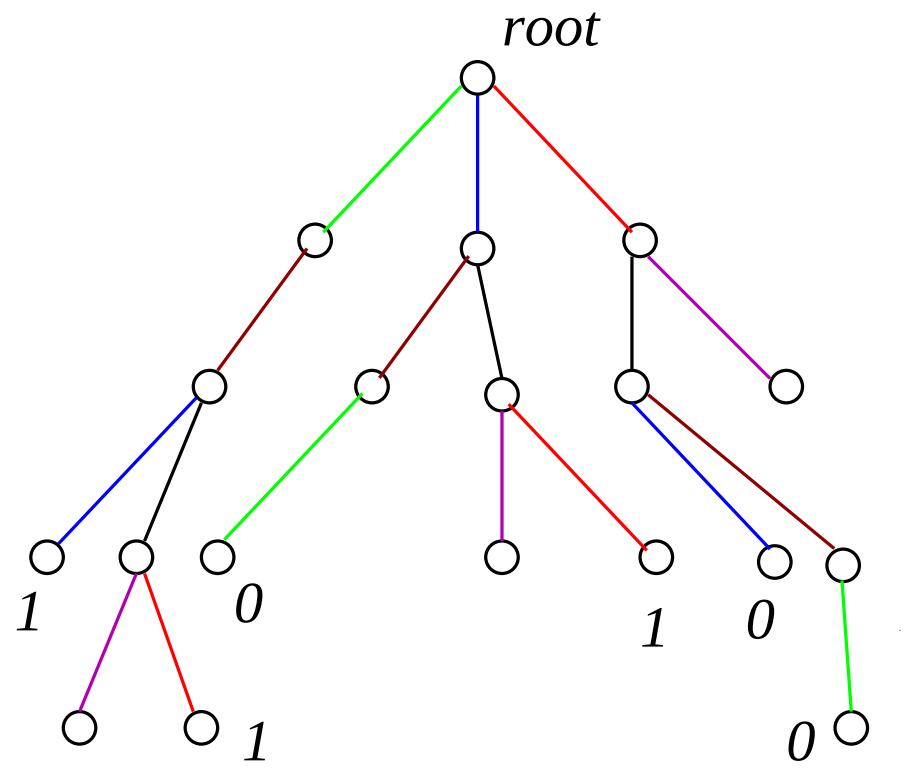
$$T^{\mathrm{SAW}}_{G,i} = \text{Truncation of } T_{G,i}(\infty) + \text{boundary conditions}$$

Lemma

$$\mu^{T^{\mathrm{SAW}}_{G,i}}(x_i) = \mu^{G}(x_i)\,.$$

Weitz's Self Avoiding Walk tree





As you can see... $|\mathsf{T}^{\mathrm{SAW}}_{G,i}| = \exp\{\Theta(n)\}$

An algorithm

Truncate $\mathsf{T}^{\mathrm{SAW}}_{G,i}$ at depth $\Theta(\log n)$

Theorem (Weitz 2006)

Assume G has degree bounded by k, and Gibbs measure on k-regular trees is unique.
Approximate counting can be performed in polynomial time.

An algorithm

Truncate $\mathsf{T}^{\mathrm{SAW}}_{G,i}$ at depth $\Theta(\log n)$

Theorem (Weitz 2006)

Assume G has degree bounded by k, and Gibbs measure on k-regular trees is unique.
Approximate counting can be performed in polynomial time.

How general is this strategy?

Theorem (Gamarnik-Katz 2007)

Pretty general.

[Uses appropriate 'backtracking' tree]

How practical is this strategy?

Complexity $= \Theta(n^{\text{big exponent}})$

big exponent depends on b

$$\sup_{x,x'} \left| \mu^{\mathsf{T}}(x_\emptyset = 1 | x_{\partial \mathsf{T}(t)}) - \mu^{\mathsf{T}}(x_\emptyset = 1 | x_{\partial \mathsf{T}(t)}) \right| \le A\, e^{-b\, t}\,.$$

[Lu-Measson-Montanari 2008]

How practical is this strategy?

Complexity $= \Theta(n^{\text{big exponent}})$

big exponent depends on b

$$\sup_{x,x'} \left| \mu^{\mathsf{T}}(x_\emptyset = 1 | x_{\partial \mathsf{T}(t)}) - \mu^{\mathsf{T}}(x_\emptyset = 1 | x_{\partial \mathsf{T}(t)}) \right| \le A\, e^{-b\, t}\,.$$

[Lu-Measson-Montanari 2008]

How practical is this strategy?

Complexity $= \Theta(n^{\text{big exponent}})$

big exponent depends on b

$$\sup_{x,x'} \left| \mu^{\mathsf{T}}(x_\emptyset = 1 | x_{\partial \mathsf{T}(t)}) - \mu^{\mathsf{T}}(x_\emptyset = 1 | x_{\partial \mathsf{T}(t)}) \right| \le A\, e^{-b\, t}\,.$$

[Lu-Measson-Montanari 2008]

Beyond uniqueness?

Why should we hope for more?

Example: NAE-SAT

One clause: $(x_1 \vee \overline{x}_{16} \vee x_{71}) \wedge (\overline{x}_1 \vee x_{16} \vee \overline{x}_{71})$

$$\mu^G(x) = \text{uniform measure over solutions}$$

Why should we hope for more?

Theorem (Montanari-Restrepo-Tetali 2009)

For $\alpha \le \alpha_*(k) = 2^{k-1}\log 2\{1 + o_k(1)\}$,

$$(G_n, \mu^{G_n}) \stackrel{\text{locally}}{\longrightarrow} (\mathsf{T}, \mu^{\mathsf{T}})\,.$$

Theorem (Achlioptas-Moore 2002)

For NAE-SAT

$$\alpha_{\rm s}(k) = 2^{k-1}\log 2\big\{1 + o_k(1)\big\}\,.$$

Proof: Second moment method.

Why should we hope for more?

Theorem (Montanari-Restrepo-Tetali 2009)

For $\alpha \le \alpha_*(k) = 2^{k-1}\log 2\{1 + o_k(1)\}$,

$$(G_n, \mu^{G_n}) \stackrel{\text{locally}}{\longrightarrow} (\mathsf{T}, \mu^{\mathsf{T}})\,.$$

Theorem (Achlioptas-Moore 2002)

For NAE-SAT

$$\alpha_{\rm s}(k) = 2^{k-1}\log 2\big\{1 + o_k(1)\big\}\,.$$

Proof: Second moment method.

Why should we hope for more?

Theorem (Montanari-Restrepo-Tetali 2009)

For $\alpha \le \alpha_*(k) = 2^{k-1}\log 2\{1 + o_k(1)\}$,

$$(G_n, \mu^{G_n}) \stackrel{\text{locally}}{\longrightarrow} (\mathsf{T}, \mu^{\mathsf{T}}) .$$

Theorem (Achlioptas-Moore 2002)

For NAE-SAT

$$\alpha_{\rm s}(k) = 2^{k-1}\log 2\big\{1 + o_k(1)\big\} .$$

Proof: Second moment method.

What is happening? Notions of correlation decay

Uniqueness:

$$\sup_{x,x'} \sum_{x_i} \left| \mu(x_i | x_{\sim i,r}) - \mu(x_i | x'_{\sim i,r}) \right| \to 0$$

Extremality:

$$\sum_{x_i, x_{\sim i,r}} \left| \mu(x_i, x_{\sim i,r}) - \mu(x_i)\mu(x_{\sim i,r}) \right| \to 0$$

Concentration:

$$\sum_{x_{i(1)} \cdots x_{i(\ell)}} \left| \mu(x_{i(1)}, \dots, x_{i(\ell)}) - \mu(x_{i(1)}) \cdots \mu(x_{i(\ell)}) \right| \to 0$$

What is happening? Notions of correlation decay

Uniqueness:

$$\sup_{x,x'} \sum_{x_i} \left| \mu(x_i | x_{\sim i,r}) - \mu(x_i | x'_{\sim i,r}) \right| \to 0$$

Extremality:

$$\sum_{x_i, x_{\sim i,r}} |\mu(x_i, x_{\sim i,r}) - \mu(x_i)\mu(x_{\sim i,r})| \to 0$$

Concentration:

$$\sum_{x_{i(1)} \cdots x_{i(\ell)}} \left| \mu(x_{i(1)}, \dots, x_{i(\ell)}) - \mu(x_{i(1)}) \cdots \mu(x_{i(\ell)}) \right| \to 0$$

What is happening? Notions of correlation decay

Uniqueness:

$$\sup_{x,x'} \sum_{x_i} \left| \mu(x_i | x_{\sim i,r}) - \mu(x_i | x'_{\sim i,r}) \right| \to 0$$

Extremality:

$$\sum_{x_i, x_{\sim i,r}} |\mu(x_i, x_{\sim i,r}) - \mu(x_i)\mu(x_{\sim i,r})| \to 0$$

Concentration:

$$\sum_{x_{i(1)} \cdots x_{i(\ell)}} \left| \mu(x_{i(1)}, \dots, x_{i(\ell)}) - \mu(x_{i(1)}) \cdots \mu(x_{i(\ell)}) \right| \to 0$$

What is happening? Notions of correlation decay

Uniqueness:

$$\sup_{x,x'} \sum_{x_i} \left| \mu(x_i | x_{\sim i,r}) - \mu(x_i | x'_{\sim i,r}) \right| \to 0$$

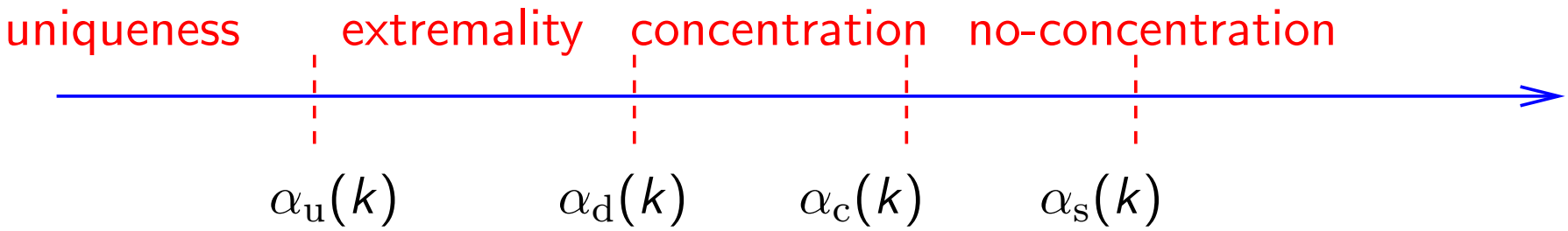
Extremality:

$$\sum_{x_i, x_{\sim i,r}} |\mu(x_i, x_{\sim i,r}) - \mu(x_i)\mu(x_{\sim i,r})| \to 0$$

Concentration:

$$\sum_{x_{i(1)} \cdots x_{i(\ell)}} \left| \mu(x_{i(1)}, \dots, x_{i(\ell)}) - \mu(x_{i(1)}) \cdots \mu(x_{i(\ell)}) \right| \to 0$$

What happens in k-SAT?



$\alpha_{\rm u}(k) = (2\log k)/k + \dots$ [proved]

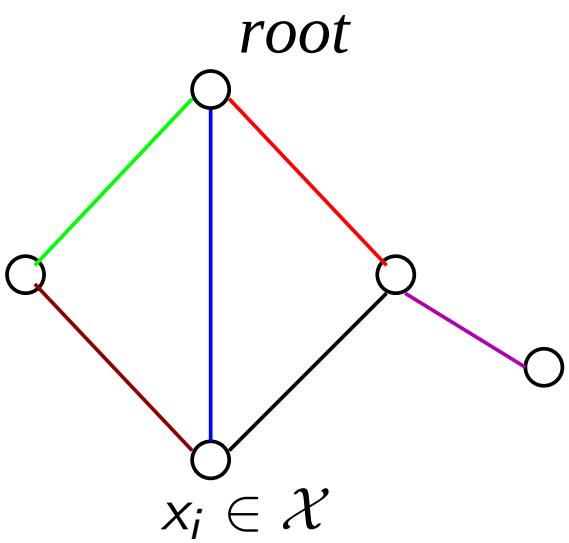
$\alpha_{\rm d}(k) = (2^k \log k)/k + \dots$ $(\alpha_{\rm d}(4) \approx 9.38)$ [proved in NAE-SAT]

$\alpha_{\rm c}(k) = 2^k \log 2 - \frac{3}{2}\log 2 + \dots$ $(\alpha_{\rm c}(4) \approx 9.547)$

$\alpha_{\rm s}(k) = 2^k \log 2 - \frac{1}{2}(1+\log 2) + \dots$ $(\alpha_{\rm s}(4) \approx 9.93)$

So what?

So what? Bethe-Peierls 'approximation'

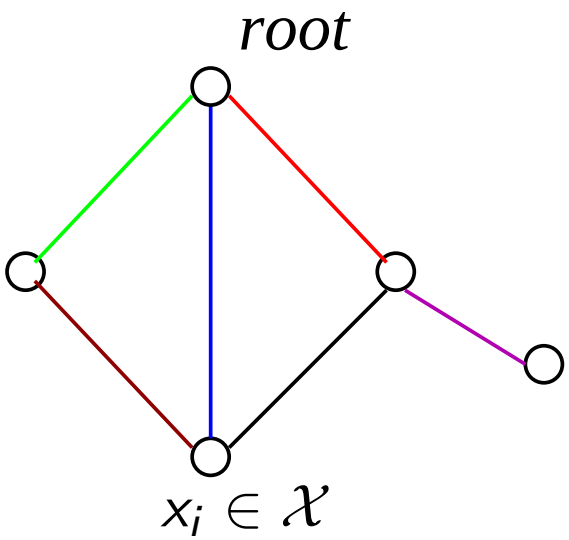


$$\mu(x) = \prod_{(i,j)\in E} \psi_{i,j}(x_i, x_j)/Z$$

Definition

A 'set of messages' is a collection $\{\nu_{i\to j}(\,\cdot\,)\}$ indexed by directed edges in G, where $\nu_{i\to j} \in \mathfrak{M}(\mathcal{X})$.

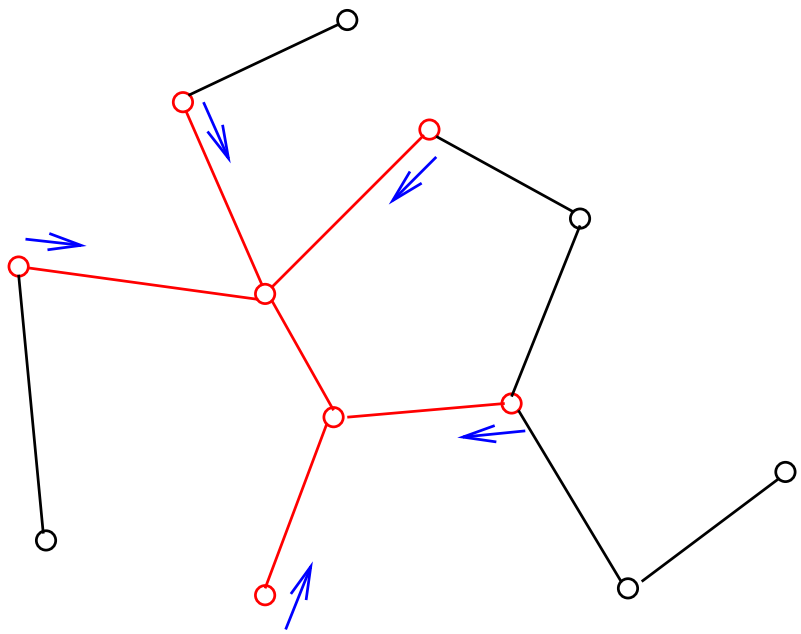
So what? Bethe-Peierls 'approximation'



$$\mu(x) = \prod_{(i,j)\in E} \psi_{i,j}(x_i, x_j)/Z$$

Definition

A 'set of messages' is a collection $\{\nu_{i\to j}(\,\cdot\,)\}$ *indexed by directed edges in* G*, where* $\nu_{i\to j} \in \mathfrak{M}(\mathcal{X})$.



Given $F \subseteq G$, $\mathrm{diam}(F) \le 2\ell \equiv$ girth, such that $\deg_F(i) = \deg_G(i)$ or ≤ 1

$$\nu_U(x_U) \equiv \frac{1}{C(\nu_U)} \prod_{(ij)\in F} \psi_{ij}(x_i, x_j) \prod_{i\in\partial F} \nu_{i\to j(i)}(x_i)\,.$$

Bethe states

Definition

A probability distribution ρ on $\mathcal{X}^V$ is an (ε, r) Bethe state, if there exists a set of messages $\{\nu_{i\to j}(\,\cdot\,)\}$ such that, for any $F \subseteq G$ with $\mathrm{diam}(F) \le 2r$

$$||\rho_U - \nu_U||_{TV} \le \varepsilon\,.$$

Theorem (Dembo-Montanari 2009)

If μ is extremal 'with rate $\delta(\,\cdot\,)$' then it an (ε, r) Bethe state for any $r < \ell$ and $\varepsilon \ge C\delta(\ell - r)$.

Bethe states

Definition

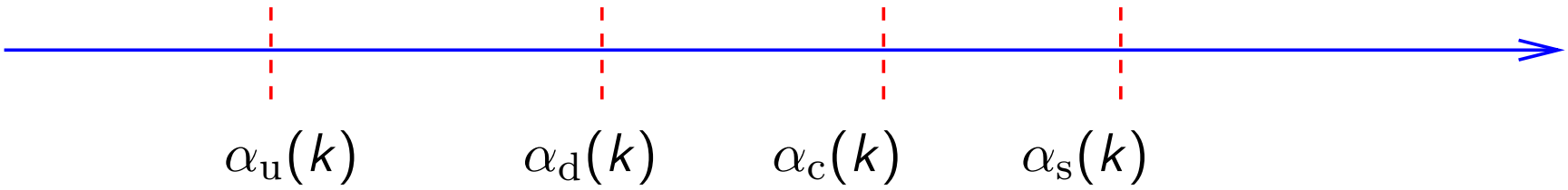
A probability distribution ρ on $\mathcal{X}^V$ is an (ε, r) Bethe state, if there exists a set of messages $\{\nu_{i\to j}(\,\cdot\,)\}$ such that, for any $F \subseteq G$ with $\mathrm{diam}(F) \le 2r$

$$||\rho_U - \nu_U||_{TV} \le \varepsilon\,.$$

Theorem (Dembo-Montanari 2009)

If μ is extremal 'with rate $\delta(\,\cdot\,)$' then it an (ε, r) Bethe state for any $r < \ell$ and $\varepsilon \ge C\delta(\ell - r)$.

Algorithms vs correlation decay

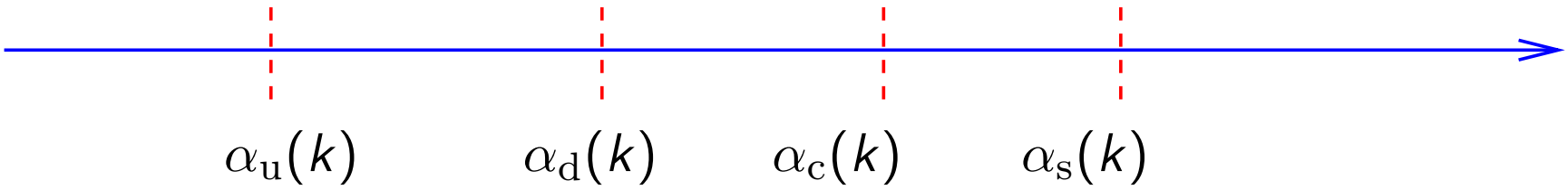


Conjecture

BP-guided decimation finds solutions up to $\alpha_*(k) \approx \alpha_{\rm d}(k)$.

Currently huge gap!

Algorithms vs correlation decay

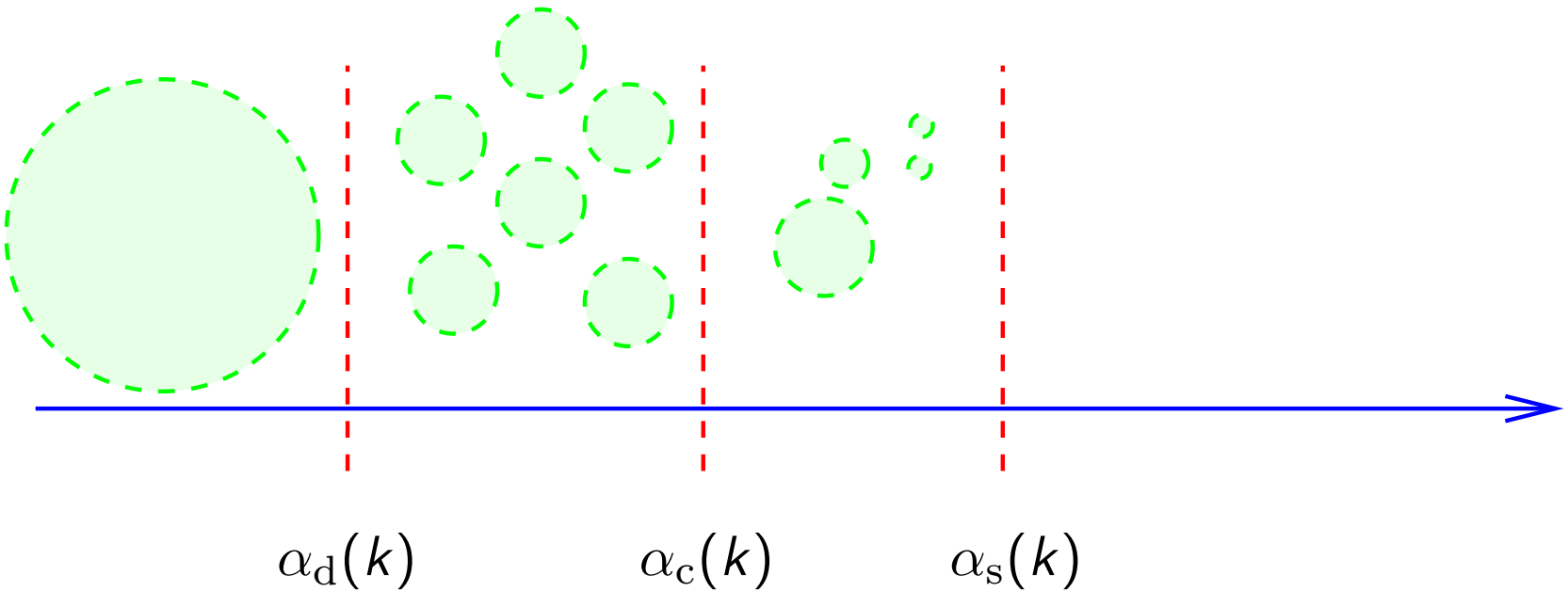


Conjecture

BP-guided decimation finds solutions up to $\alpha_*(k) \approx \alpha_{\rm d}(k)$.

Currently huge gap!

Relation with the structure of the solution space



[Biroli-Monasson-Weigt 1999]
[Mézard-Parisi-Zecchina 2001]
[Krzakala et al. 2007]
[Achlioptas-Coja 2009]

A recent application

Noisy undedetermined linear systems

$$y = A\,x_0 + w$$

Estimate $x_0 \in \mathbb{R}^N$ given (y, A).

[Signal processing: Donoho, Candes, Tao,...]
[Sketching: Indyk, Gilbert, Muthu,...]

Noisy undedetermined linear systems

$$y = A\,x_0 + w$$

Estimate $x_0 \in \mathbb{R}^N$ given (y, A).

[Signal processing: Donoho, Candes, Tao,...]
[Sketching: Indyk, Gilbert, Muthu,...]

Noisy undedetermined linear systems

$$y = A\,x_0 + w$$

Estimate $x_0 \in \mathbb{R}^N$ given (y, A).

[Signal processing: Donoho, Candes, Tao,...]
[Sketching: Indyk, Gilbert, Muthu,...]

The LASSO

$$\widehat{x}(y, A) = \operatorname{argmin}_{x \in \mathbb{R}^N} C_{A,y}(x)$$

$$C_{A,y}(x) = \lambda \|x\|_1 + \frac{1}{2}\|y - Ax\|_2^2$$

[Tibshirani 96; Chen, Donoho 95; 1000+ papers]

Wonderful, but...

$\longrightarrow$ What performance should I expect?

$\longrightarrow$ How am I supposed to choose $\mathcal{C}_{A,y}$?

$\longrightarrow$ What if I can design A?

$\longrightarrow$ Low-complexity algorithms?

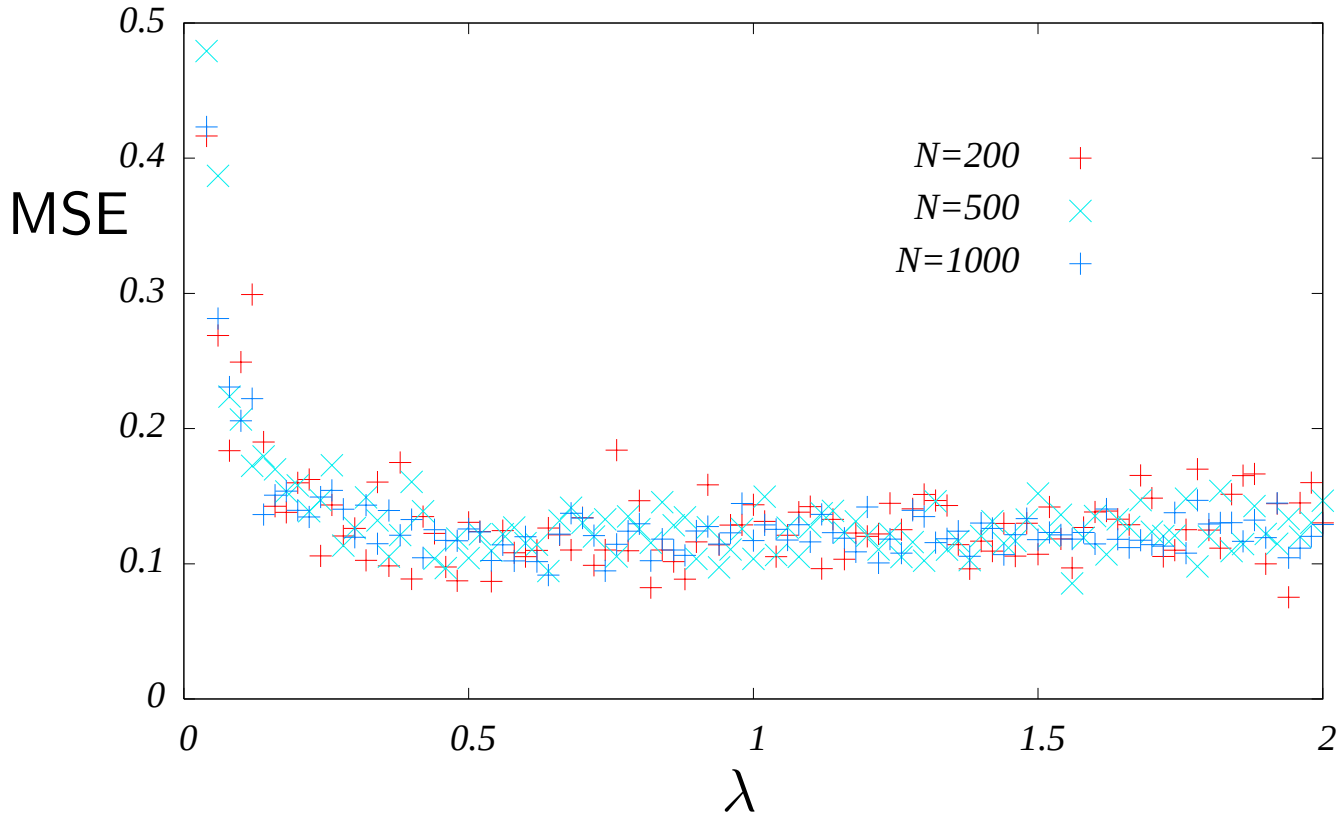
A little experiment

$A \to$ 'real' data

$$x_{0,i} = \begin{cases} +1 & \text{with prob. } 0.064, \\ 0 & \text{with prob. } 0.872, \\ -1 & \text{with prob. } 0.064, \end{cases}$$

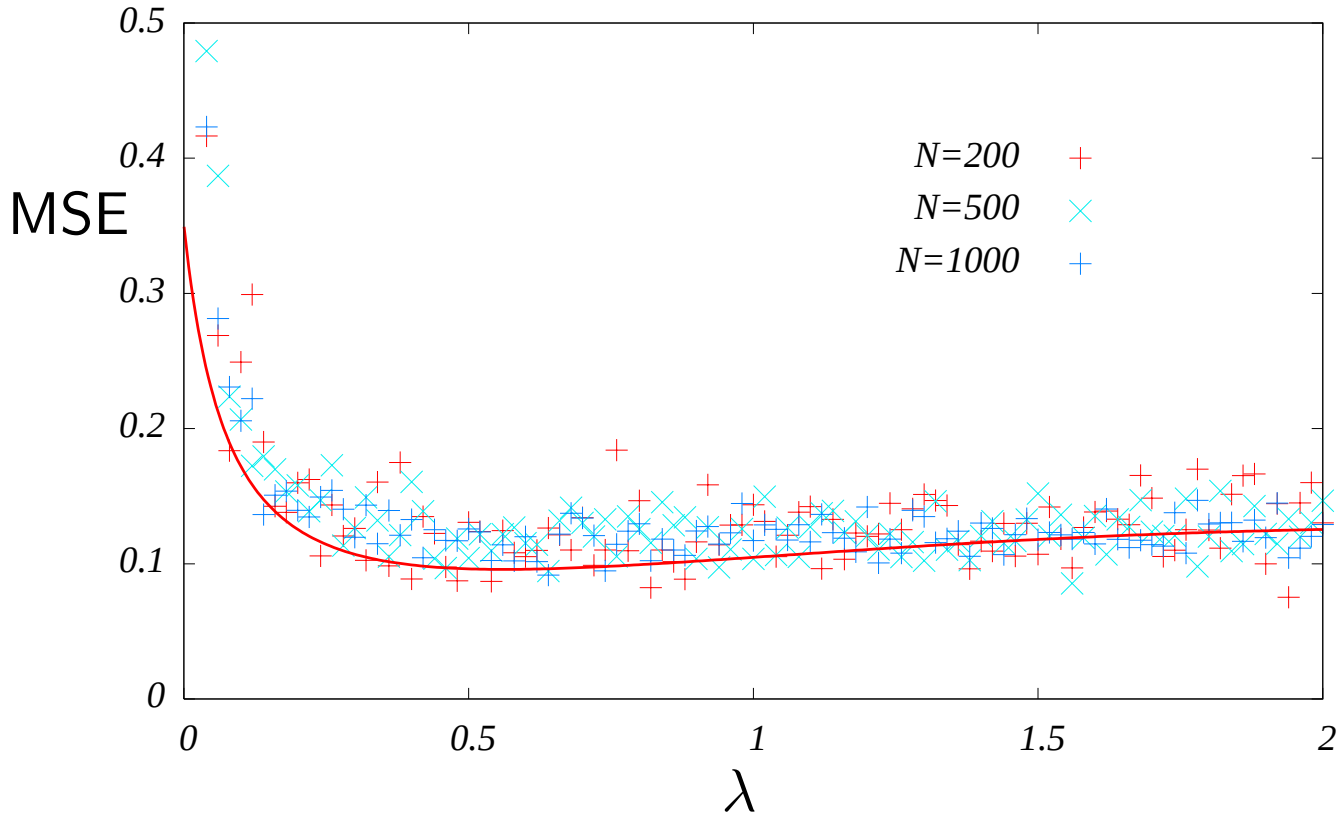
$$w_i \sim \mathrm{N}(0, 0.2)$$

Clinical data



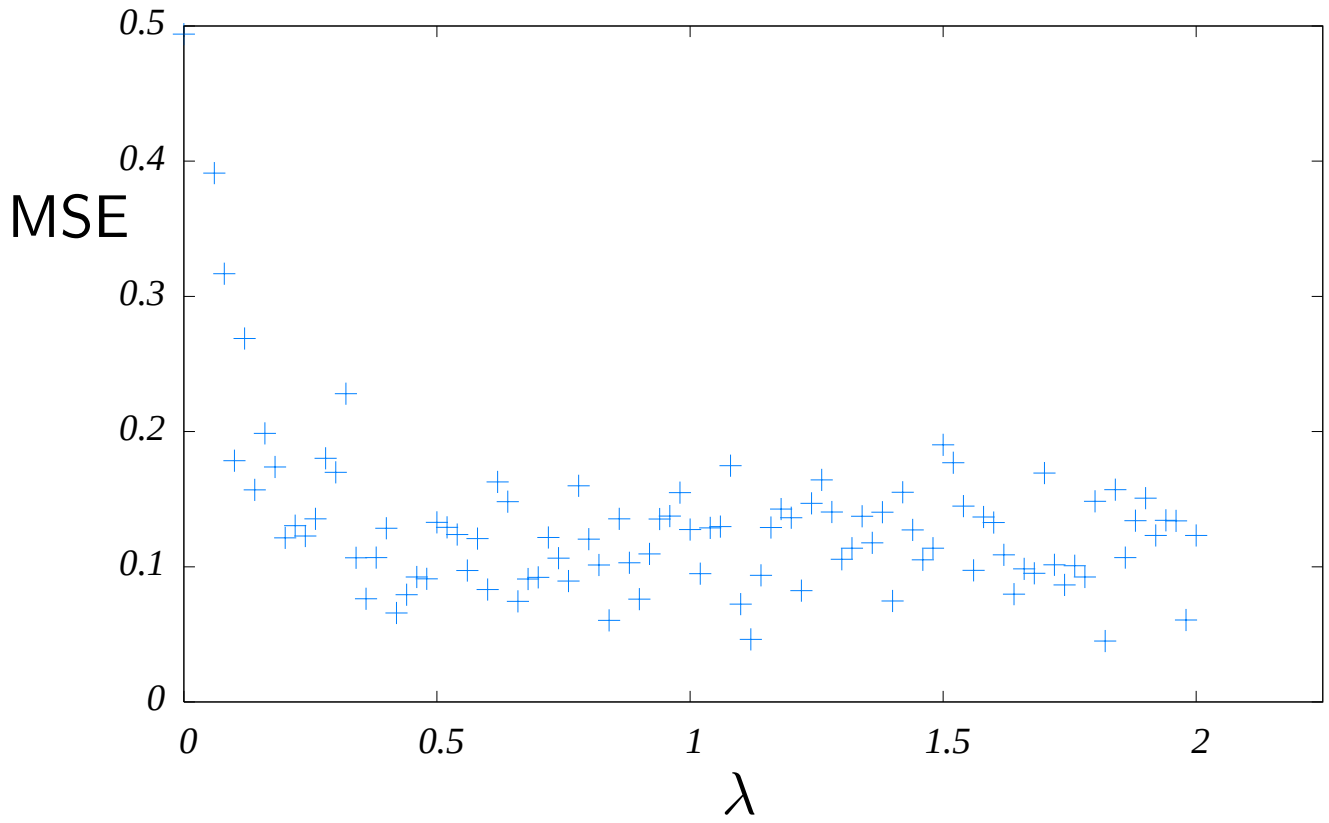
A is $n \times N$, $n = 0.64N$

Clinical data



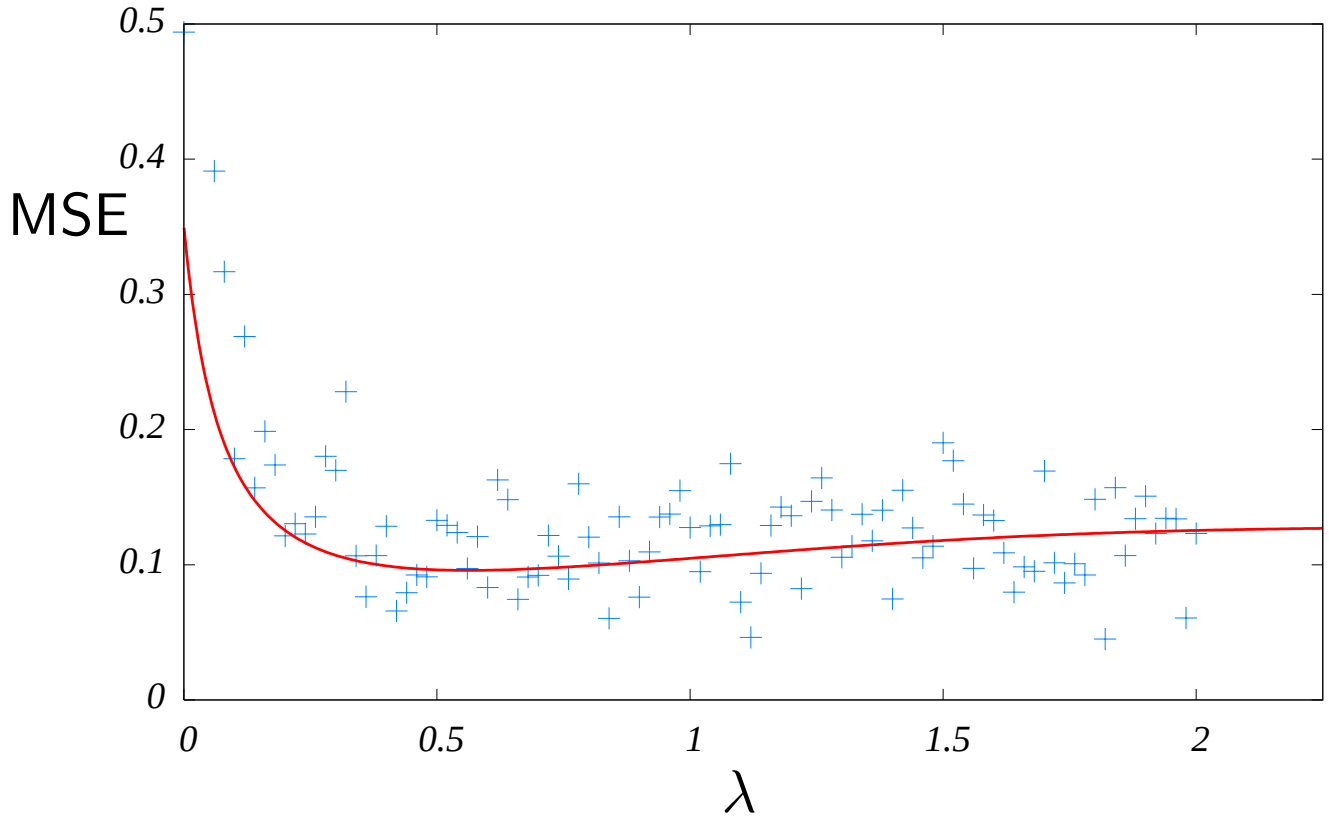
A is $n \times N$, $n = 0.64N$

Gene expression data



A is 85×200 [from Hastie, Tibshirani, Friedman]

Gene expression data



A is 85×200 [from Hastie, Tibshirani, Friedman]

A theorem

Theorem (Bayati, Montanari, 2010)

Assume $A_{ij} \sim \mathrm{N}(0, 1/n)$, $y = Ax_0 + w$, *and* $(\tau_\infty^2, \theta_\infty)$ *unique solution of*

$$\begin{aligned} \tau_\infty^2 &= \sigma^2 + \frac{1}{\delta}\mathbb{E}\{[\eta(X_0 + \tau_\infty Z; \theta_\infty) - X_0]^2\}\,, \\ \lambda &= \theta_\infty\{1 - \frac{1}{\delta}\mathbb{E}[\eta'(X_0 + \tau_\infty Z; \theta_\infty)]\} \end{aligned}$$

Then,

$$\lim_{N\to\infty} \frac{1}{N}\|\widehat{x}^{\mathrm{LASSO}}(\lambda) - x_0\|^2 = (\tau_\infty^2 - \sigma^2)\delta\,.$$

almost surely as $n \to \infty$.

Conjectured in a more general context with Donoho and Maleki

A theorem

Theorem (Bayati, Montanari, 2010)

Assume $A_{ij} \sim \mathrm{N}(0, 1/n)$, $y = Ax_0 + w$, *and* $(\tau_\infty^2, \theta_\infty)$ *unique solution of*

$$\begin{aligned} \tau_\infty^2 &= \sigma^2 + \frac{1}{\delta}\mathbb{E}\{[\eta(X_0 + \tau_\infty Z; \theta_\infty) - X_0]^2\}\,, \\ \lambda &= \theta_\infty\{1 - \frac{1}{\delta}\mathbb{E}[\eta'(X_0 + \tau_\infty Z; \theta_\infty)]\} \end{aligned}$$

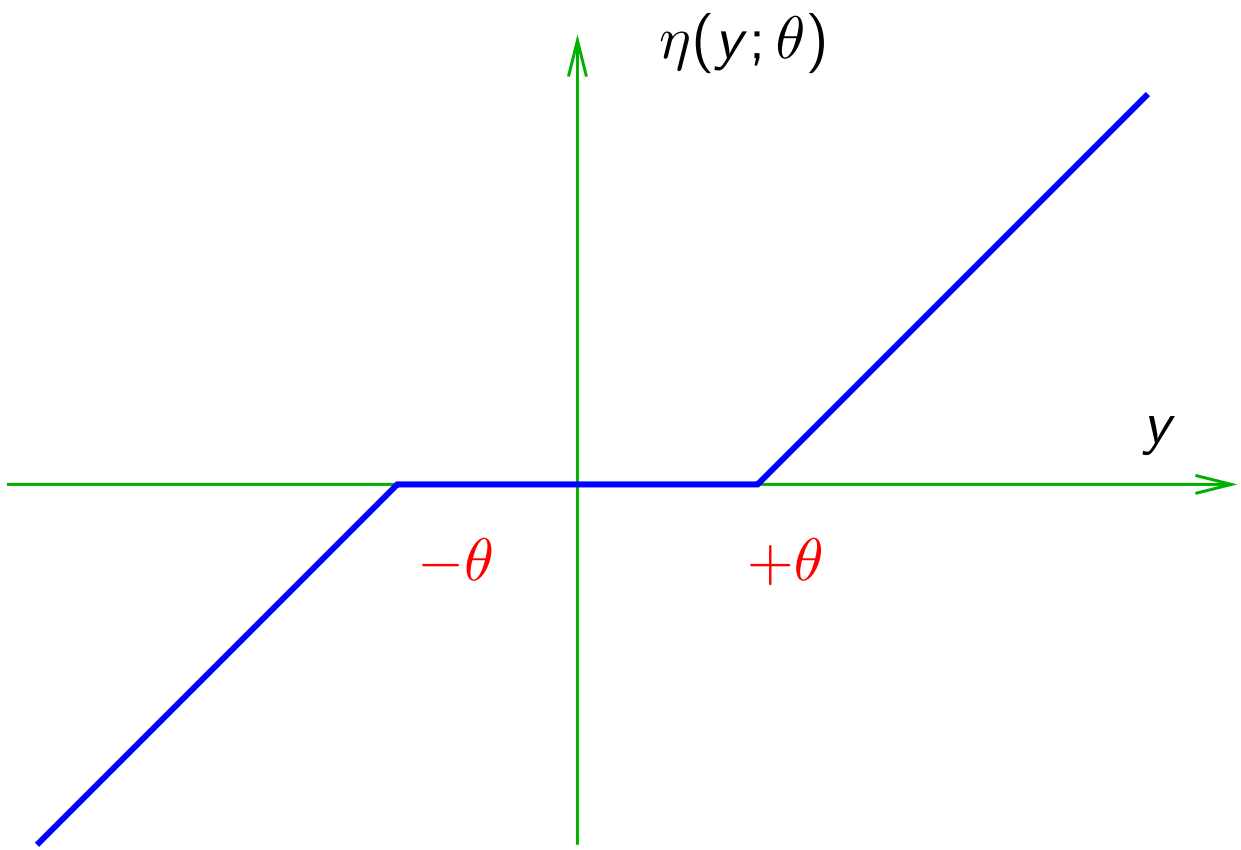
Then,

$$\lim_{N\to\infty} \frac{1}{N}\|\widehat{x}^{\mathrm{LASSO}}(\lambda) - x_0\|^2 = (\tau_\infty^2 - \sigma^2)\delta\,.$$

almost surely as $n \to \infty$.

Conjectured in a more general context with Donoho and Maleki

η



Proof structure

1. Construct a message passing algorithm tho infer x.

2. Prove that the distribution of messages converges weakly to ***

2'. and that their variance is tracked by a recursion.

3. Prove that message passing converges to the LASSO opt.

Approximate message passing algorithm

$$
\begin{aligned}
x^{t+1} &= \eta(x^t + A^T z^t; \theta_t)\,, \\
z^t &= y - A x^t + b_t z^{t-1}\,.
\end{aligned}
$$

with $\quad b_t \equiv \frac{1}{n}\sum_{i=1}^{N} \eta'(x^{t-1} + A^T z^{t-1}; \theta_t)\,.$

Approximate message passing algorithm

$$
\begin{aligned}
x^{t+1} &= \eta(x^t + A^T z^t; \theta_t)\,, \\
z^t &= y - A x^t + b_t z^{t-1}\,.
\end{aligned}
$$

with $\quad b_t \equiv \frac{1}{n} \sum_{i=1}^{N} \eta'(x^{t-1} + A^T z^{t-1}; \theta_t)\,.$

Approximate message passing algorithm

$$
\begin{aligned}
x^{t+1} &= \eta(x^t + A^T z^t; \theta_t)\,, \\
z^t &= y - Ax^t + b_t z^{t-1}\,.
\end{aligned}
$$

Graph is dense.

No local weak limit.

No 'edge' variables: only $O(n)$ messages.

Onsager term

Approximate message passing algorithm

$$
\begin{aligned}
x^{t+1} &= \eta(x^t + A^T z^t; \theta_t)\,, \\
z^t &= y - Ax^t + b_t z^{t-1}\,.
\end{aligned}
$$

Graph is dense.

No local weak limit.

No 'edge' variables: only $O(n)$ messages.

Onsager term

Approximate message passing algorithm

$$
\begin{aligned}
x^{t+1} &= \eta(x^t + A^T z^t; \theta_t)\,, \\
z^t &= y - Ax^t + b_t z^{t-1}\,.
\end{aligned}
$$

Graph is dense.

No local weak limit.

No 'edge' variables: only $O(n)$ messages.

Onsager term

Approximate message passing algorithm

$$
\begin{aligned}
x^{t+1} &= \eta(x^t + A^T z^t; \theta_t)\,, \\
z^t &= y - Ax^t + b_t z^{t-1}\,.
\end{aligned}
$$

Graph is dense.

No local weak limit.

No 'edge' variables: only $O(n)$ messages.

Onsager term

Approximate message passing algorithm

$$
\begin{aligned}
x^{t+1} &= \eta(x^t + A^T z^t; \theta_t)\,, \\
z^t &= y - Ax^t + b_t z^{t-1}\,.
\end{aligned}
$$

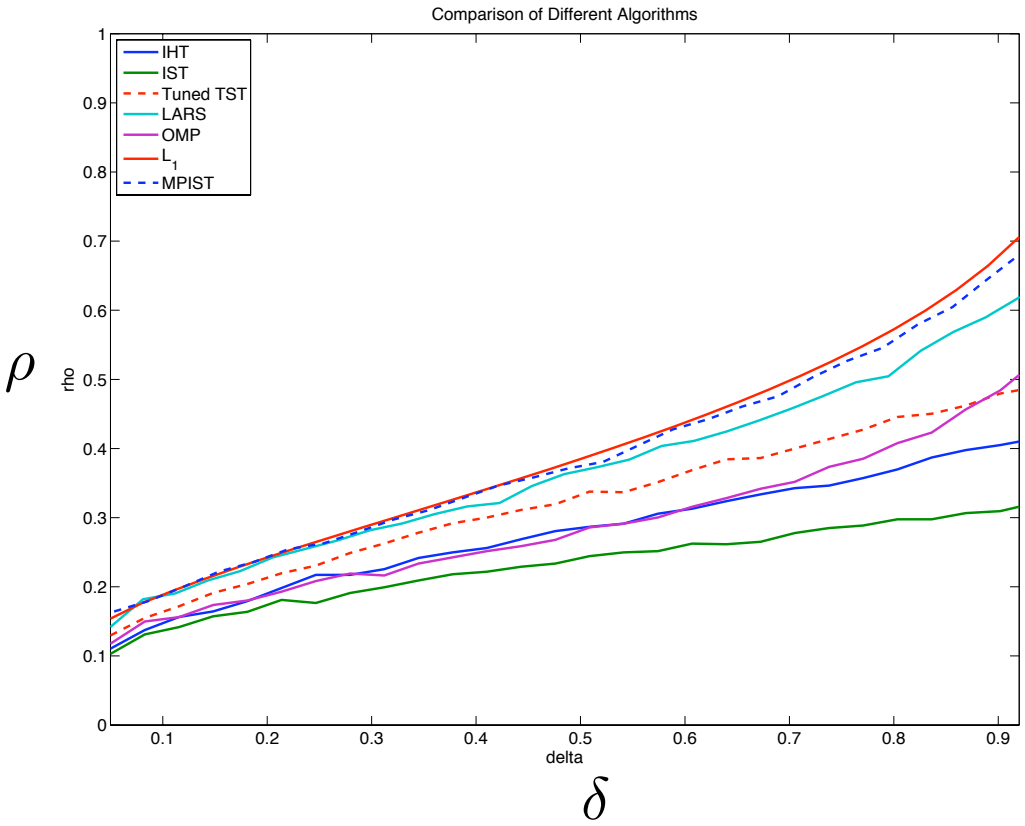
Graph is dense.

No local weak limit.

No 'edge' variables: only $O(n)$ messages.

Onsager term

A large gain in performances



$$
\begin{aligned}
x^{t+1} &= \eta(x^t + A^T z^t; \theta_t)\,, \\
z^t &= y - Ax^t + b_t z^{t-1}\,.
\end{aligned}
$$

Conclusion

I did not talk about

- Gaussian graphical models. (Weiss)
- Free energies $\longrightarrow$ Generalized BP. (Yedidia-Freeman-Weiss)
- Relation with convex relaxations. (Wainwright-Jordan, Bayati et al)
- Message passing to find game-theoretical equilibria. (Kanoria et al. arXiv:1004.2079)

Conclusion: A success looking for theoreticians

- A *super-heuristics*:
  - Subsumes many natural heuristics.
  - Easy to design/optimize.
- (Can be) used almost everywhere.
- No example in which it 'beats' standard methods.

Thanks!

Conclusion: A success looking for theoreticians

- A *super-heuristics*:
  - Subsumes many natural heuristics.
  - Easy to design/optimize.
- (Can be) used almost everywhere.
- No example in which it 'beats' standard methods.

Thanks!