Shared Information for a Markov Chain on a Tree

Sagnik Bhattacharya, Prakash Narayan
ECE-ISR@UMD

ISIT 2022

July 1, 2022

#1: Capturing dependence among multiple rvs – *Shared Information*

How to capture dependence among multiple rvs?

Shared information

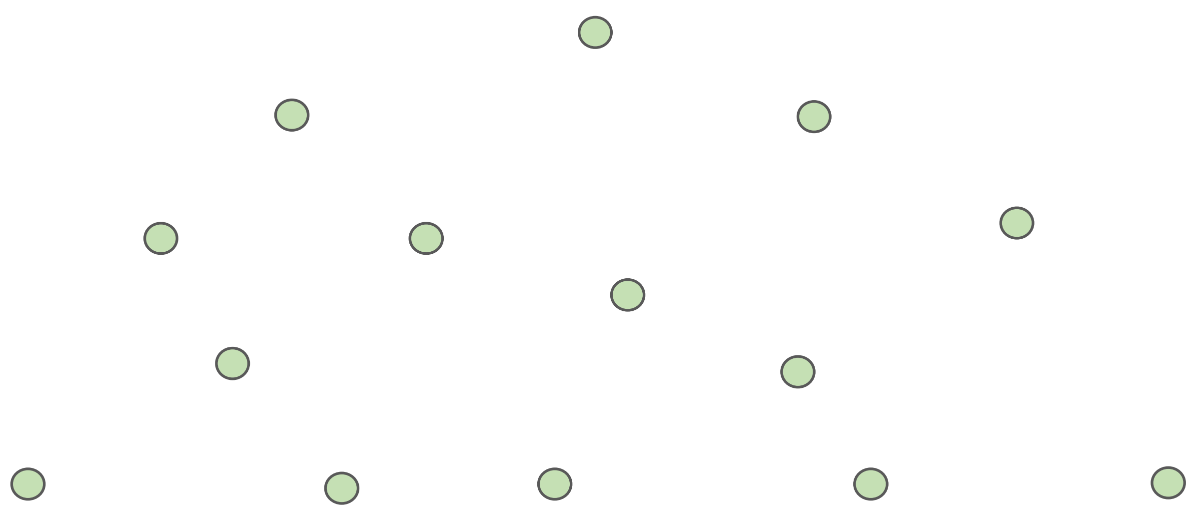
$$\mathrm{SI}(X_{\mathcal{M}}) = \min_{2 \leq k \leq m} \min_{\pi=(\pi_u, u=1,\cdots,k)} \frac{1}{k-1} D(P_{X_{\mathcal{M}}} \parallel \prod_{u=1}^{k} P_{X_{\pi_u}})$$

Shared information

$$\begin{aligned}\mathrm{SI}(X_{\mathcal{M}}) &= \min_{2\le k\le m} \min_{\pi=(\pi_u, u=1,\cdots,k)} \frac{1}{k-1} D(P_{X_{\mathcal{M}}} \parallel \prod_{u=1}^{k} P_{X_{\pi_u}}) \\ &= \min_{2\le k\le m} \min_{\pi=(\pi_u, u=1,\cdots,k)} \frac{1}{k-1} \left[\sum_{u=1}^{k} \mathrm{H}(X_{\pi_u}) - \mathrm{H}(X_{\mathcal{M}})\right]\end{aligned}$$

Shared information

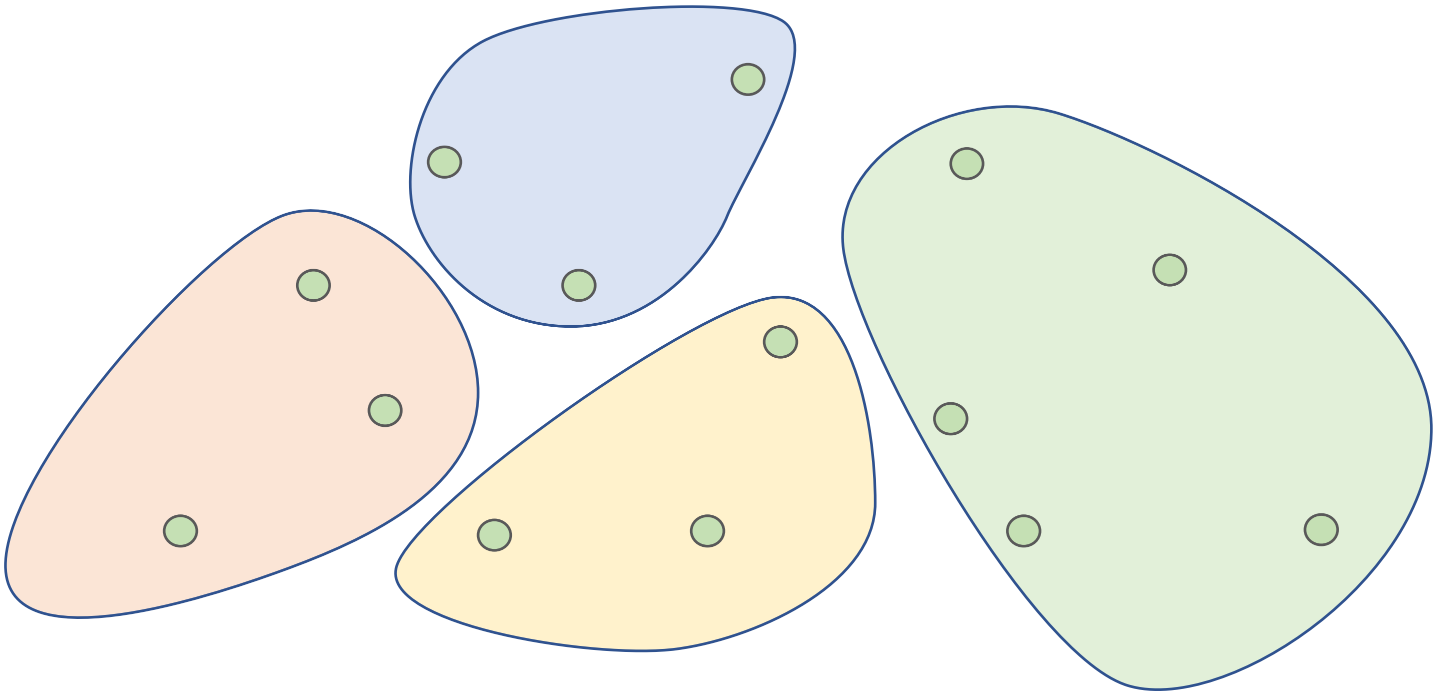
$$
\begin{aligned}
\mathrm{SI}(X_{\mathcal{M}}) &= \min_{2 \le k \le m} \; \min_{\pi=(\pi_u, u=1,\cdots,k)} \; \frac{1}{k-1} D\Big(P_{X_{\mathcal{M}}} \,\Big\|\, \prod_{u=1}^{k} P_{X_{\pi_u}}\Big) \\
&= \min_{2 \le k \le m} \; \min_{\pi=(\pi_u, u=1,\cdots,k)} \; \frac{1}{k-1} \left[\sum_{u=1}^{k} \mathrm{H}(X_{\pi_u}) - \mathrm{H}(X_{\mathcal{M}}) \right]
\end{aligned}
$$



Shared information

$$\mathrm{SI}(X_{\mathcal{M}}) = \min_{2\le k\le m} \min_{\pi=(\pi_u, u=1,\cdots,k)} \frac{1}{k-1} D(P_{X_{\mathcal{M}}} \parallel \prod_{u=1}^{k} P_{X_{\pi_u}})$$

$$= \min_{2\le k\le m} \min_{\pi=(\pi_u, u=1,\cdots,k)} \frac{1}{k-1} \left[\sum_{u=1}^{k} \mathrm{H}(X_{\pi_u}) - \mathrm{H}(X_{\mathcal{M}})\right]$$



Some special cases

- Two rvs

$$\mathrm{SI}(X_1, X_2) = \text{mutual information } \mathrm{I}(X_1 \wedge X_2)$$

Some special cases

- Two rvs

$$\mathrm{SI}(X_1, X_2) = \text{mutual information } \mathrm{I}(X_1 \wedge X_2)$$

- Three rvs
  - Minimum of

$$\mathrm{I}(X_1 \wedge X_2, X_3) \qquad I(X_2 \wedge X_1, X_3) \qquad \mathrm{I}(X_3 \wedge X_1, X_2)$$

$$\frac{1}{2}\left[\mathrm{H}(X_1) + \mathrm{H}(X_2) + \mathrm{H}(X_3) - \mathrm{H}(X_1, X_2, X_3)\right]$$

Shared information - operational significance

- Secret key capacity of a multiterminal source model
  - [Csiszár-Narayan, 2004]

Shared information - operational significance

- Secret key capacity of a multiterminal source model
    - [Csiszár-Narayan, 2004]
    - [Chan, 2008]
    - [Chan-Zheng, 2016]
    - [Chan, 2011]

Shared information - operational significance

- Secret key capacity of a multiterminal source model
  - [Csiszar-Narayan, 2004]
  - [Chan, 2008]
  - [Chan-Zheng, 2016]
  - [Chan, 2011]
- Hypothesis Testing
  - [Tyagi-Watanabe, 2016]

Shared information - operational significance

- Secret key capacity of a multiterminal source model
  - [Csiszar-Narayan, 2004]
  - [Chan, 2008]
  - [Chan-Zheng, 2016]
  - [Chan, 2011]
- Hypothesis Testing
  - [Tyagi-Watanabe, 2016]
- Extensively studied in [Chan-Bashabsheh-Ebrahimi-Kaced-Liu, 2015]

Shared information - operational significance

- Secret key capacity of a multiterminal source model
  - [Csiszar-Narayan, 2004]
  - [Chan, 2008]
  - [Chan-Zheng, 2016]
  - [Chan, 2011]
- Hypothesis Testing
  - [Tyagi-Watanabe, 2016]
- Extensively studied in [Chan-Bashabsheh-Ebrahimi-Kaced-Liu, 2015]
- Clustering
  - [Chan-Bashabsheh-Zhou-Kaced-Liu, 2016]

Shared information - computation

$$\mathrm{SI}(X_{\mathcal{M}}) = \min_{2 \le k \le m} \min_{\pi=(\pi_u, u=1,\cdots,k)} \frac{1}{k-1} D(P_{X_{\mathcal{M}}} \parallel \prod_{u=1}^{k} P_{X_{\pi_u}})$$

- When underlying pmf is known, there is an efficient algorithm to compute SI
  - Submodular optimization [CBEKL15]

Shared information - computation

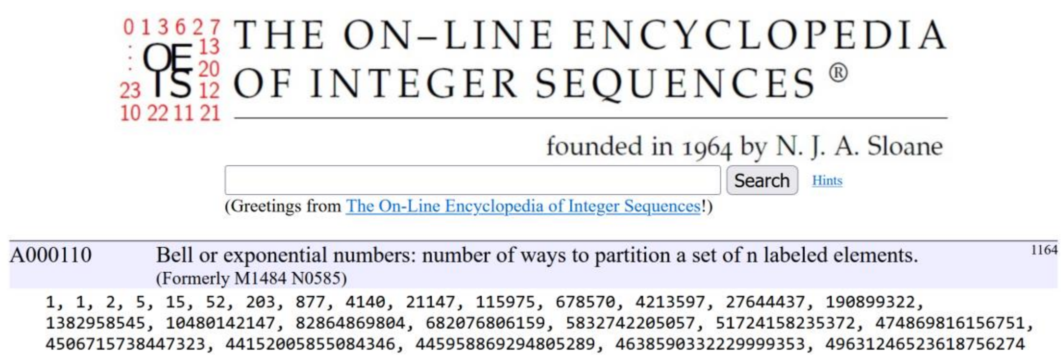
$$\mathrm{SI}(X_{\mathcal{M}}) = \min_{2\le k\le m} \min_{\pi=(\pi_u, u=1,\cdots,k)} \frac{1}{k-1} D(P_{X_{\mathcal{M}}} \parallel \prod_{u=1}^{k} P_{X_{\pi_u}})$$

- When underlying pmf is known, there is an efficient algorithm to compute SI
  - Submodular optimization [CBEKL15]
- When pmf is unknown, estimation involves prohibitively massive search space.

Shared information - computation

$$\mathrm{SI}(X_{\mathcal{M}}) = \min_{2\le k\le m} \min_{\pi=(\pi_u, u=1,\cdots,k)} \frac{1}{k-1} D(P_{X_{\mathcal{M}}} \parallel \prod_{u=1}^{k} P_{X_{\pi_u}})$$

- When underlying pmf is known, there is an efficient algorithm to compute SI
  - Submodular optimization [CBEKL15]
- When pmf is unknown, estimation involves prohibitively massive search space.

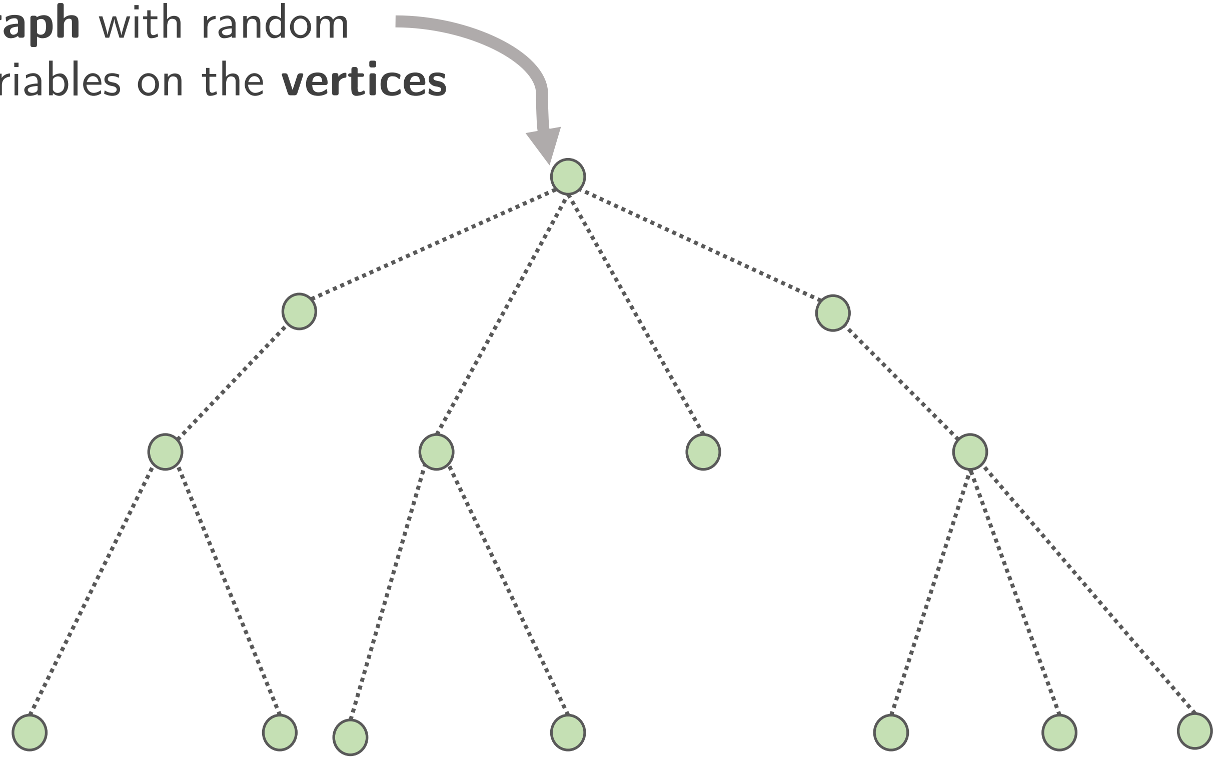


Want: simpler forms in special cases

#2: Markov Chain on a Tree (MCT)

Markov chain on a tree (MCT)

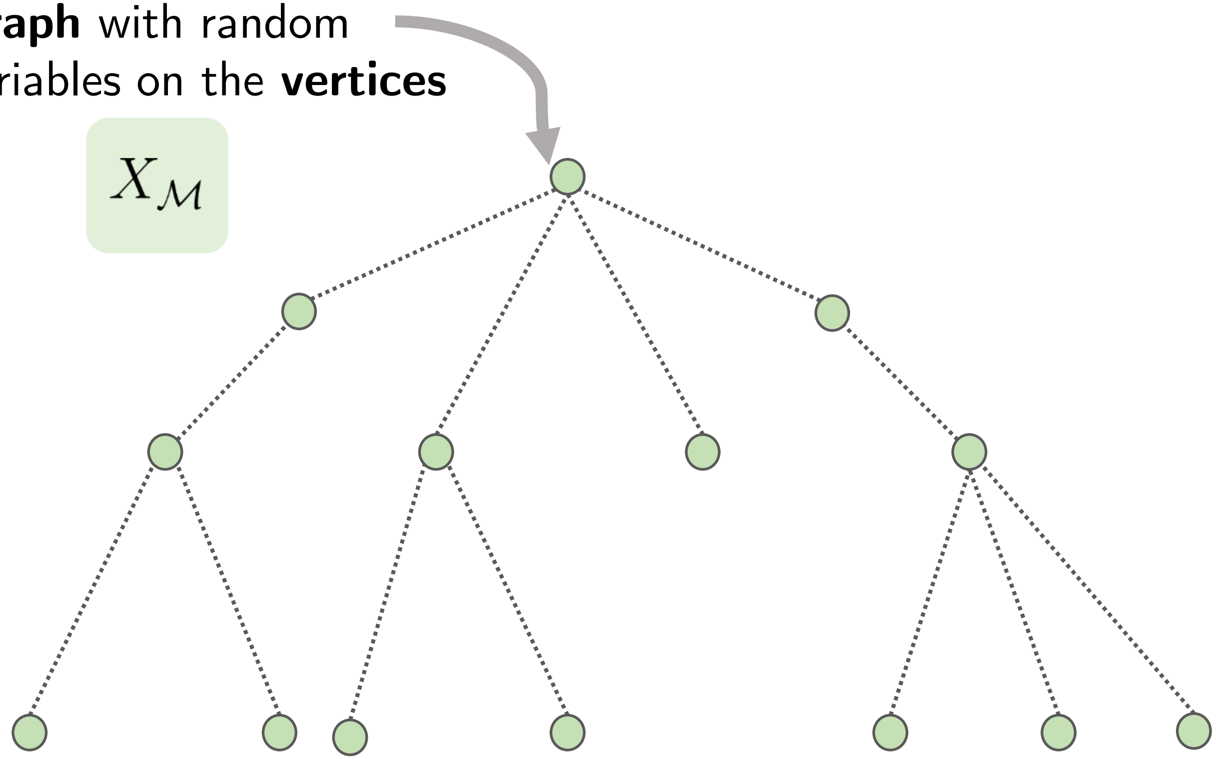
Graph with random
variables on the **vertices**



H.-O. Georgii, Gibbs Measures and Phase Transitions. De Gruyter, 2011.

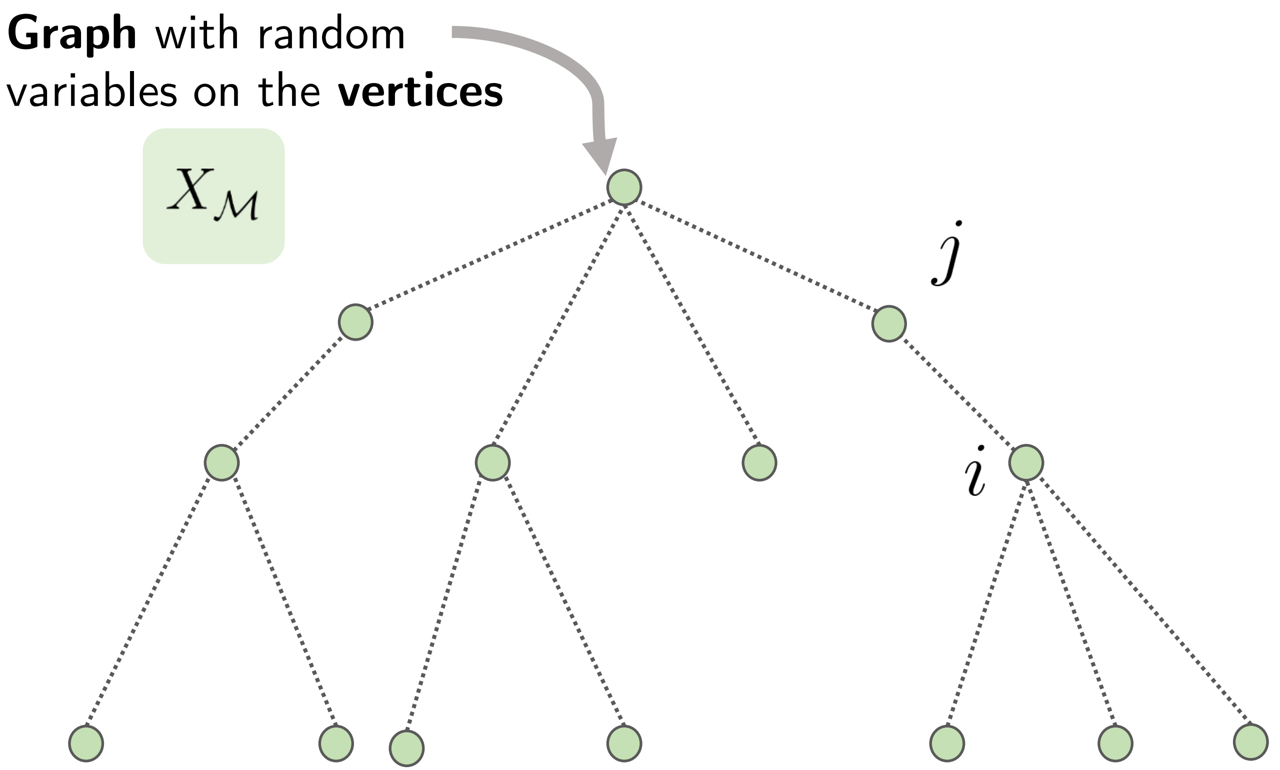
Markov chain on a tree (MCT)

Graph with random
variables on the **vertices**



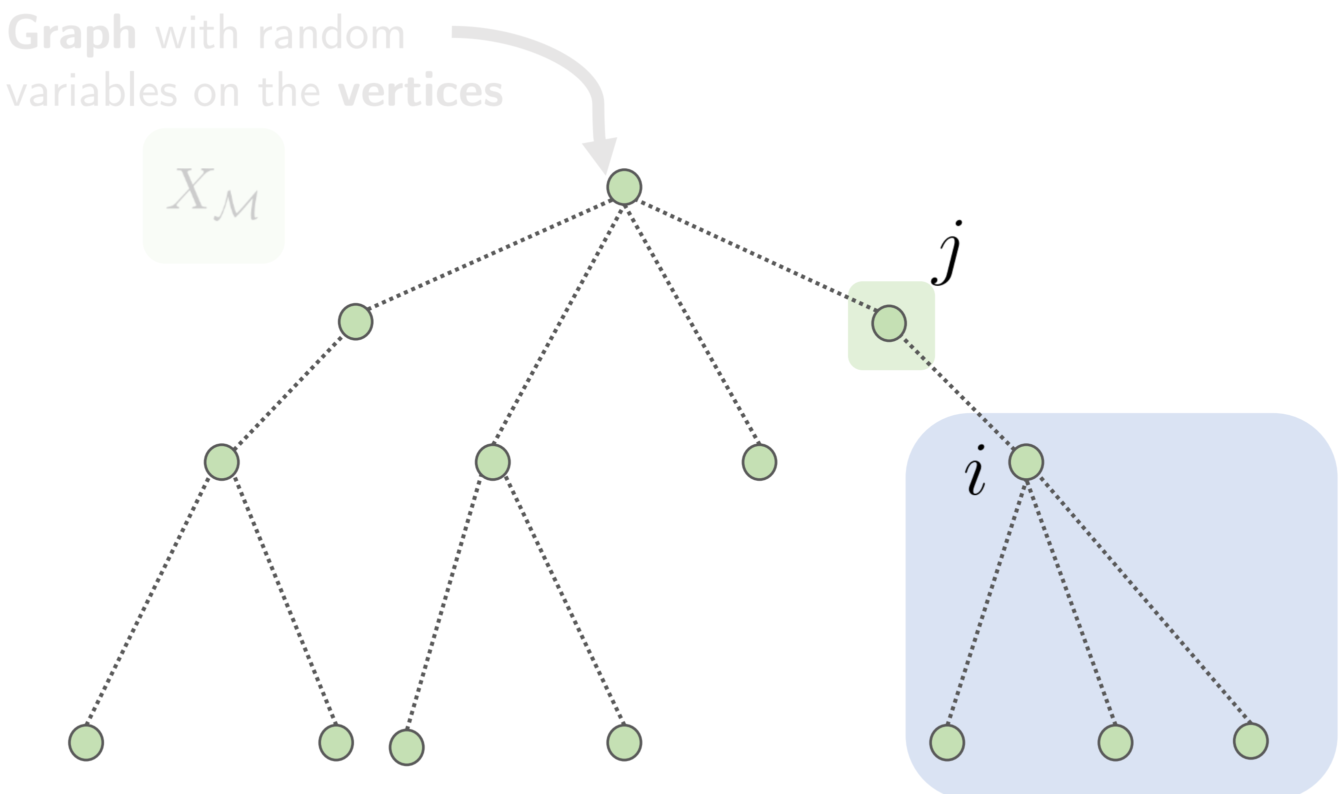
H.-O. Georgii, Gibbs Measures and Phase Transitions. De Gruyter, 2011.

Markov chain on a tree (MCT)



H.-O. Georgii, Gibbs Measures and Phase Transitions. De Gruyter, 2011.

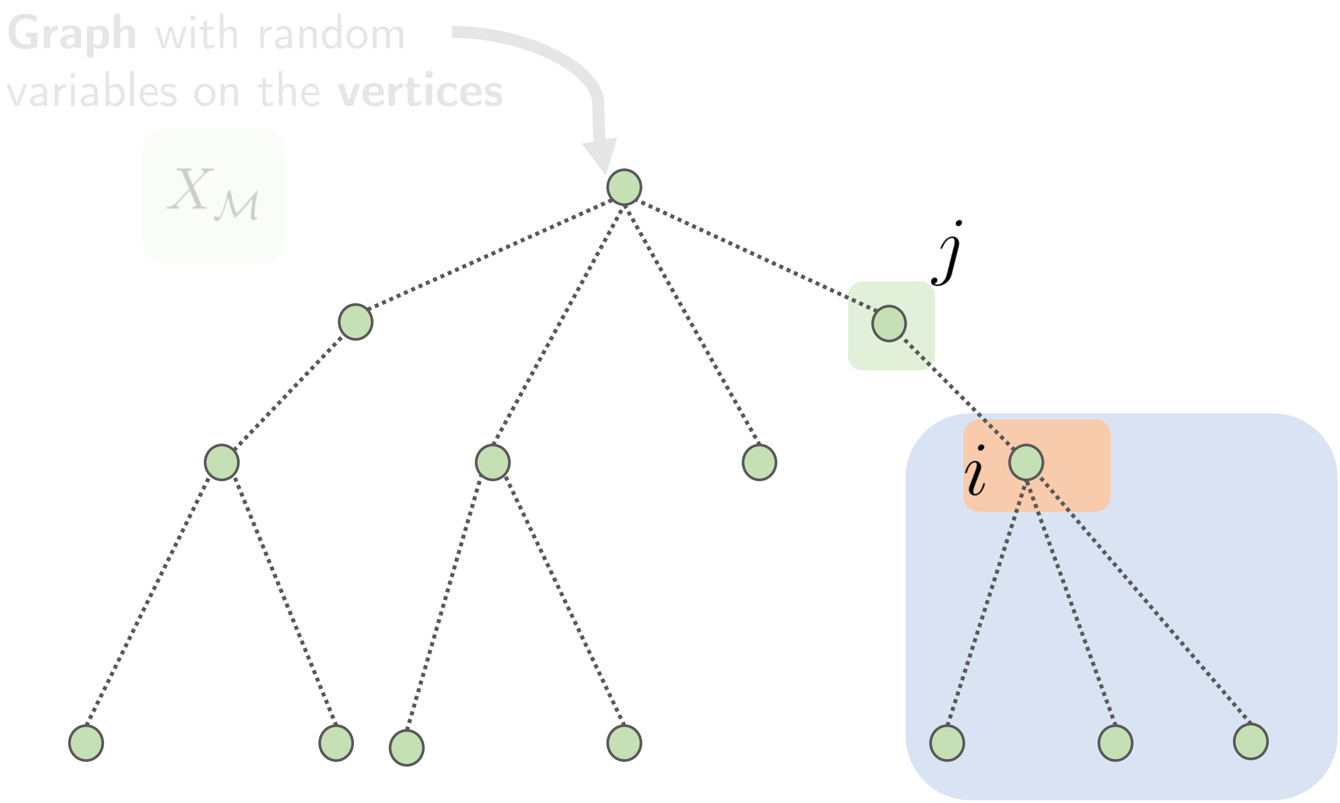
Markov chain on a tree (MCT)



$$P_{X_j \mid \mathcal{B}(i \leftarrow j)}$$

H.-O. Georgii, Gibbs Measures and Phase Transitions. De Gruyter, 2011.

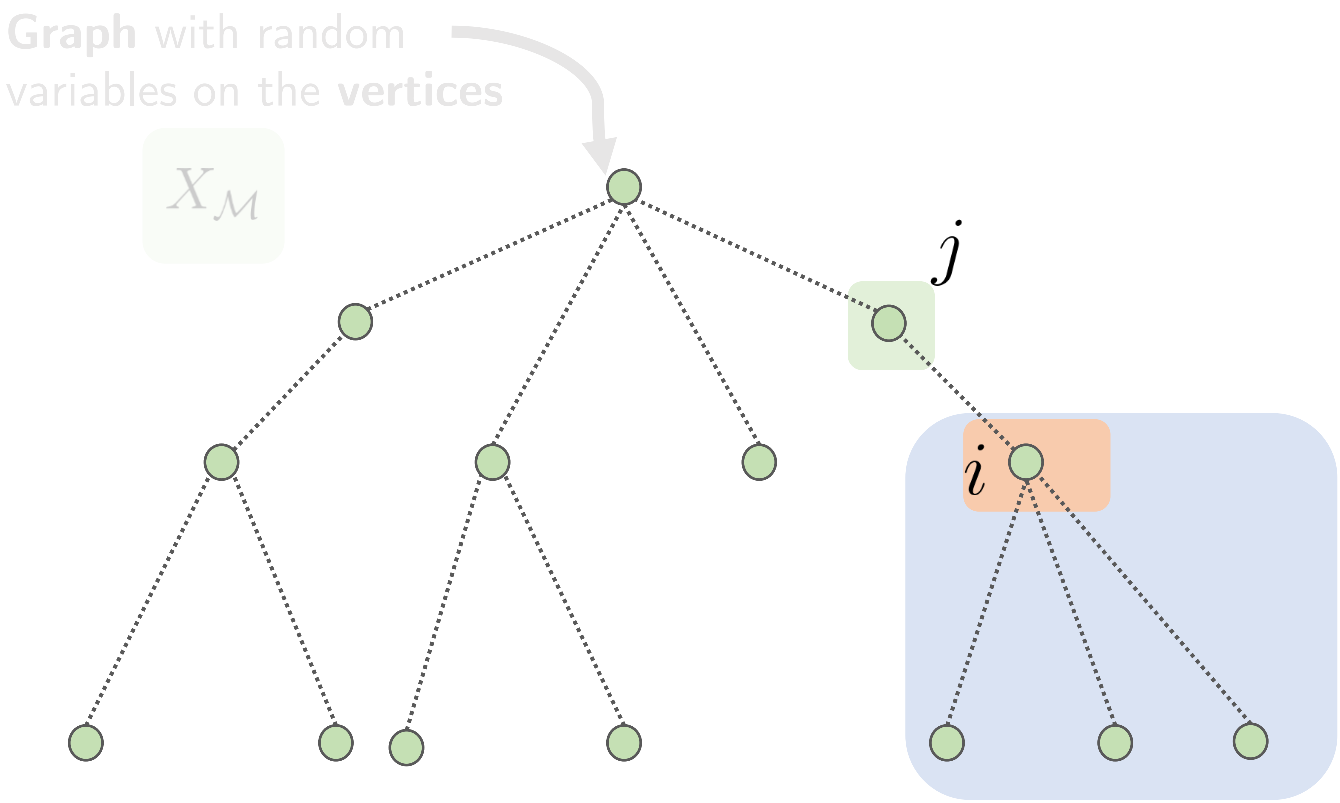
Markov chain on a tree (MCT)



$$P_{X_j \mid \mathcal{B}(i \leftarrow j)}$$

H.-O. Georgii, Gibbs Measures and Phase Transitions. De Gruyter, 2011.

Markov chain on a tree (MCT)



$$P_{X_j \mid \mathcal{B}(i \leftarrow j)} = P_{X_j \mid X_i}$$

H.-O. Georgii, Gibbs Measures and Phase Transitions. De Gruyter, 2011.

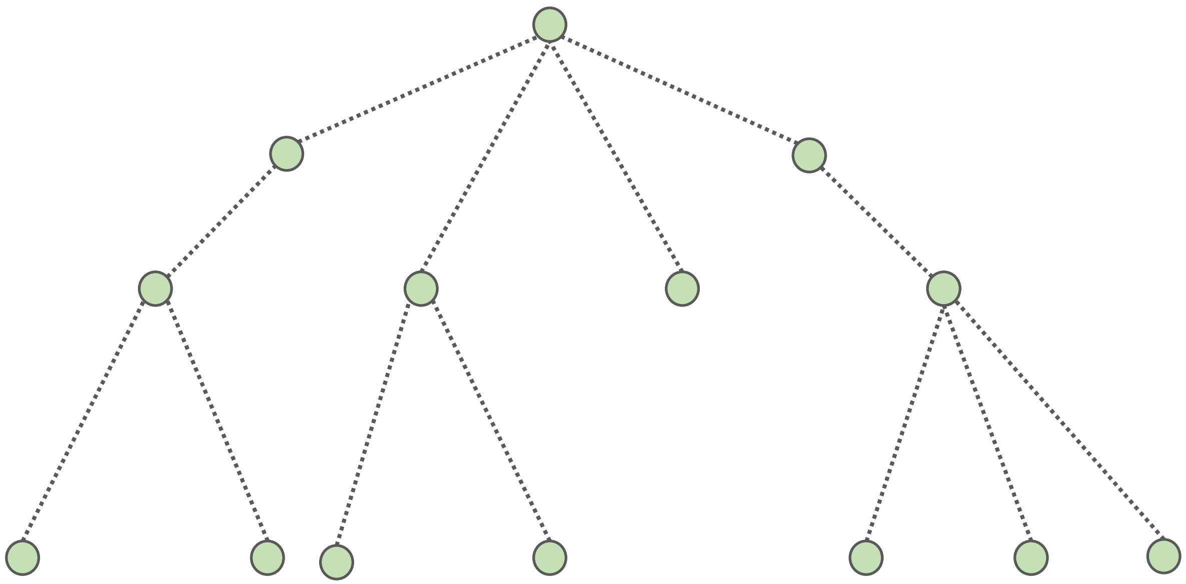
#3: SI in an MCT

SI in an MCT

- 2-partition achieves SI [Csiszár-Narayan, 2004]

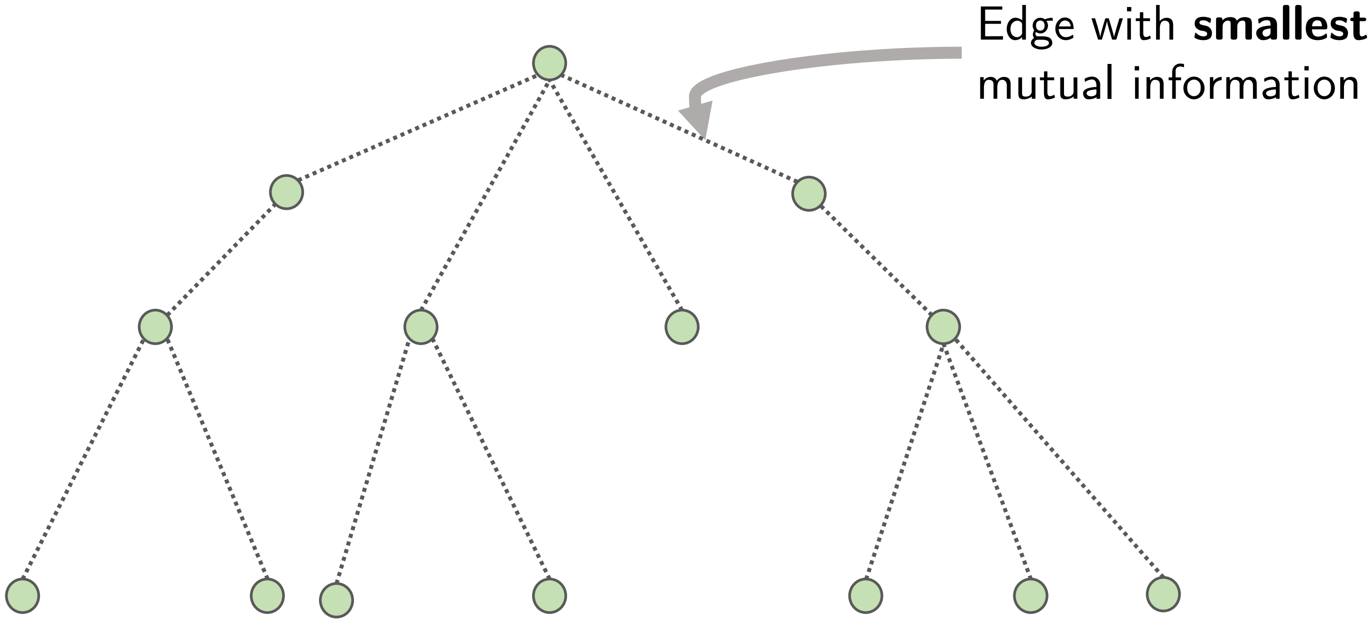
SI in an MCT

- 2-partition achieves SI [Csiszár-Narayan, 2004]



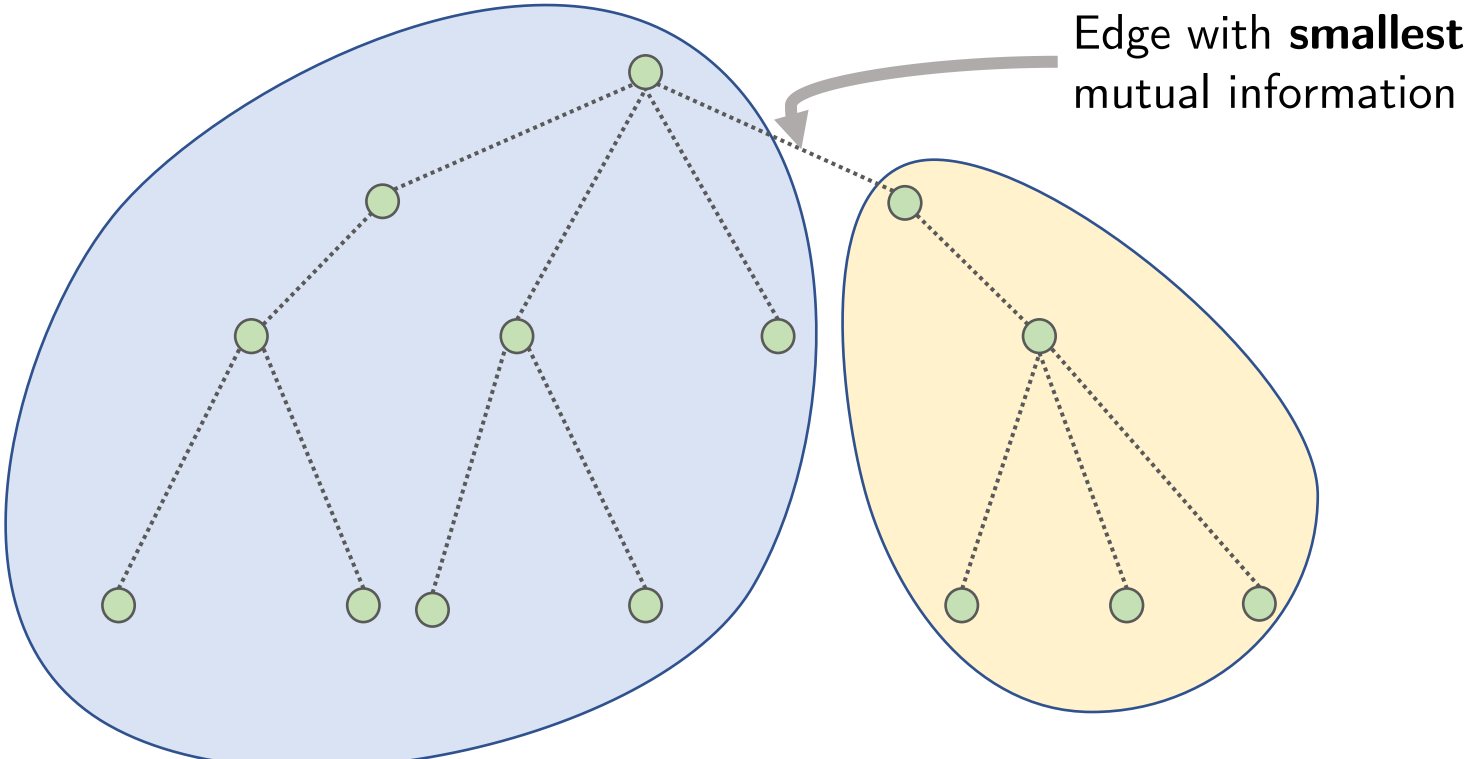
SI in an MCT

- 2-partition achieves SI [Csiszár-Narayan, 2004]



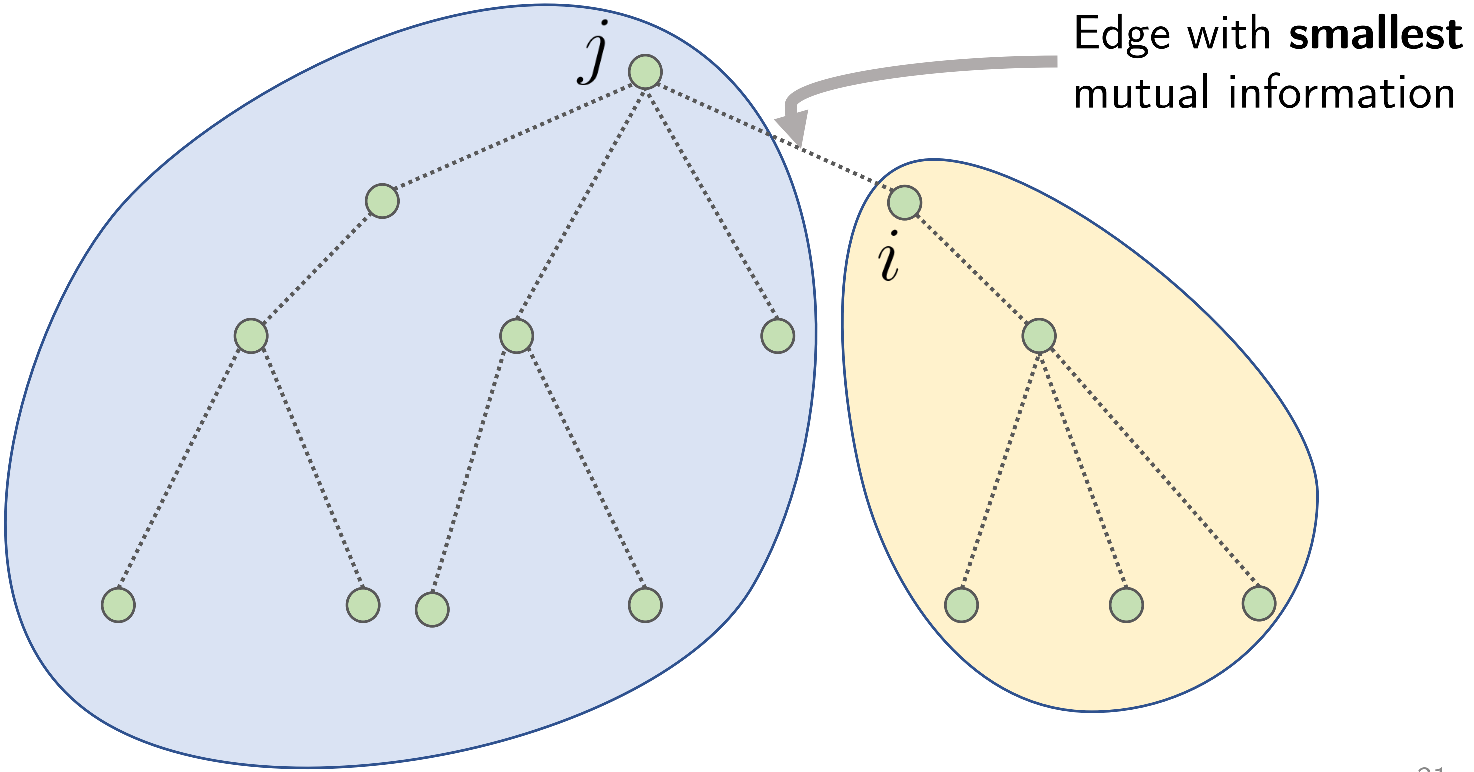
SI in an MCT

- 2-partition achieves SI [Csiszár-Narayan, 2004]



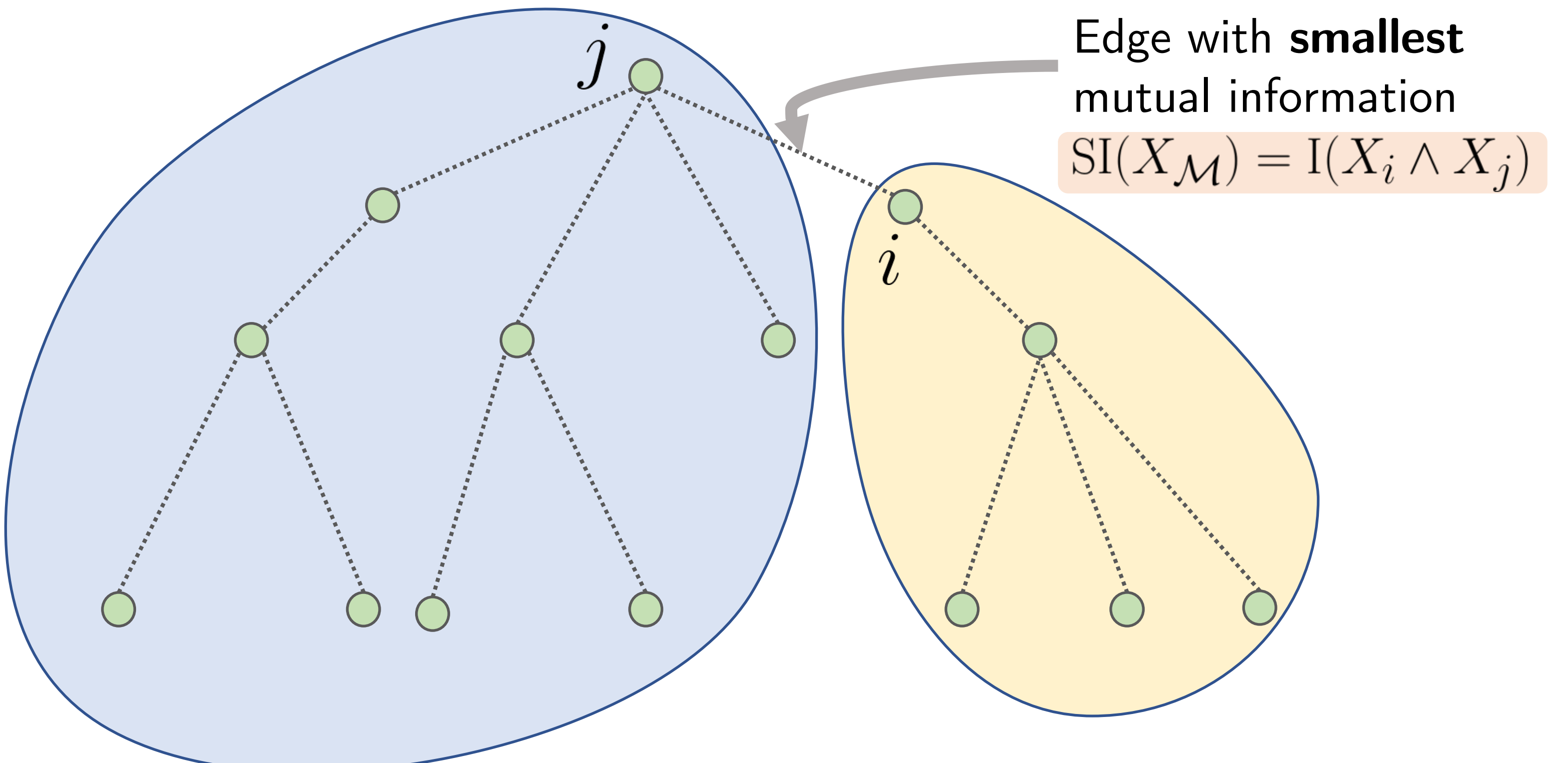
SI in an MCT

- 2-partition achieves SI [Csiszár-Narayan, 2004]



SI in an MCT

- 2-partition achieves SI [Csiszár-Narayan, 2004]



Shared information – simpler forms in special cases

$$\mathrm{SI}(X_{\mathcal{M}}) = \min_{2\leq k\leq m} \min_{\pi=(\pi_u,u=1,\cdots,k)} \frac{1}{k-1} D(P_{X_{\mathcal{M}}} \parallel \prod_{u=1}^{k} P_{X_{\pi_u}})$$

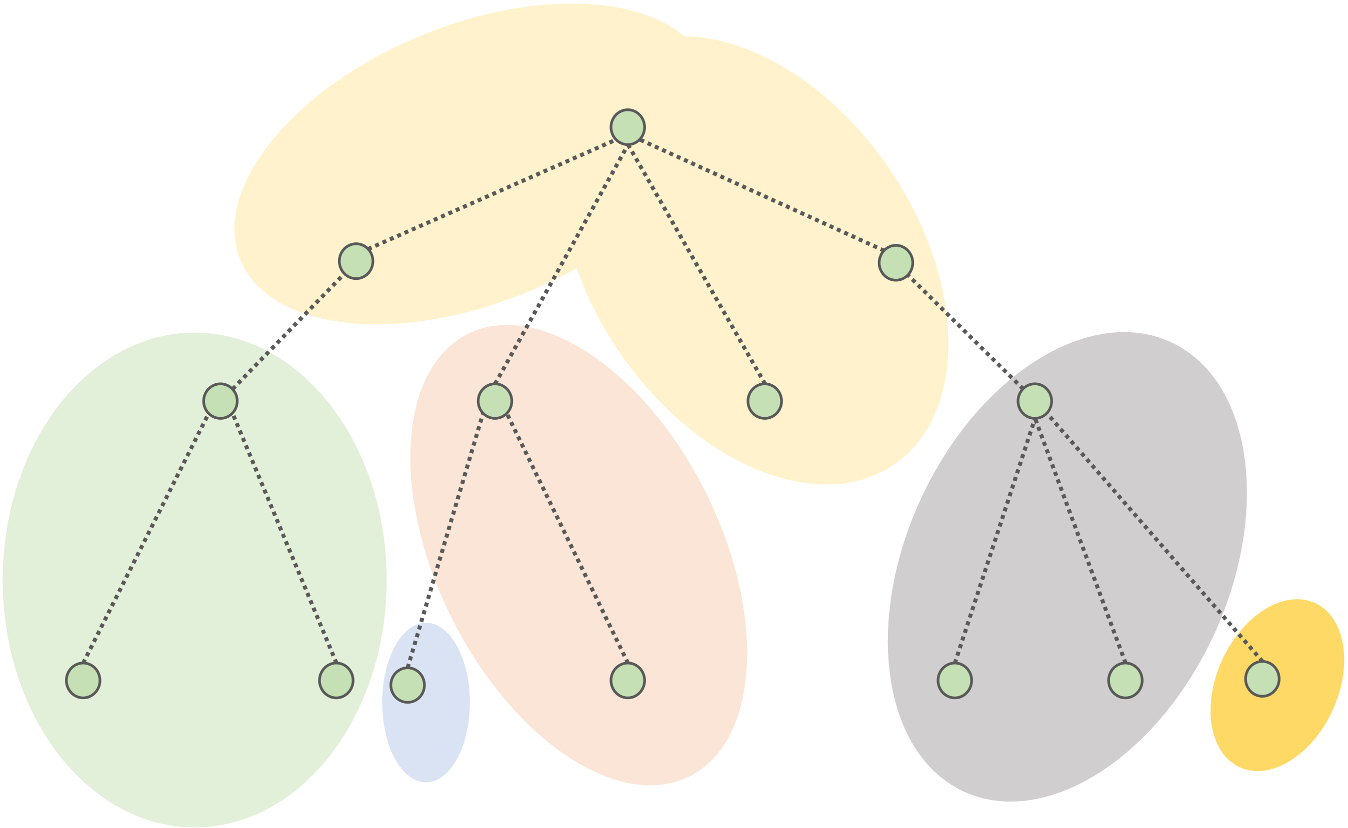
- Upper bound is clear*
  - Choose that special partition

Shared information – simpler forms in special cases

$$\mathrm{SI}(X_{\mathcal{M}}) = \min_{2 \le k \le m} \min_{\pi=(\pi_u, u=1,\cdots,k)} \frac{1}{k-1} D(P_{X_{\mathcal{M}}} \parallel \prod_{u=1}^{k} P_{X_{\pi_u}})$$

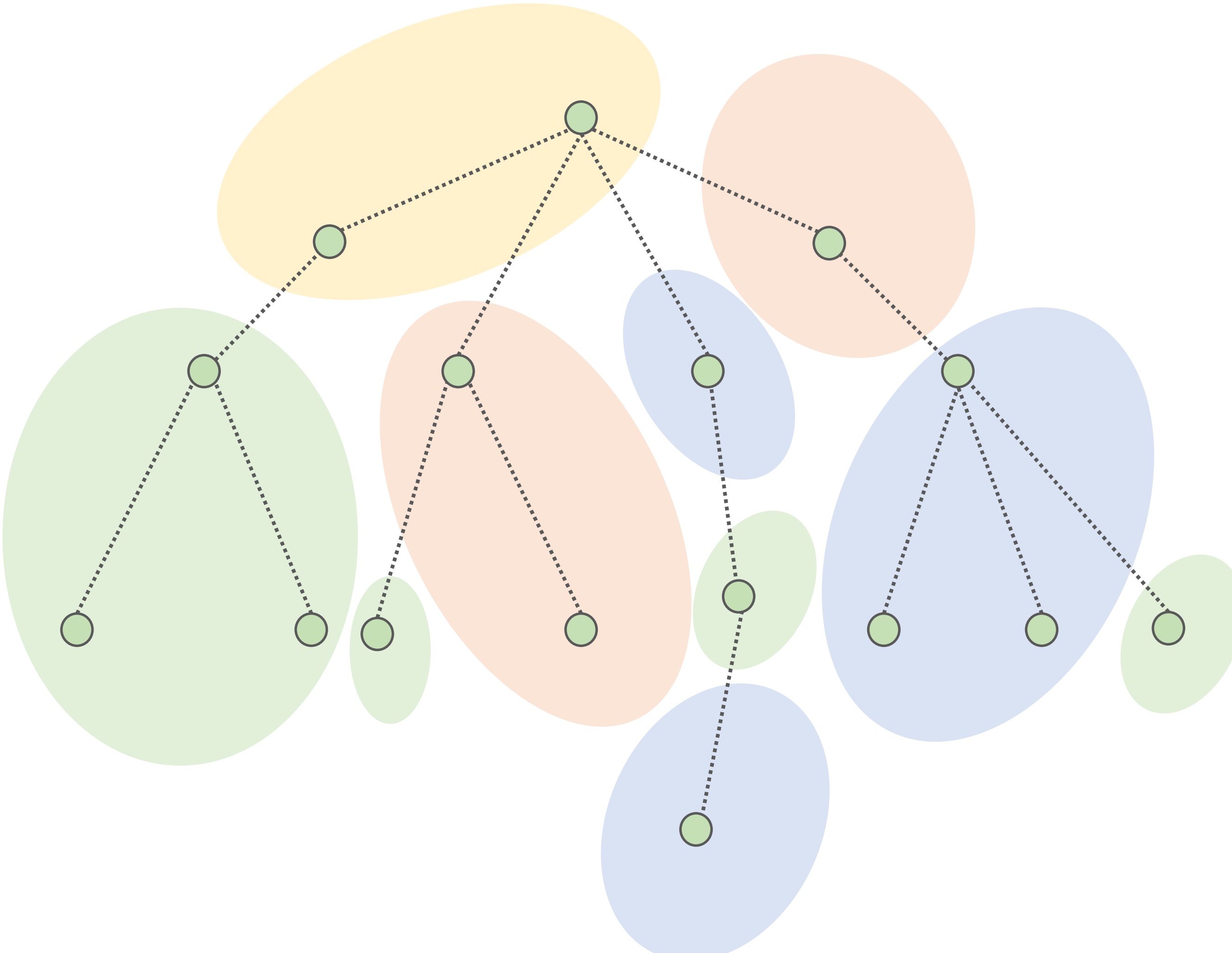
- Upper bound is clear*
  - Choose that special partition
- Need: lower bound
  - Original proof from secret-key capacity [Csiszár-Narayan, 2004]
  - Different in spirit from [Chan-Bashabsheh-Zhou-Kaced-Liu, 2016]
  - Easy* when atoms of the partition are connected

Connected atoms

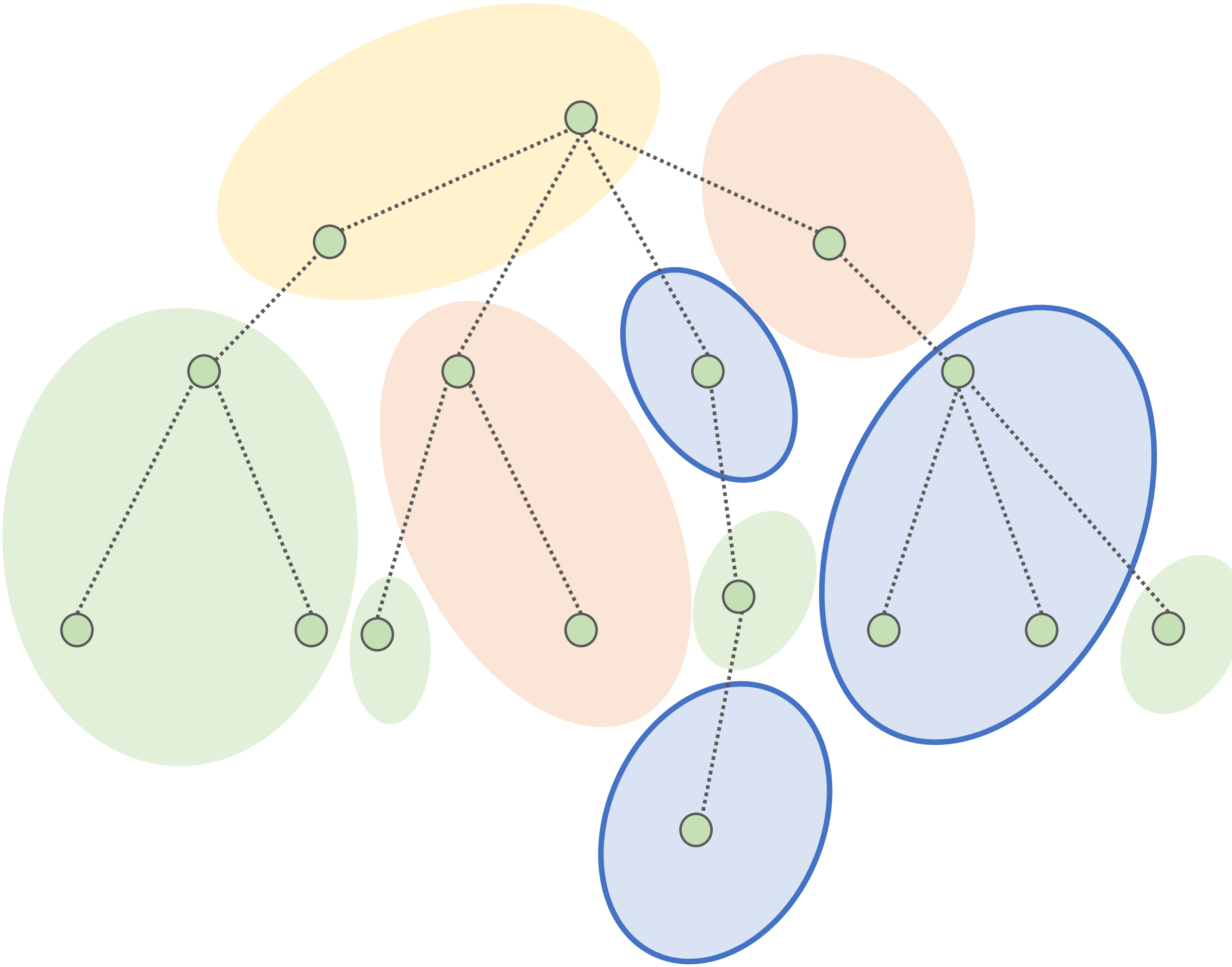


Idea: reduce nonconnected atom case to this case

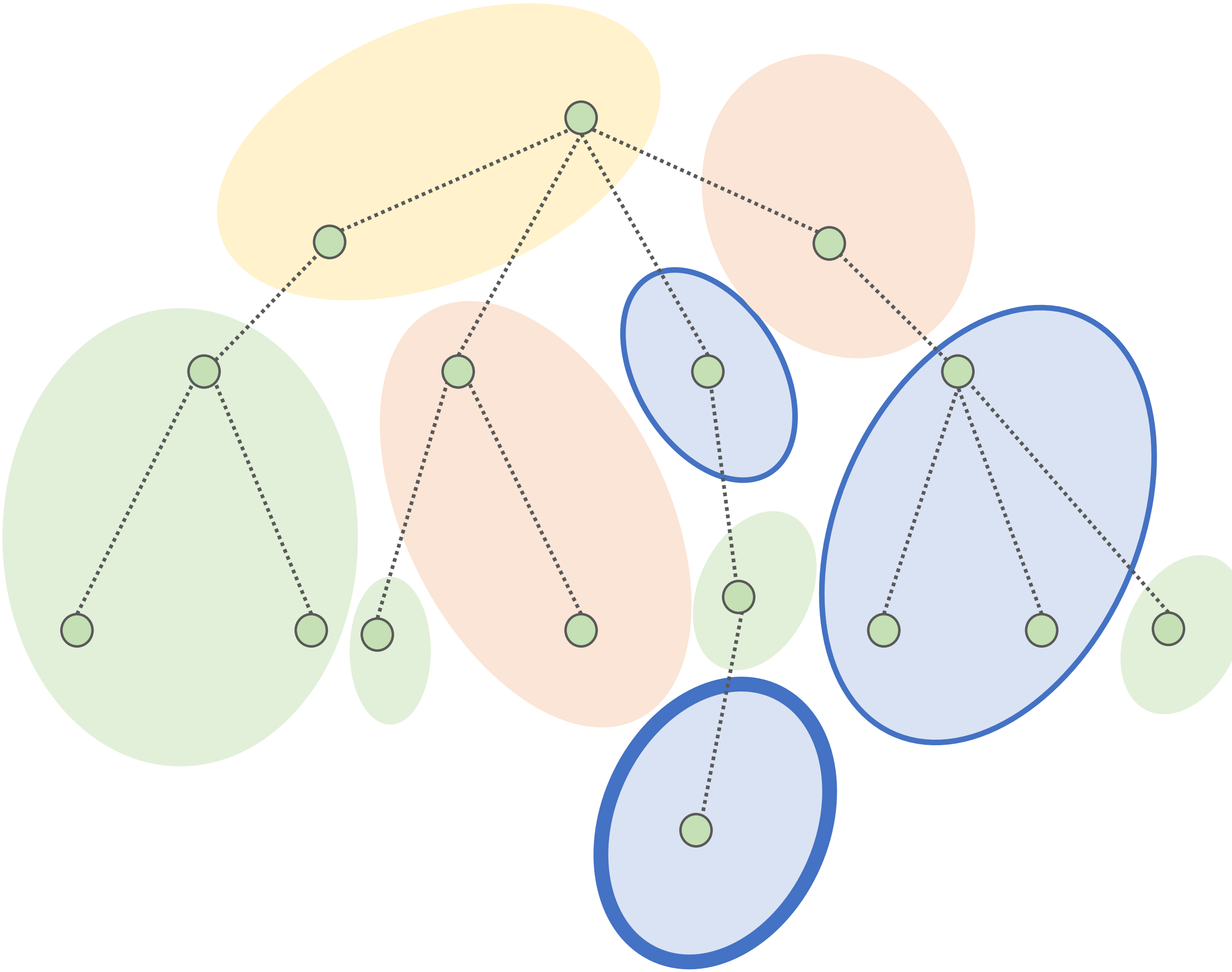
Idea: reduce nonconnected atom case to this case



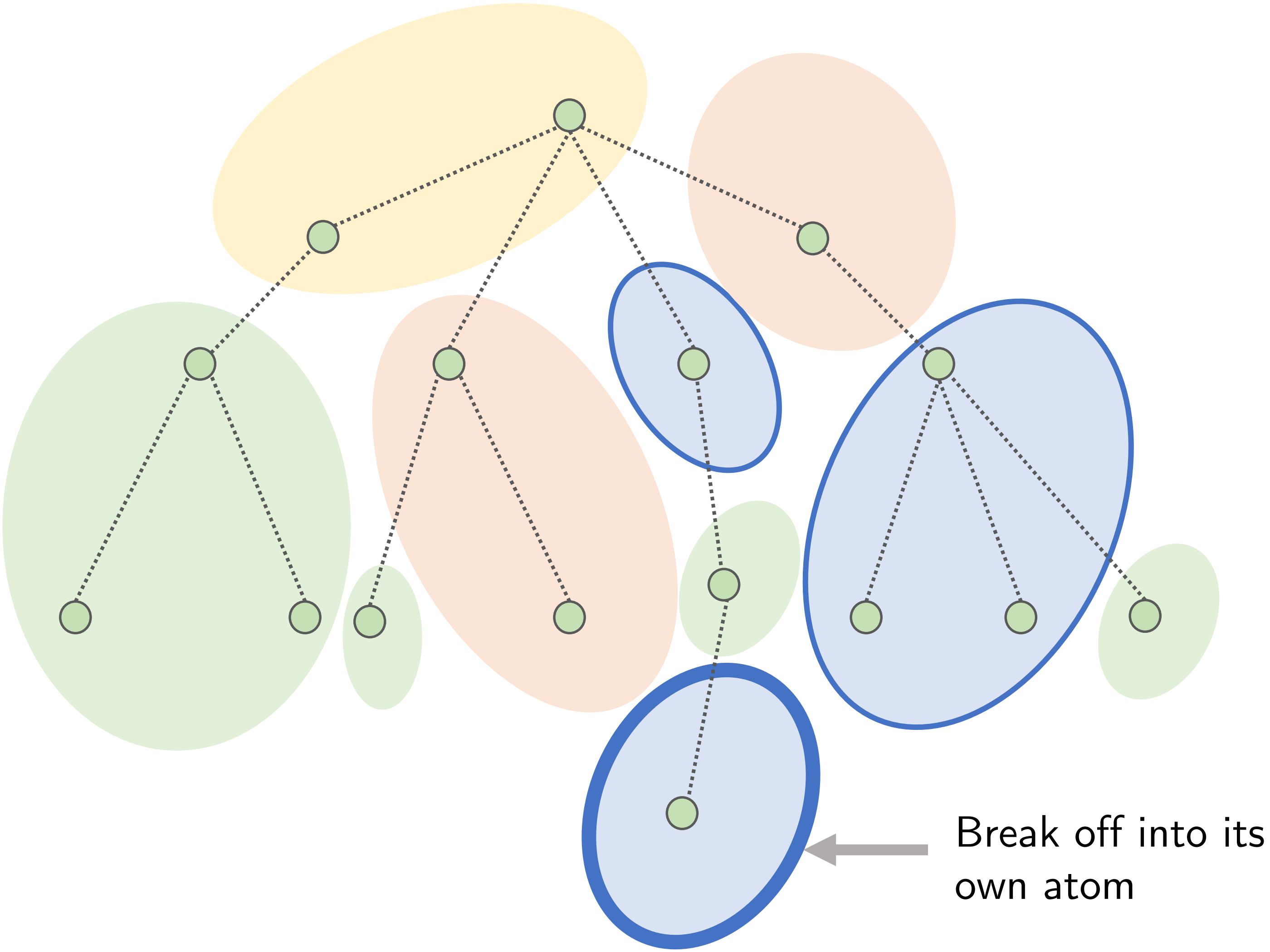
Idea: reduce nonconnected atom case to this case



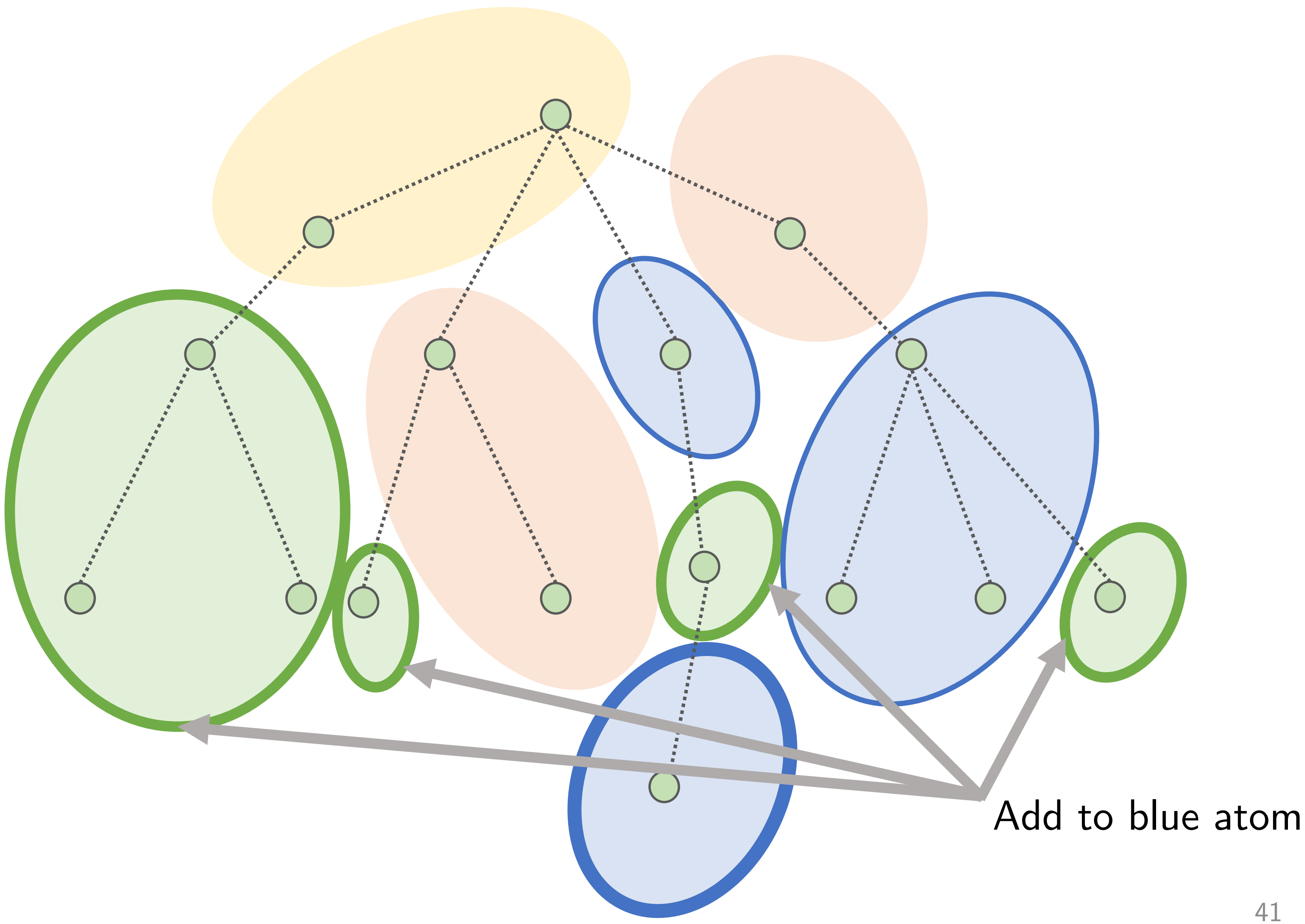
Idea: reduce nonconnected atom case to this case

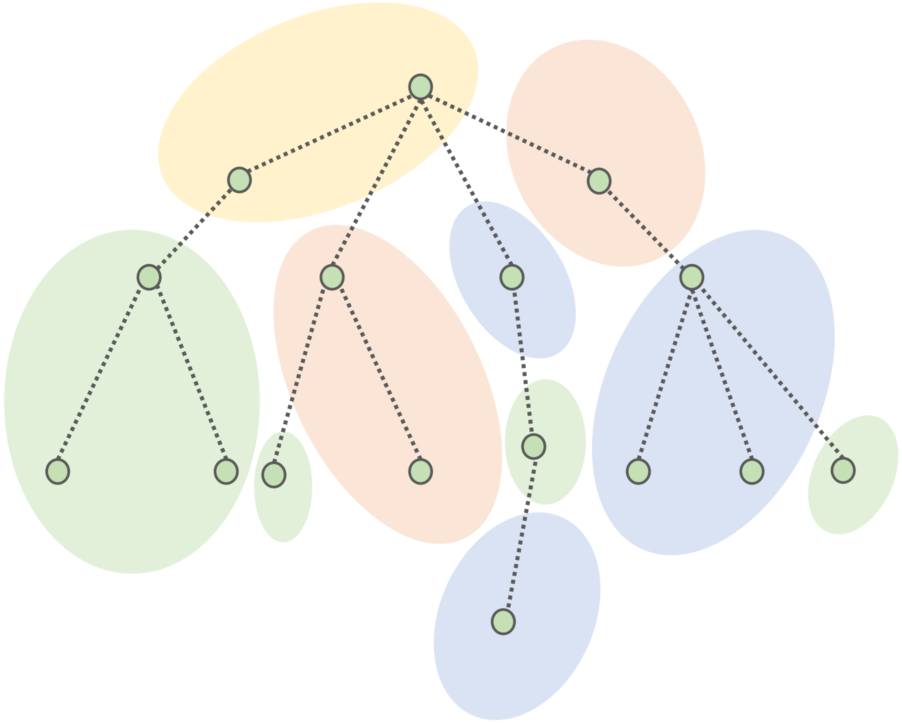


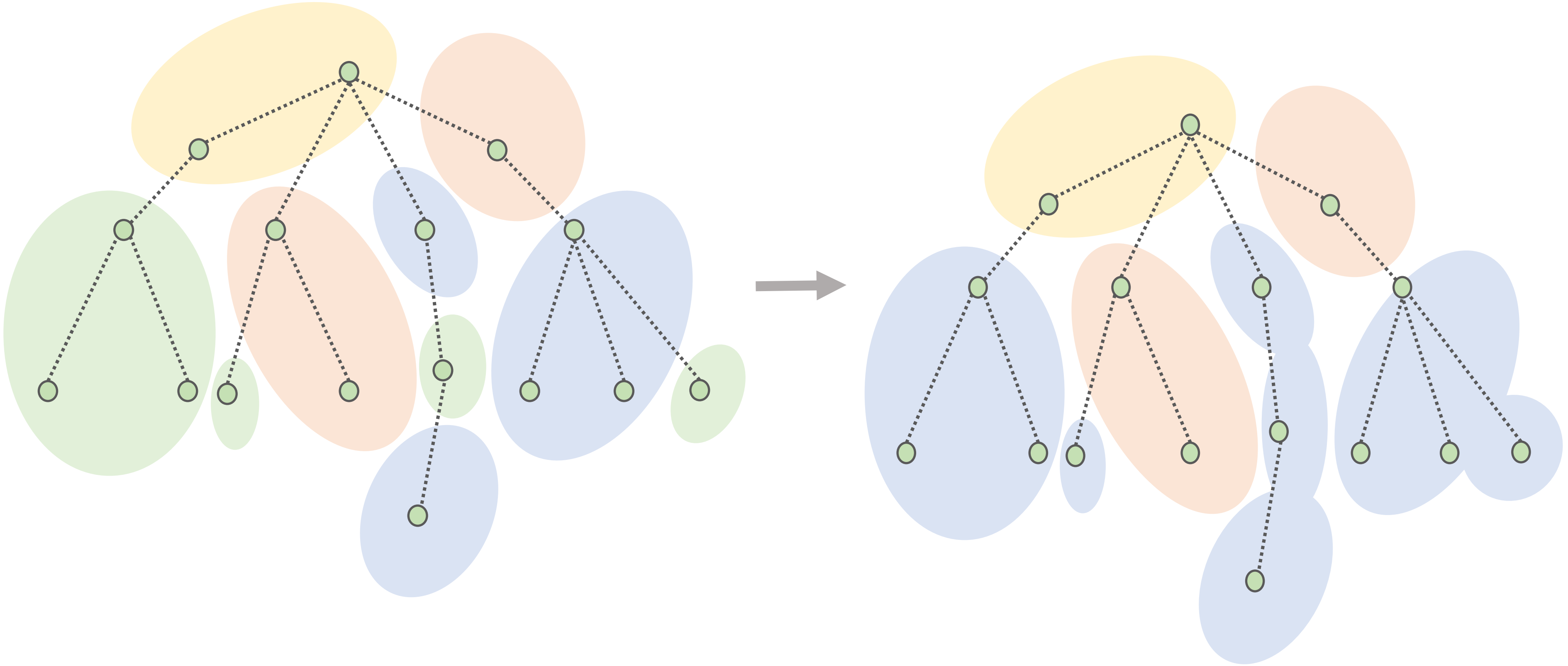
Idea: reduce nonconnected atom case to this case

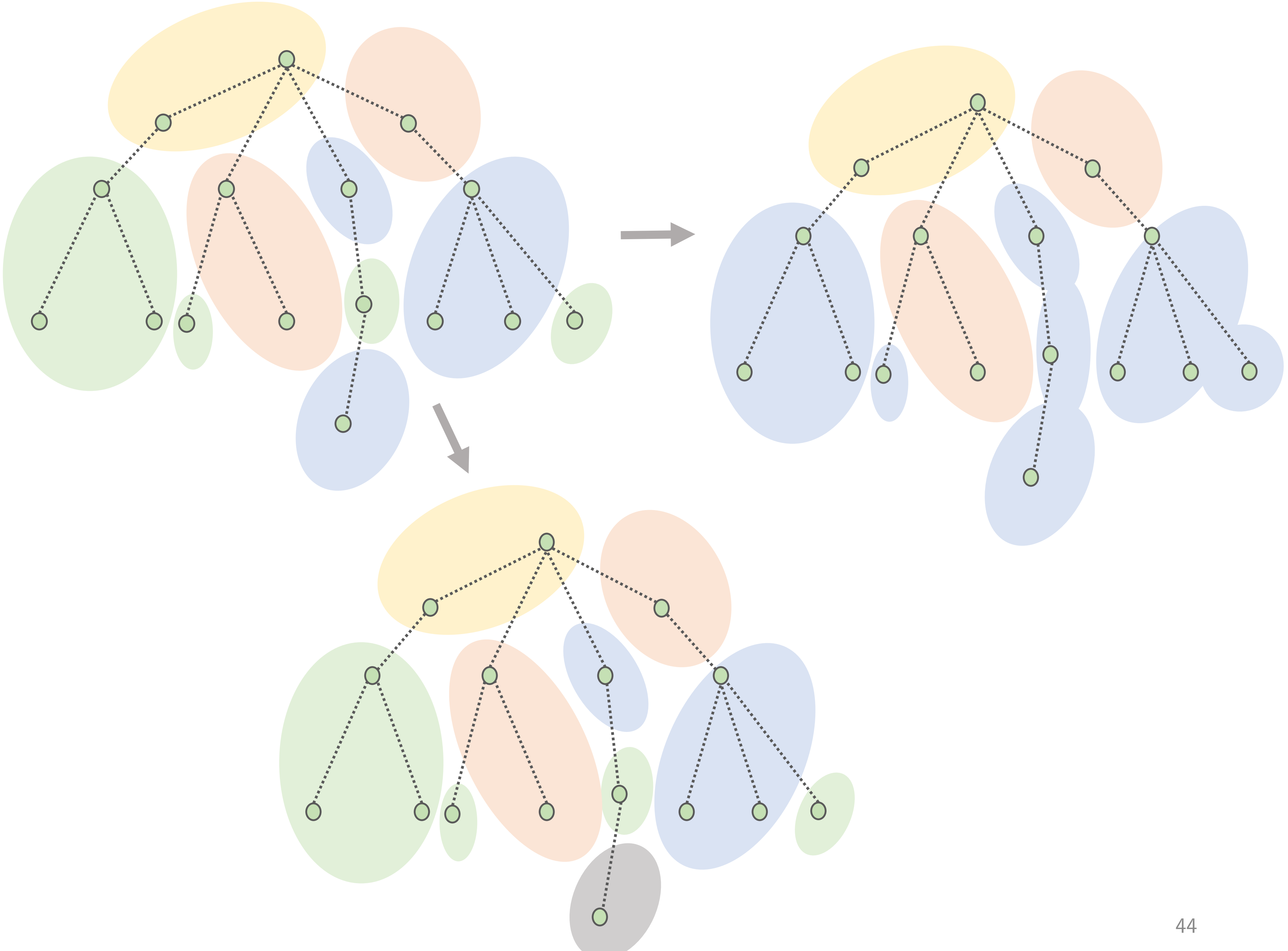


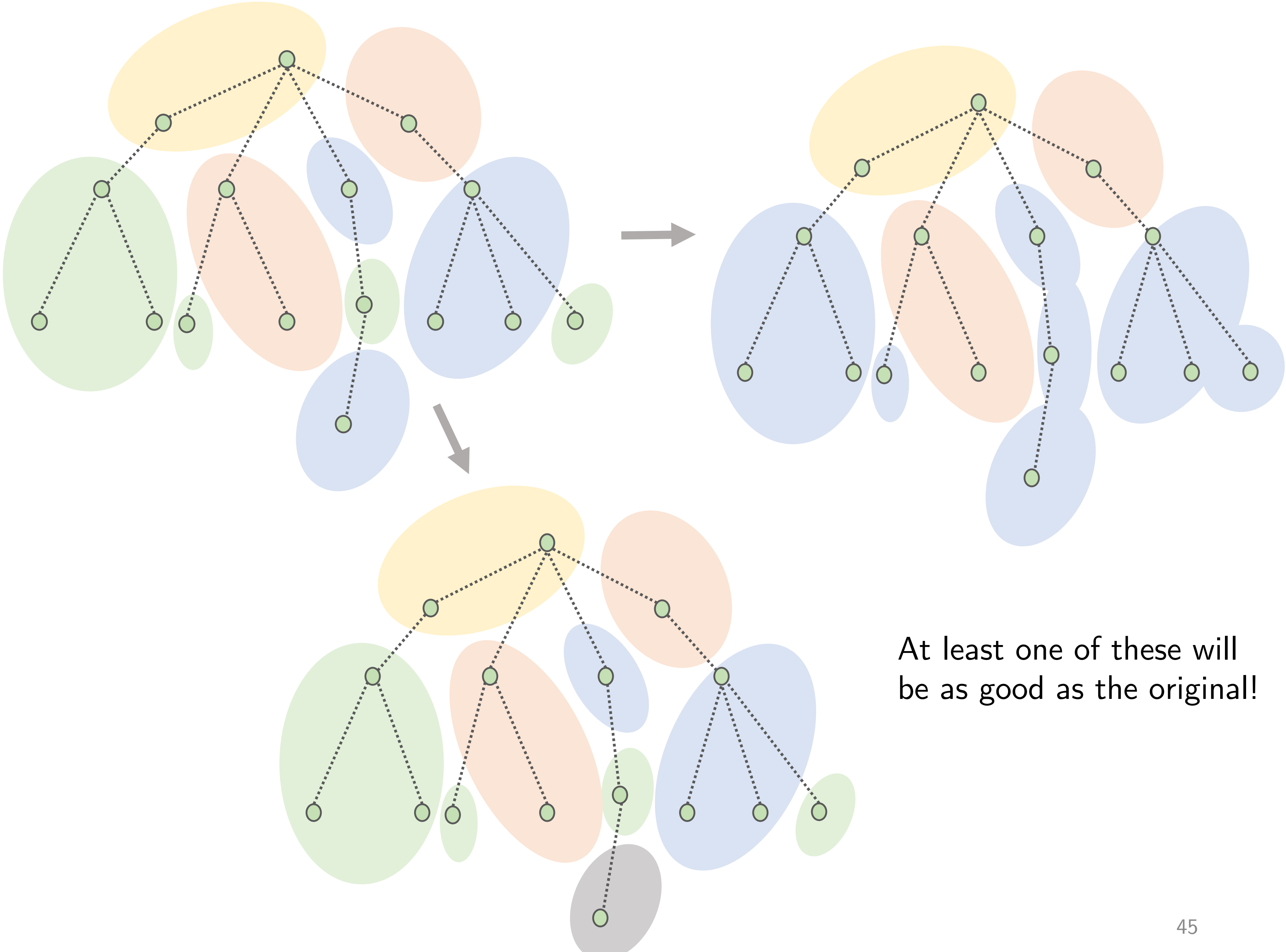
Idea: reduce nonconnected atom case to this case











At least one of these will be as good as the original!

#4: Estimating SI in an MCT

Estimating SI in an MCT

- Must find the edge with the lowest mutual information
    - Reduces to best arm identification [Audibert-Bubeck-Munos, 2010]

Estimating SI in an MCT

- Must find the edge with the lowest mutual information
  - Reduces to best arm identification [Audibert-Bubeck-Munos, 2010]
- Proposal: a **correlated** bandits algorithm

Estimating SI in an MCT

- Must find the edge with the lowest mutual information
  - Reduces to best arm identification [Audibert-Bubeck-Munos, 2010]
- Proposal: a **correlated** bandits algorithm
  - For every edge, sample both random variables at the same time
  - **Uniform sampling:** each pair is sampled equally often

Estimating SI in an MCT

- Must find the edge with the lowest mutual information
  - Reduces to best arm identification [Audibert-Bubeck-Munos, 2010]
- Proposal: a **correlated** bandits algorithm
  - For every edge, sample both random variables at the same time
  - **Uniform sampling:** each pair is sampled equally often
- **Key challenge:** estimators for mutual information are always **biased**.

MI estimation

- The **empirical mutual information** [Goppa, 1975]

$$\mathrm{I}_{\mathsf{EMI}}^{(n)}(\boldsymbol{x} \wedge \boldsymbol{y}) = \mathrm{H}(P_{\boldsymbol{x}}^{(n)}) + \mathrm{H}(P_{\boldsymbol{y}}^{(n)}) - \mathrm{H}(P_{\boldsymbol{xy}}^{(n)})$$

MI estimation

- The **empirical mutual information** [Goppa, 1975]

$$\mathrm{I}^{(n)}_{\mathsf{EMI}}(\boldsymbol{x} \wedge \boldsymbol{y}) = \mathrm{H}(P^{(n)}_{\boldsymbol{x}}) + \mathrm{H}(P^{(n)}_{\boldsymbol{y}}) - \mathrm{H}(P^{(n)}_{\boldsymbol{xy}})$$

- [Paninski, 2003]

$$-\log\left(1 + \frac{|\mathcal{X}| - 1}{n}\right)\left(1 + \frac{|\mathcal{Y}| - 1}{n}\right) \leq \mathrm{Bias}(\mathrm{I}^{(n)}_{\mathsf{EMI}}(\boldsymbol{X} \wedge \boldsymbol{Y})) \leq \log\left(1 + \frac{|\mathcal{X}|\,|\mathcal{Y}| - 1}{n}\right)$$

MI estimation

- The **empirical mutual information** [Goppa, 1975]

$$\mathrm{I}^{(n)}_{\mathsf{EMI}}(\boldsymbol{x} \wedge \boldsymbol{y}) = \mathrm{H}(P^{(n)}_{\boldsymbol{x}}) + \mathrm{H}(P^{(n)}_{\boldsymbol{y}}) - \mathrm{H}(P^{(n)}_{\boldsymbol{xy}})$$

- [Paninski, 2003]

$$-\log\left(1 + \frac{|\mathcal{X}| - 1}{n}\right)\left(1 + \frac{|\mathcal{Y}| - 1}{n}\right) \leq \mathrm{Bias}(\mathrm{I}^{(n)}_{\mathsf{EMI}}(\boldsymbol{X} \wedge \boldsymbol{Y})) \leq \log\left(1 + \frac{|\mathcal{X}|\,|\mathcal{Y}| - 1}{n}\right)$$

- Using McDiarmid's inequality, similar to [Antos-Kontoyiannis, 2001]

$$P_{XY}\left(\mathrm{I}^{(n)}_{\mathsf{EMI}}(\boldsymbol{X} \wedge \boldsymbol{Y}) - \mathbb{E}_{P_{XY}}\left[\mathrm{I}^{(n)}_{\mathsf{EMI}}(\boldsymbol{X} \wedge \boldsymbol{Y})\right] \geq \epsilon\right) \leq \exp\left(-\frac{2n\epsilon^2}{36\log^2 n}\right)$$

Misidentification probability

$$P_{X_{\mathcal{M}}}\left(\hat{e}_N(X_{\mathcal{M}}^N) \neq (\bar{i}, \bar{j})\right) \leq 2\,|\mathcal{E}| \exp\left(\frac{-(N/\,|\mathcal{E}|)\Delta_1^2}{648\log^2(N/\,|\mathcal{E}|)}\right)$$

if

$$N > |\mathcal{E}| \max\left\{\frac{|\mathcal{X}|^2 - 1}{2^{\Delta_1/3} - 1}, \frac{|\mathcal{X}| - 1}{2^{\Delta_1/6} - 1}\right\},$$

Misidentification probability

Number of edges

$$P_{X_{\mathcal{M}}}\left(\hat{e}_N(X_{\mathcal{M}}^N) \neq (\bar{i}, \bar{j})\right) \leq 2\,|\mathcal{E}|\exp\left(\frac{-(N/\,|\mathcal{E}|)\Delta_1^2}{648\log^2(N/\,|\mathcal{E}|)}\right)$$

if

$$N > |\mathcal{E}|\max\left\{\frac{|\mathcal{X}|^2-1}{2^{\Delta_1/3}-1}, \frac{|\mathcal{X}|-1}{2^{\Delta_1/6}-1}\right\},$$

Misidentification probability

Difference between
the smallest and
second-smallest MIs

Number of edges

$$P_{X_{\mathcal{M}}}\left(\hat{e}_N(X_{\mathcal{M}}^N) \neq (\bar{i},\bar{j})\right) \leq 2\,|\mathcal{E}|\exp\left(\frac{-(N/\,|\mathcal{E}|)\Delta_1^2}{648\log^2(N/\,|\mathcal{E}|)}\right)$$

if

$$N > |\mathcal{E}|\max\left\{\frac{|\mathcal{X}|^2-1}{2^{\Delta_1/3}-1},\frac{|\mathcal{X}|-1}{2^{\Delta_1/6}-1}\right\},$$

Misidentification probability

Difference between the smallest and second-smallest MIs

Number of edges

$$P_{X_{\mathcal{M}}}\left(\hat{e}_N(X_{\mathcal{M}}^N) \neq (\bar{i},\bar{j})\right) \leq 2\,|\mathcal{E}|\exp\left(\frac{-(N/\,|\mathcal{E}|)\Delta_1^2}{648\log^2(N/\,|\mathcal{E}|)}\right)$$

if

$$N > |\mathcal{E}|\max\left\{\frac{|\mathcal{X}|^2-1}{2^{\Delta_1/3}-1},\frac{|\mathcal{X}|-1}{2^{\Delta_1/6}-1}\right\},$$

Reduces bias below Δ_1

Key takeaways

- New proof for SI in an MCT
- Best-arm identification using biased estimators

Directions for future work

- **Better estimators** lead to better bounds
- **Lower bounds** for best-arm identification using biased estimators