

# Negative Dependence, Stable Polynomials etc in ML

Part 1

**STEFANIE JEGELKA & SUVRIT SRA**

Dept of EECS & CSAIL  
Massachusetts Institute of Technology

Neural information Processing Systems, 2018

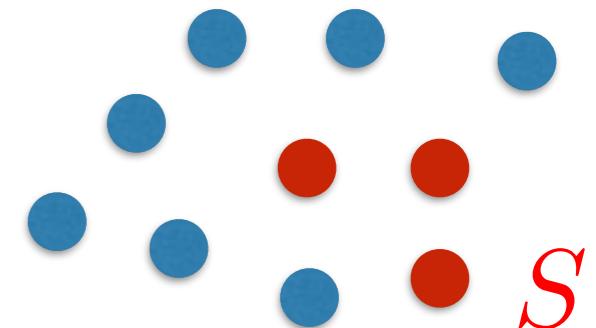


[ml.mit.edu](http://ml.mit.edu)



# Negative Dependence

Discrete probability measure  $\mu(S)$ ,  $S \subseteq V$   
Equivalently:  $n$  binary random variables  $X_i$



## negative dependence



0



1



0



1



1



0

$$\mathbb{P}(\text{smartphone} \in S) \leq \mathbb{P}(\text{smartphone} \in S)\mathbb{P}(\text{smartphone} \in S)$$

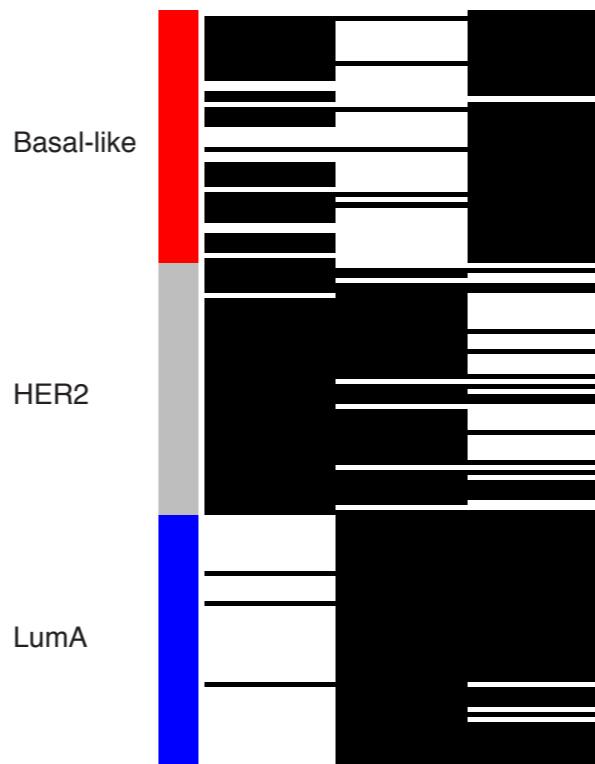
$$\mathbb{P}(\text{smartphone} \in S \mid \text{smartphone} \in S) \leq \mathbb{P}(\text{smartphone} \in S)$$

# Negative dependence - where?

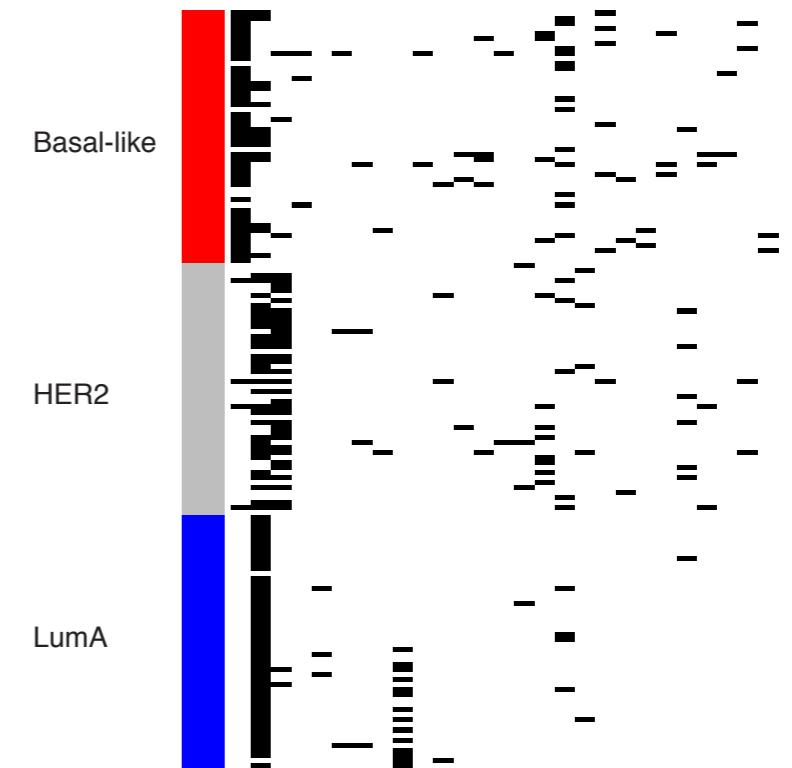


Priors on diversity  
Interpretability

**Negative Dependence**

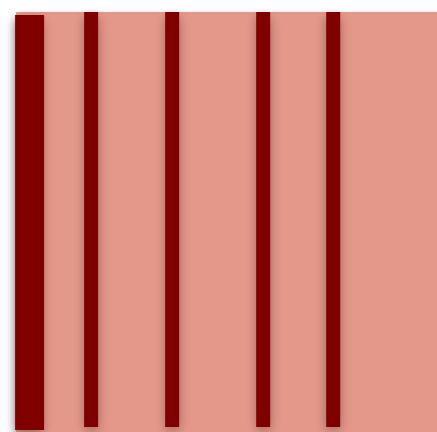


**Indian Buffet Process**

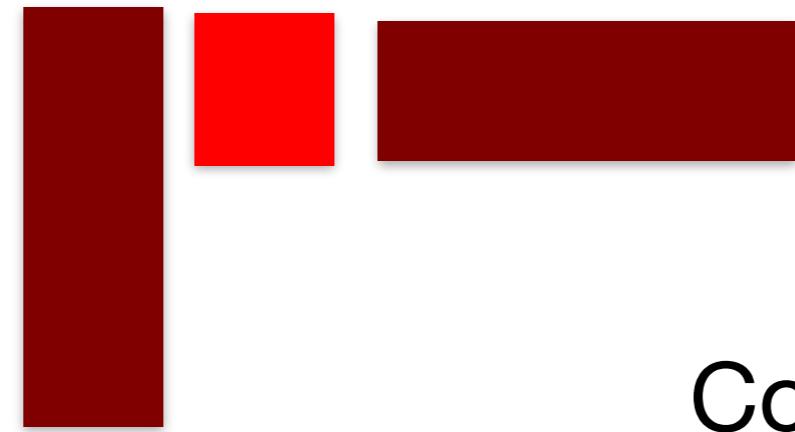


(Xu, Müller, Telesca 2015)

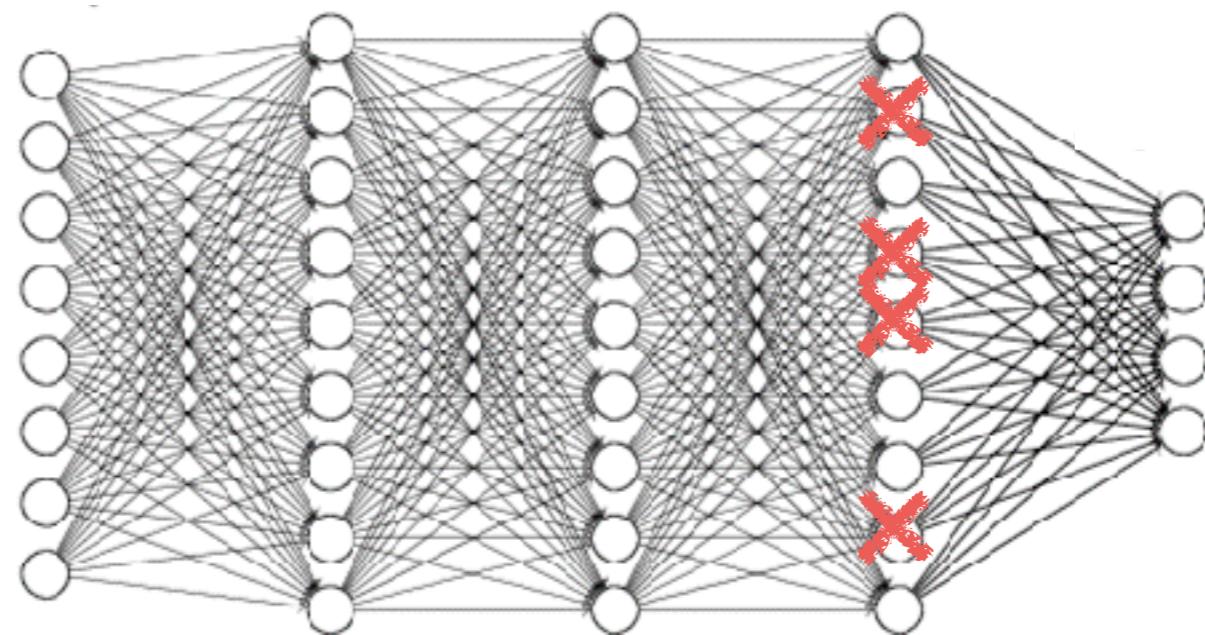
# Negative dependence - where?



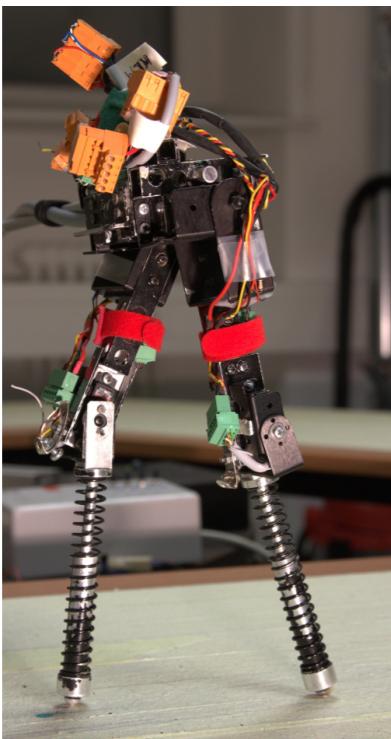
$\approx$



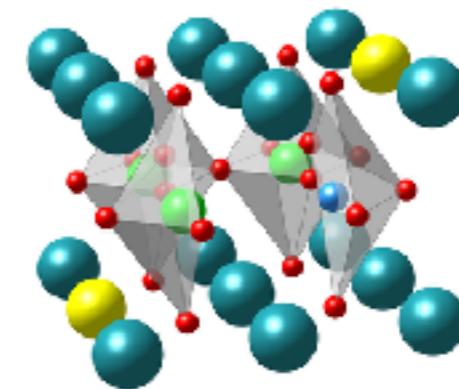
Compressing data  
and models



# Negative dependence - where?



Exploration,  
active learning,  
experimental design,  
Bayesian Optimization



1

Intro &  
Theory

2

Theory &  
Applications

## Introduction

Prominent example: Determinantal Point Processes

## Stronger notions of negative dependence

## Implications: Sampling

—BREAK—

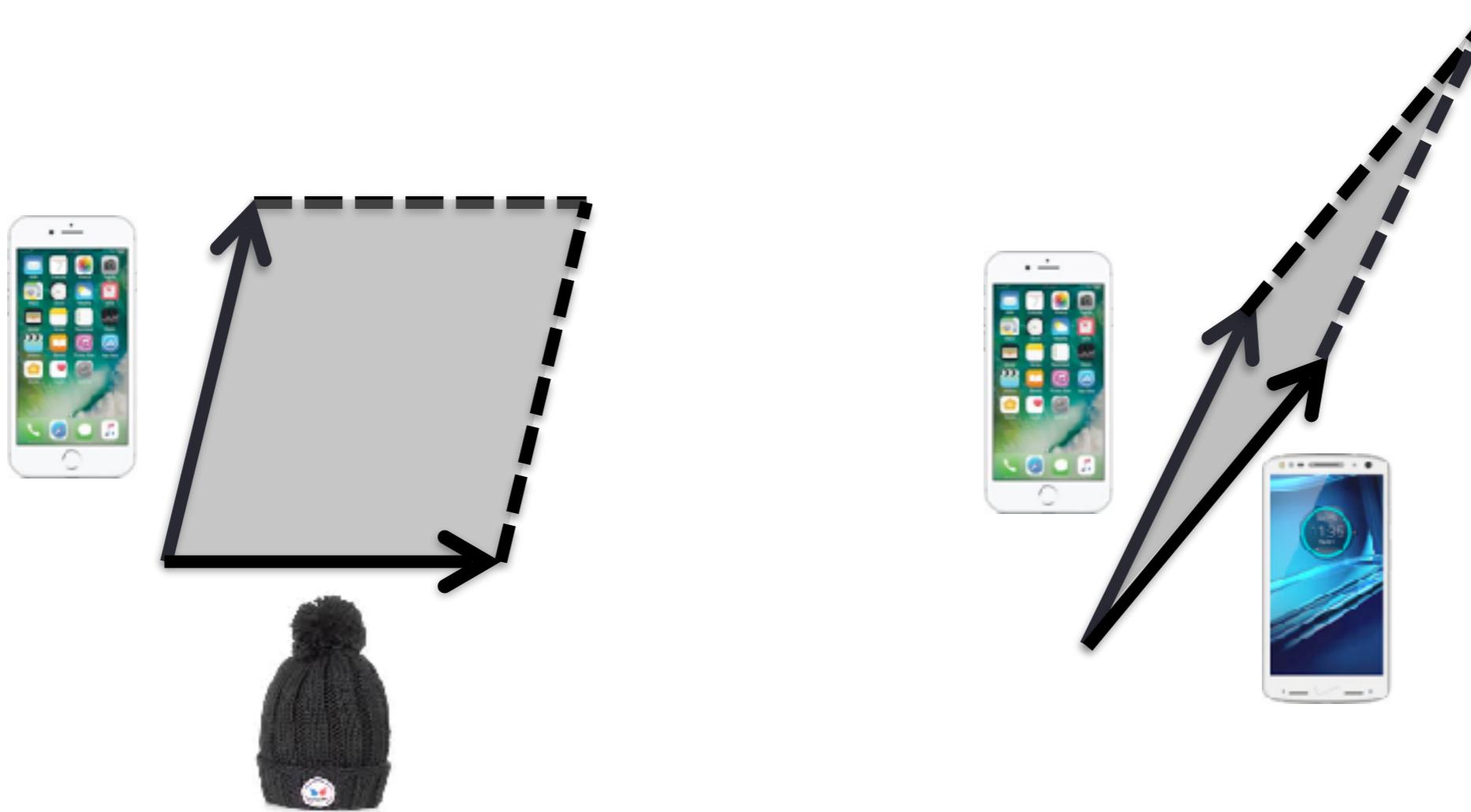
## Approximating partition functions

## Learning a DPP (and some variants)

## Applications

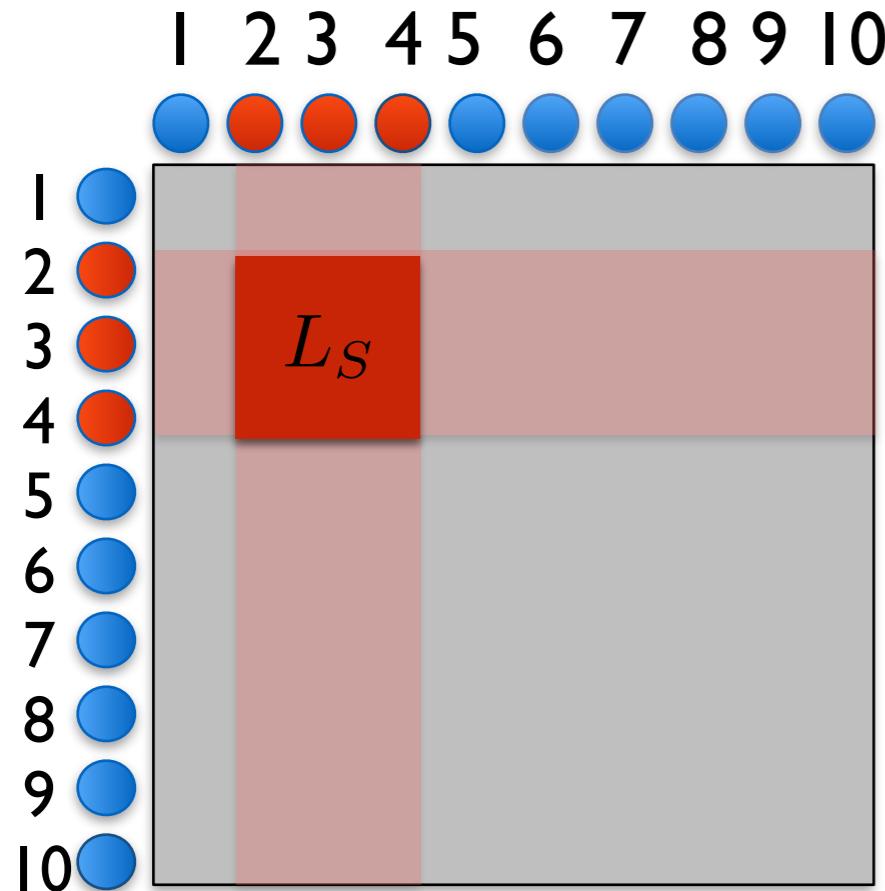
## Perspectives and wrap-up

# Capturing Diversity: Determinantal Point Processes



$$\mu(S) \propto \text{Vol}^2(\{v_i\}_{i \in S})$$

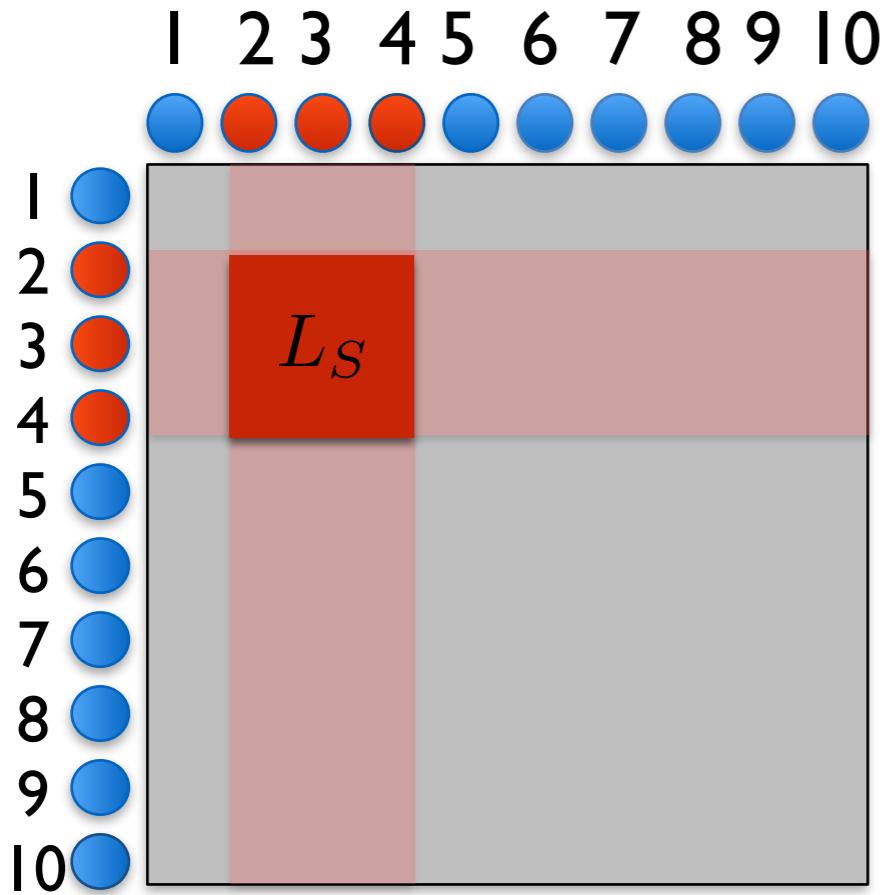
# Determinantal Point Processes



PSD similarity matrix  $L$

$$\mu(S) \propto \det(L_S)$$

# Determinantal Point Processes (DPPs)

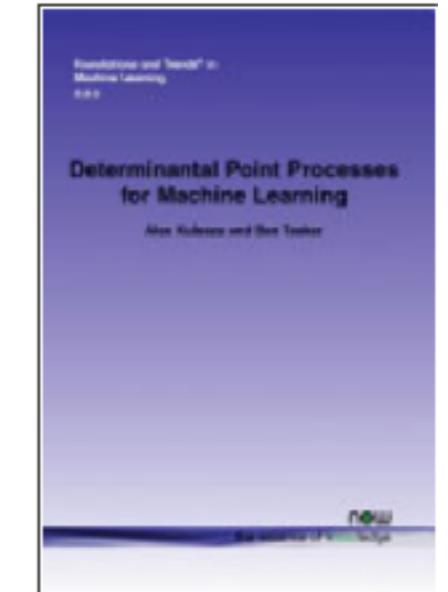


PSD similarity matrix  $L$

$$\mu(S) = \frac{\det(L_S)}{\det(L + I)}$$

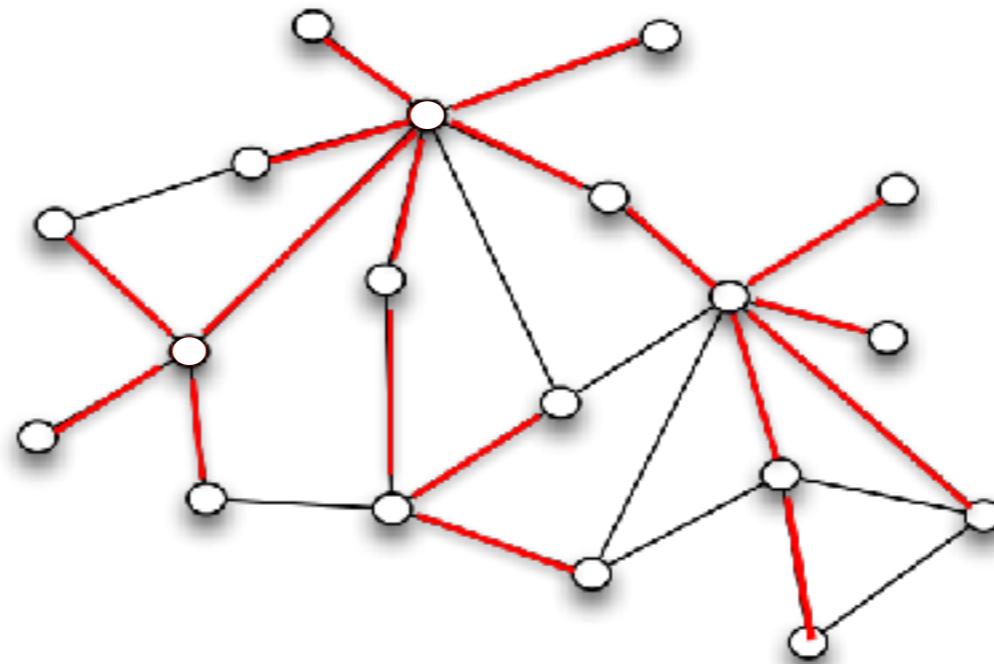
- ◆ Macchi 1975: “fermion processes”
- ◆ Borodin & Olshanski 2000: “Determinantal Point Process”
- ◆ Introduction to ML: Kulesza & Taskar

(Hough, Krishnapur, Peres, Virág 2006; Lyons 2014; Lyons & Peres 2016; Pemantle 2000)



# Combinatorial Examples

random spanning trees

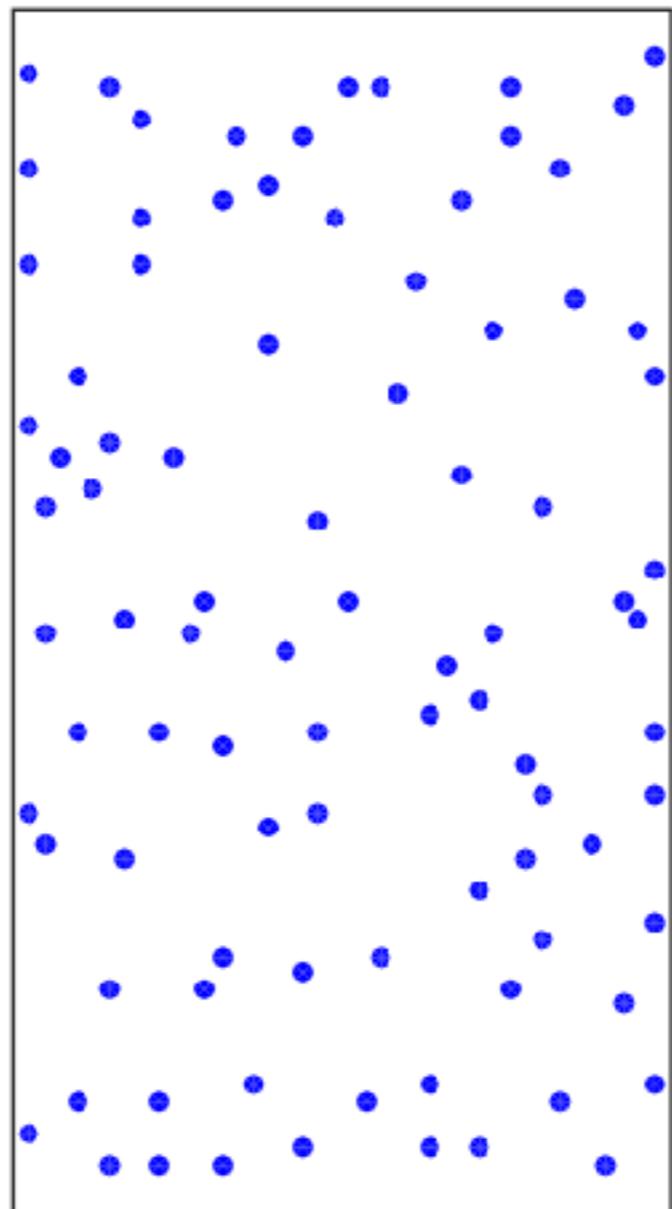


ascents in random sequences

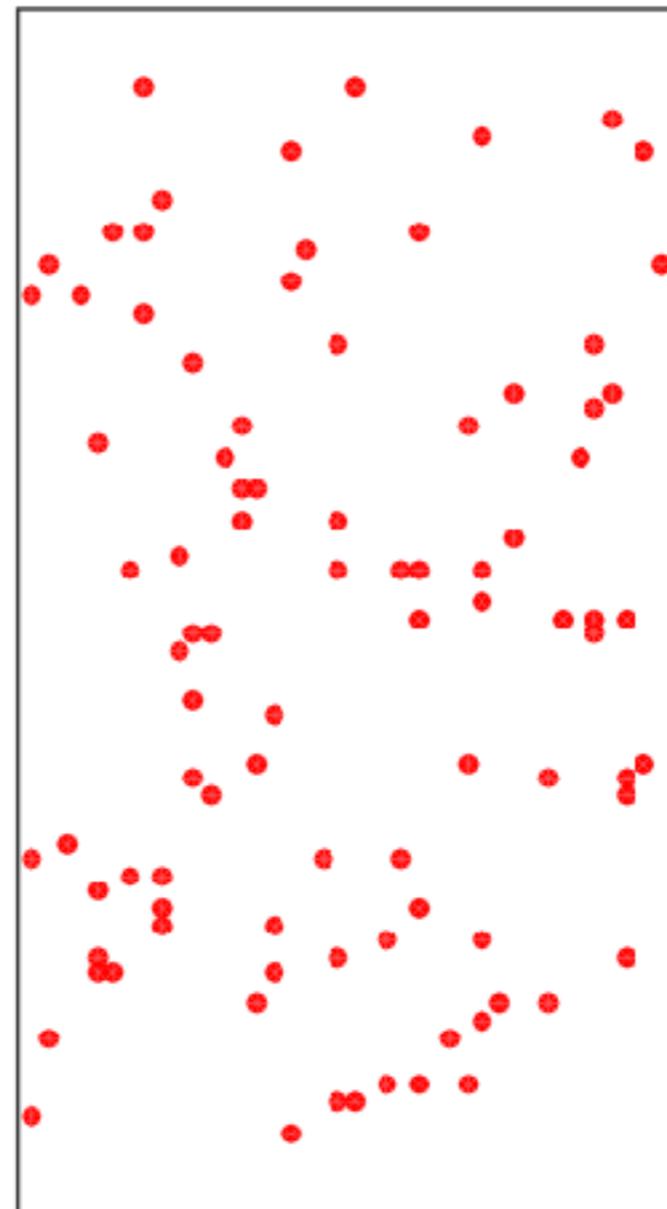
0 **2** **8** 0 **9** 7 4 **5** 2 **4** **9** 5 5 2 **4**

# DPP samples

DPP

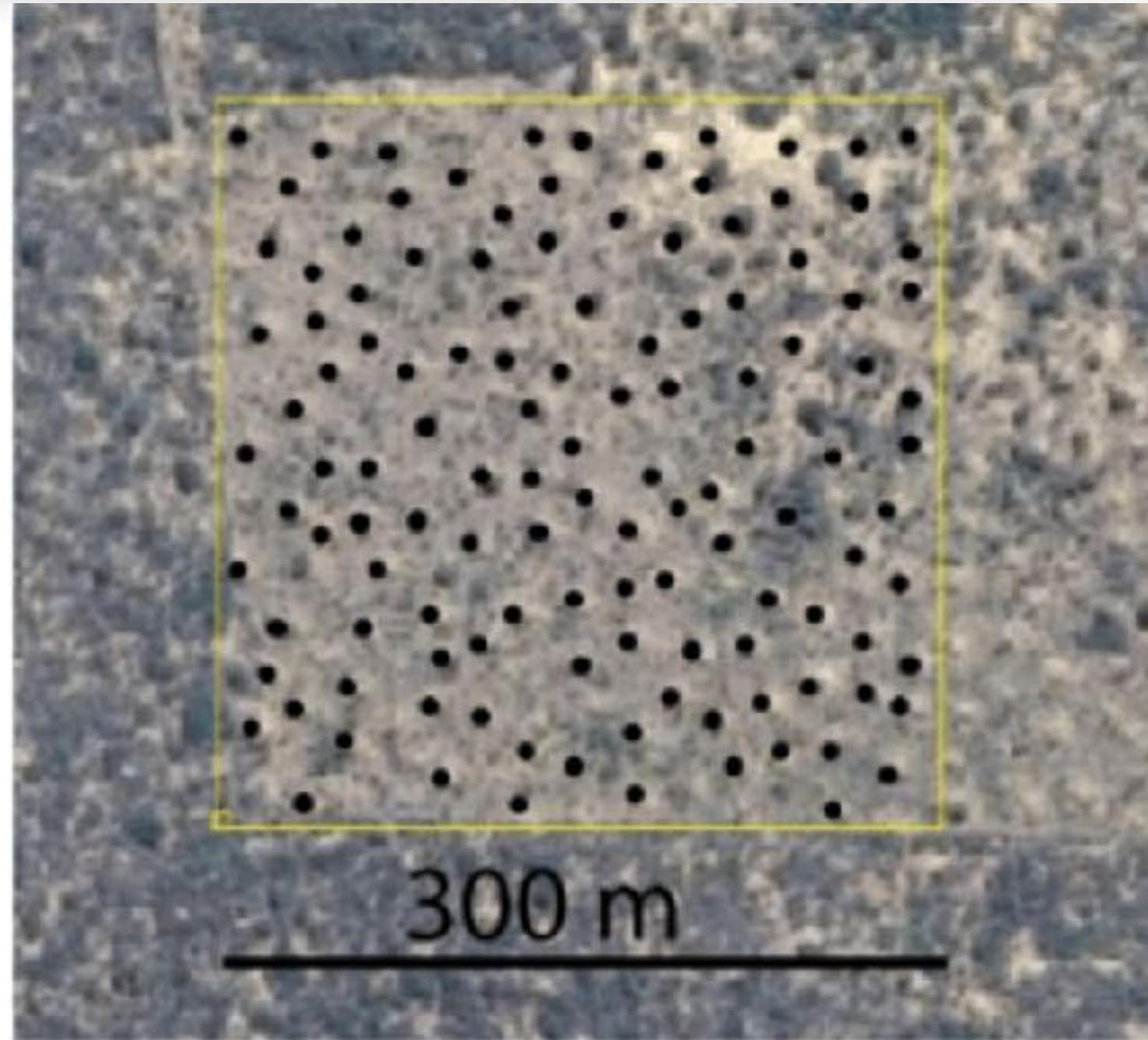


uniform



$$l_{ij} = \exp\left(-\frac{1}{2\sigma^2} \|x_i - x_j\|^2\right)$$

# Repulsion in nature

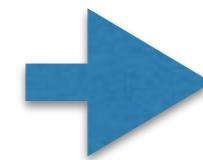


termite mounds in Brazil

(Martin, Funch, Hanson, Yoo, *Current Biology* 2018,  
<http://djalil.chafai.net/blog/2018/11/23/yet-another-determinantal-point-process-in-nature/>, )

1

Intro &  
Theory



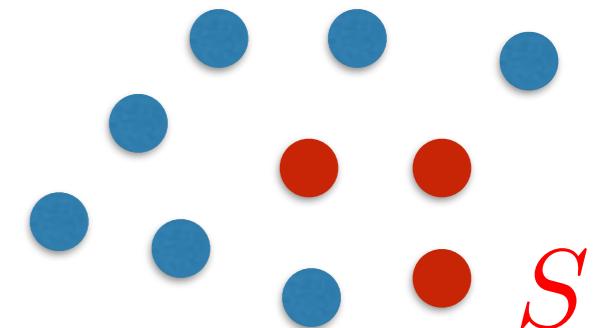
## Stronger notions of negative dependence

Relations to polynomials

## Implications: Sampling

# Negative Dependence

Discrete probability measure  $\mu(S)$ ,  $S \subseteq V$   
Equivalently:  $n$  binary random variables  $X_i$



**negative dependence**



0



1



0



1



1



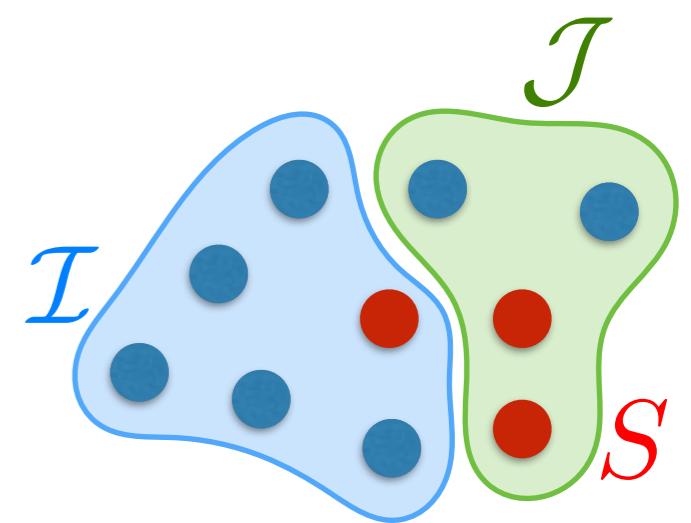
0

$$\mathbb{P}(\text{Smartphone} \in S) \leq \mathbb{P}(\text{Smartphone} \in S)\mathbb{P}(\text{Smartphone} \in S)$$

# Stronger Notions

$$\mathbb{P}(\text{ } \in S) \leq \mathbb{P}(\text{ } \in S)\mathbb{P}(\text{ } \in S)$$

equivalently:  $\mathbb{E}X_i X_j \leq \mathbb{E}X_i \mathbb{E}X_j$

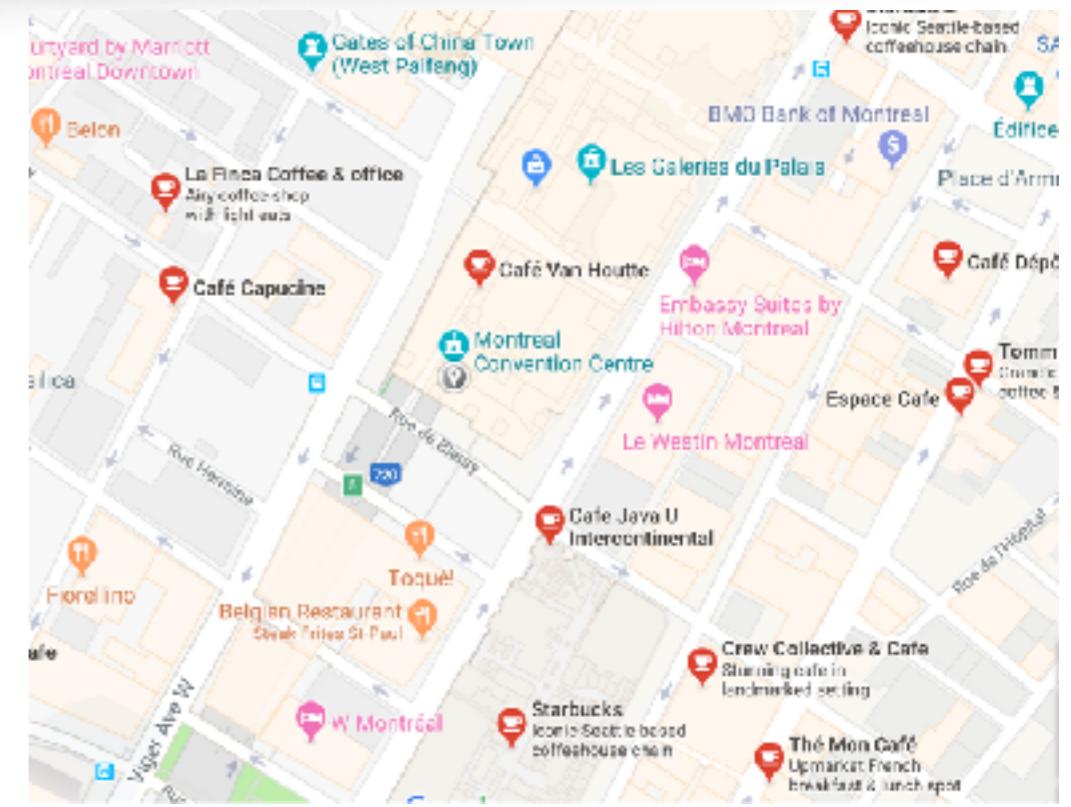


## Negative Association

For all monotone increasing functions  $G(S), H(S)$

$$\mathbb{E}[G(S \cap \mathcal{I})H(S \cap \mathcal{J})] \leq \mathbb{E}G(S \cap \mathcal{I})\mathbb{E}H(S \cap \mathcal{J})$$

# Example



$$B_i = \begin{cases} 1 & \text{if thirsty researcher goes to coffee shop } i \\ 0 & \text{otherwise} \end{cases}$$

# Detour: Positive Dependence

---

## Positive Association:

For all monotone increasing functions  $G(S), H(S)$

$$\mathbb{E}[G(S)H(S)] \geq \mathbb{E}G(S)\mathbb{E}H(S)$$

## FKG Lattice Condition, Multivariate Totally Positive: log-supermodularity

$$\mu(S)\mu(T) \leq \mu(S \cup T)\mu(S \cap T) \quad \forall S, T \subseteq V$$

implies positive association (*Fortuin, Kasteleyn, Ginibre 1971*)

**Analog does not hold for negative dependence!**

negative association implies log-submodularity, but not conversely

# Towards a theory of negative dependence

Journal of Mathematical Physics 41, 1371 (2000); <https://doi.org/10.1063/1.533200>

Robin Pemantle

The “right notion” should imply:

- ◆ Negative Association
- ◆ Stochastic Covering
- ◆ Log-concave rank sequences  $a_k = \mathbb{P}(|S| = k)$
- ◆ Closed under “natural” operations
  - ◆ conditioning
  - ◆ marginalization
  - ◆ products ...

# Generating Polynomial

$$q_\mu(z) = \sum_{S \subseteq V} \mu(S) \prod_{i \in S} z_i, \quad z \in \mathbb{C}^n$$



Example:  
2 items

$$q_\mu(z) = \mu_\emptyset + \mu_1 z_1 + \mu_2 z_2 + \mu_{1,2} z_1 z_2$$

**Operations on polynomial = operations on distribution**

**Obtain coefficient**  $\mu_1 = \mu(\{1\})$

1. differentiate wrt  $z_1$ :

$$\frac{\partial}{\partial z_1} q_\mu(z) = \mu_1 + \mu_{1,2} z_2$$

2. set  $z = 0$

$$\left. \frac{\partial}{\partial z_1} q_\mu(z) \right|_{z=0}$$

# Generating Polynomial

$$q_\mu(z) = \sum_{S \subseteq V} \mu(S) \prod_{i \in S} z_i, \quad z \in \mathbb{C}^n$$



**Example:**  
2 items

$$q_\mu(z) = \mu_\emptyset + \mu_1 z_1 + \mu_2 z_2 + \mu_{1,2} z_1 z_2$$

**Marginalization:**  $\pi(\{1\}) = \mathbb{P}(1 \in S)$

Set  $z_2 = 1$

$$q_\mu(z_1, 1) = [\mu_\emptyset + \mu_2] + [\mu_1 + \mu_{1,2}] z_1 = q_\pi(z_1)$$

$\mathbb{P}(1 \notin S) \quad \mathbb{P}(1 \in S)$

**Properties of polynomial**  $\Leftrightarrow$  **Properties of distribution**

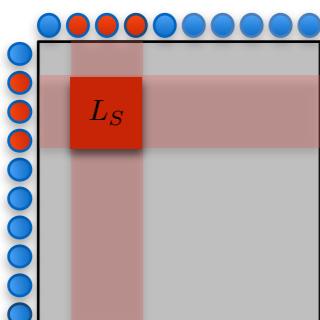
# Strongly Rayleigh Measures

$$q_\mu(z) = \sum_{S \subseteq V} \mu(S) \prod_{i \in S} z_i, \quad z \in \mathbb{C}^n$$

$\mu(S)$  is **Strongly Rayleigh (SR)** if  $q_\mu(z)$  is *real stable*:

$$\operatorname{Im}(z_i) > 0 \quad \forall i \Rightarrow q_\mu(z) \neq 0$$

SR implies almost all conditions laid out by Pemantle!  
(*Borcea, Bränden, Liggett 2009*)



**Determinantal Point Process:**  $q_\mu(z)$  essentially multivariate variant of the characteristic polynomial  $\det(z_i I - L)$

**Operations on polynomial  $\Leftrightarrow$  Operations on distribution**

# Real Stable Polynomials

---

$$\operatorname{Im}(z_i) > 0 \quad \forall i \Rightarrow q_\mu(z) \neq 0$$

- ♦ Deep mathematical connections
- ♦ Univariate case goes back to Newton (at least)
- ♦ Powerful class of closure properties

Generating polynomial  $q(z)$  is SR is equivalent to

$$\frac{\partial q(x)}{\partial z_i} \frac{\partial q(x)}{\partial z_j} \geq q(x) \frac{\partial^2 q(x)}{\partial z_i \partial z_j}, \quad x \in \mathbb{R}^n \quad \forall i, j$$

$$\Rightarrow \mu(S)\mu(T) \geq \mu(S \cup T)\mu(S \cap T) \quad \forall S, T \subseteq V$$

# Nice properties of SR

---

- ◆ **Closed** under marginalization,  
conditioning on  $|S| = k, X_i = 1, X_i = 0$   
...
- ◆ Implies many other types of **negative dependence**,  
e.g. negative association and  $\mathbb{E} \prod_i X_i \leq \prod_i \mathbb{E} X_i$
- ◆ **Concentration of measure**:  $X_i$  behave like  
independent random variables  
sum  $\sum_{i=1}^n X_i$ , Lipschitz functions  $F(S)$ , matrices  
(Panconesi, Srinivasan 1997, Dubhashi, Ranjan 1998, Farcomeni 2008,  
Pemantle, Peres 2011, Kyng, Song 2018, Garbe, Vondrak 2018,...)

# Algorithmic Implications

---

- ♦ Sampling

(Féder, Mihail 1992, Jerrum, Son 2002, Jerrum, Son, Tetali, Vigoda 2004, Anari, Gharan, Rezaei 2016, Li, Jegelka, Sra 2016)

- ♦ Approximate partition functions, permanents, volumes, counting

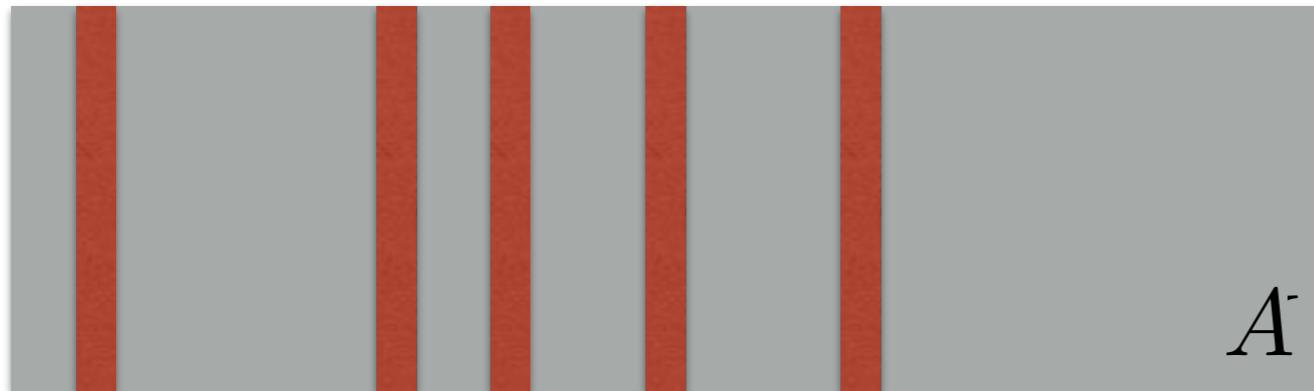
(Gurvits 2006, Nikolov-Singh 2016, Straszak-Vishnoi 2016, Anari, Gharan, Saberi, Singh 2016, Anari, Gharan 2017...)

- ♦ Approximation Algorithms

(Panconesi, Srinivasan 1997, Hayes 2003, Considine, Byers, Mitzenmacher 2004, Asadpour, Goemans, Madry, Oveis-Gharan, Saberi 2010, Gharan, Saberi, Singh 2011, Spielman, Srivastava 2011, ...)

# Other SR measures

## Dual Volume Sampling



$$P(S) \propto \det(A_S A_S^\top)$$

**NOT a DPP ... but Strongly Rayleigh!**

proof: closeness properties of polynomials

(Avron, Boutsidis 2013, Li, Jegelka, Sra 2017, Derezhinski, Warmuth 2017)

25

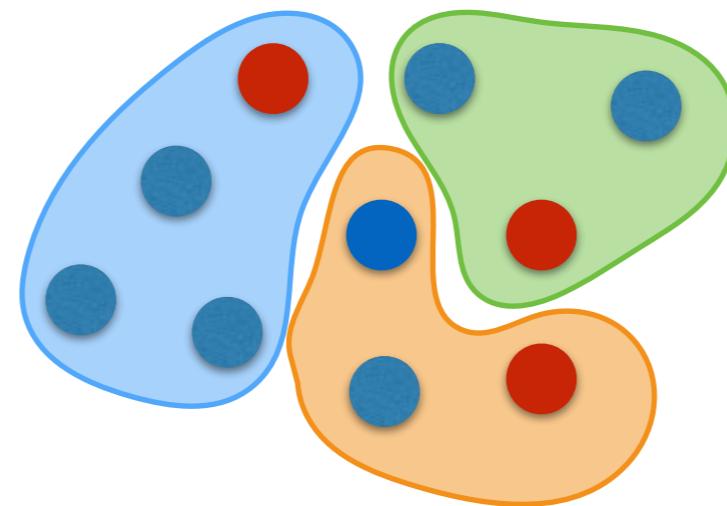
# .... and limits

$\mu(S)$  is SR. What about  $\mu(S)^p$ ? In general not SR!

(Kulesza, Taskar 2012, Zou, Adams 2012, Gillenwater 2014, Anari, Gharan 2017, Mariet, Sra, Jegelka 2018)

Conditioning on combinatorial constraints: at most one item per group. In general not SR!

(Celis, Deshpande, Kathuria, Straszak, Vishnoi 2017, Celis, Keswani, Straszak, Deshpande, Kathuria, Vishnoi 2018)



## 1

Intro &  
Theory

### Introduction

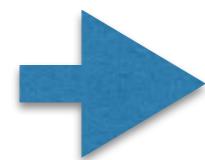
Examples

Determinantal Point Processes

### Stronger notions of negative dependence

Relations to polynomials

Strongly Rayleigh measures

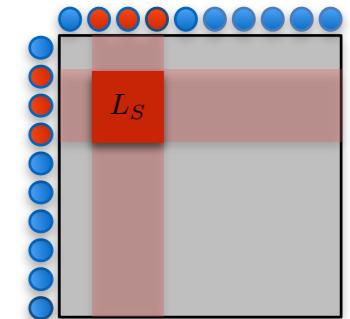


### Implications: Sampling

# Sampling for Determinantal Point Processes

**EVD/SVD** (*Hough, Krishnapur, Peres, Virág 2006*)  $O(n^3)$

- ◆ sample subspace via eigenvalues
- ◆ sequentially sample items



Faster via  $A$ : “dual” sampling (*Kulesza, Taskar 2010*)

$$L = A^\top A$$

$\downarrow$   
 $n \times d$

## Acceleration: approximate $L$ or $A$

volume-preserving sketching (*Magen, Zouzias 2008; Deshpande, Rademacher 2010, Gillenwater, Kulesza, Taskar 2012*)

Nyström (*Affandi, Kulesza, Fox, Taskar 2013*)

MCMC (*Bardenet, Hardy 2016*)

Coresets (*Li, Jegelka, Sra 2016*)

R-DPP (*Derezinski 2018*)

**Other approaches** (*Derezinski, Warmuth 2018, Derezinski, Warmuth, Hsu 2018, Mariet, Sra 2016, ...*)

# Sampling for general SR measures

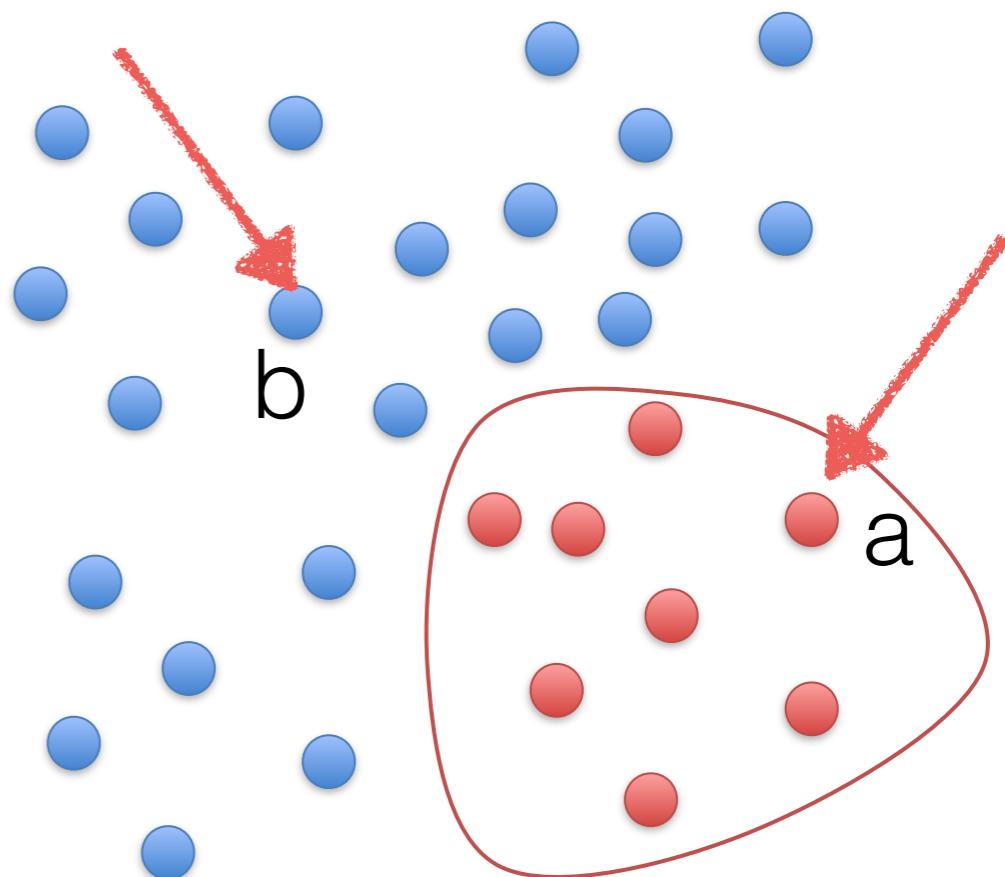
## Markov Chain Monte Carlo (MCMC)

in iteration  $t$ :

sample  $a \in S_{t-1}$ ,  $b \notin S_{t-1}$  uniformly at random

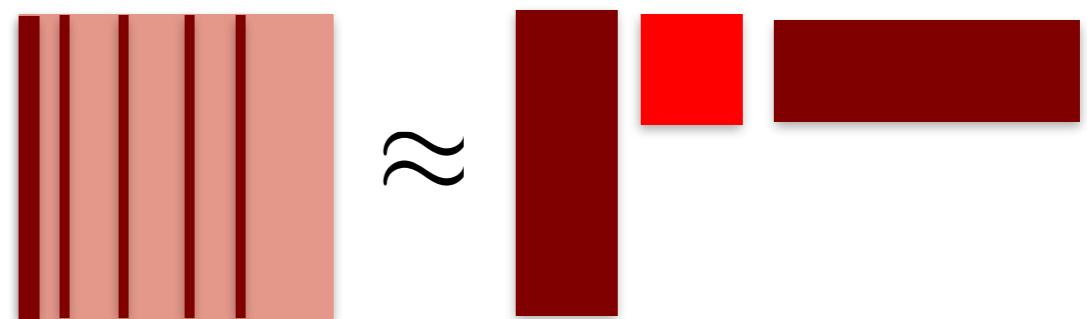
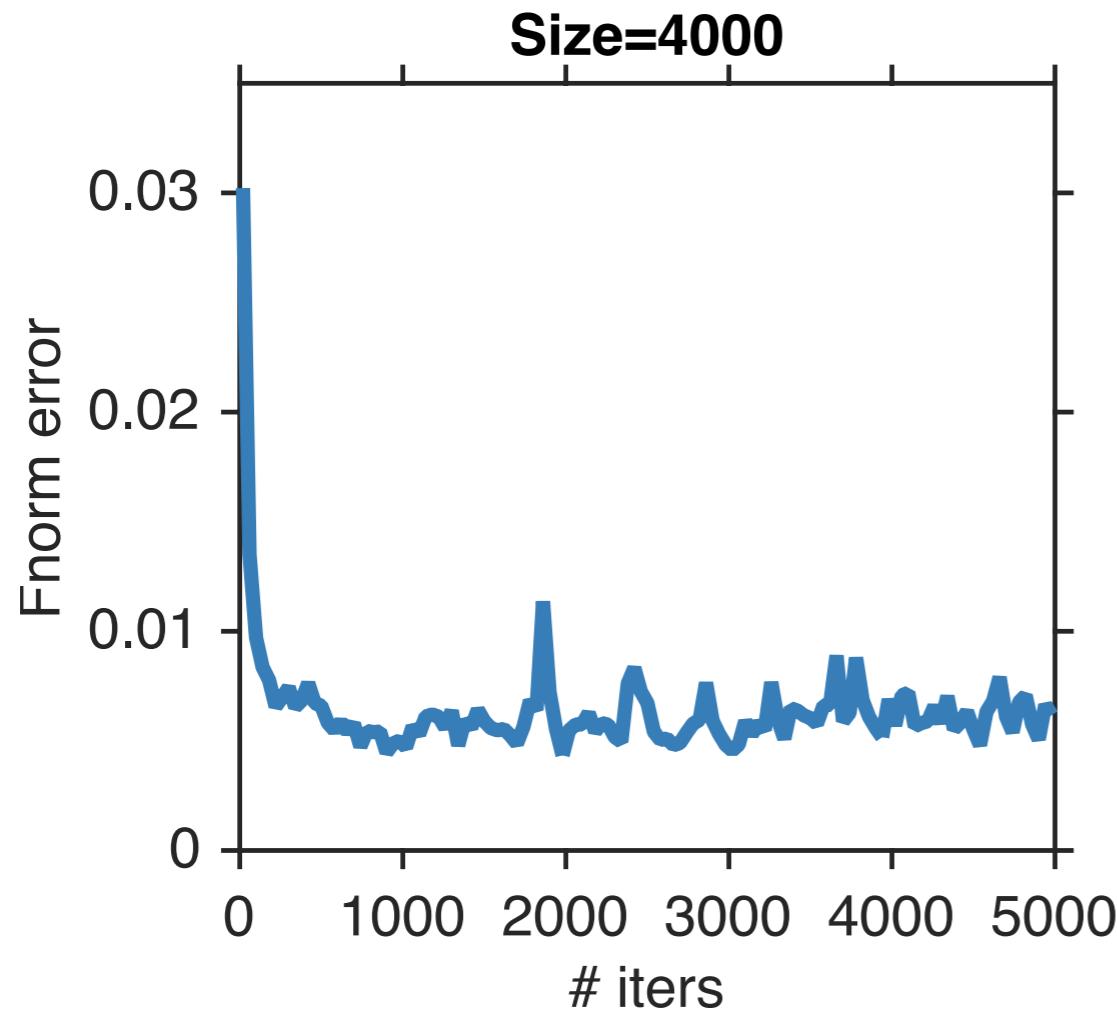
randomly do one of

- ▶ swap: remove  $b$ , add  $a$
- ▶ add  $b$
- ▶ remove  $a$
- ▶ retain  $S_t = S_{t-1}$



if conditioning on  $|S| = k$ : only swaps

# Sampling empirically



# How many iterations?

Mixing time: #iterations until  $\|\mu^t - \mu\|_{\text{TV}} \leq \epsilon$

balanced matroids:

(Féder, Mihail 1992, Jerrum, Son 2002)

Jerrum, Son, Tetali, Vigoda, 2004)

$$O\left(nk \log\left(\frac{1}{\epsilon\mu(S_0)}\right)\right)$$

$|S| = k$

fixed-cardinality SR:

(Anari, Oveis-Gharan, Rezaei 2016)

arbitrary SR:

(Li, Jegelka, Sra 2016)

$$O\left(n^2 \log\left(\frac{1}{\epsilon\mu(S_0)}\right)\right)$$

# Sampling: from fixed-cardinality to general SR



$$V_{\text{new}} = V \cup V'$$

- sample  $n$  out of  $2n$ , but use only  $T \cap V$
- extend measure to “shadow set”
- **Key:** measure on  $2n$  is Strongly Rayleigh by closedness properties of real stable polynomials (“symmetric homogenization”)

(Li, Jegelka, Sra 2016) 32

# Sampling

---

- ◆ generic method: MCMC;  
different algorithms for special cases (DPP, dual volume)
- ◆ Accelerating each iteration of DPP-MCMC: lazy computations with Gauss quadrature (*Li,Sra,Jegelka 2016*)
- ◆ Continuous DPP sampling  
(*Affandi,Fox,Taskar 2013, Gharan,Rezaei 2018, ...*)
- ◆ SR sufficient for fast mixing but not necessary: e.g., complete log concavity (*Anari,Liu,Gharan,Vinzant 2018*)
- ◆ Sampling log-submodular distributions (negative lattice condition) (*Rebeschini,Karbasi 2015, Gotovos,Hassani, Krause 2018, Gotovos,Hassani,Krause,Jegelka 2018*)

# Outline

1

Intro &  
Theory

2

Theory &  
Applications

## Introduction

Examples, Determinantal Point Processes

## Stronger notions of negative dependence

Strongly Rayleigh measures and real stable polynomials

## Implications: Sampling

— BREAK —

## Approximating partition functions

## Learning a DPP (and some variants)

## Applications

## Perspectives and wrap-up