Privacy-preserving Mechanisms for Correlated Data

Kamalika Chaudhuri

University of California, San Diego

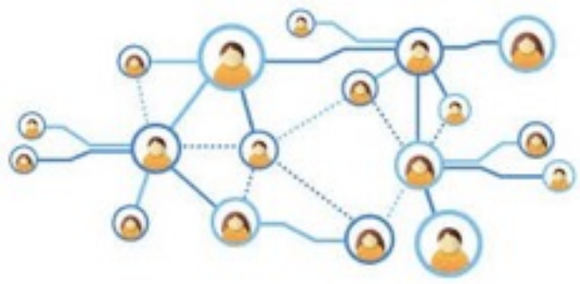
Joint work with Shuang Song and Yizhen Wang

Sensitive Data

Medical Records

Search Logs

Social Networks



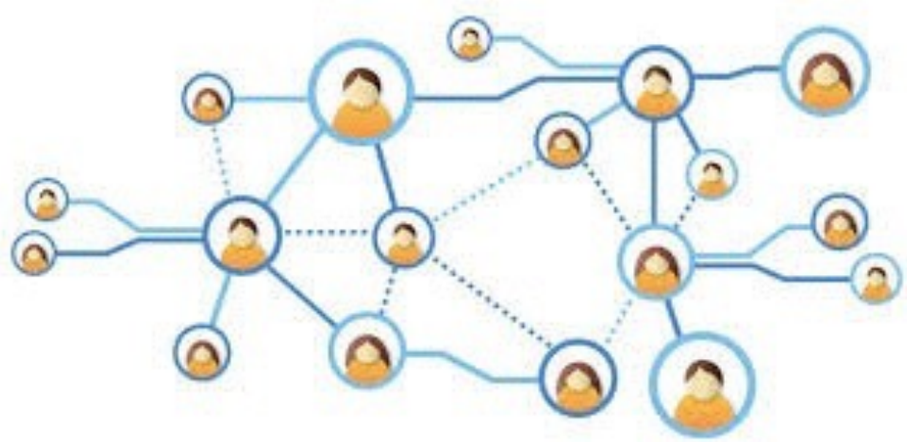
Talk Agenda:

How do we **analyze sensitive** data while still preserving **privacy**?

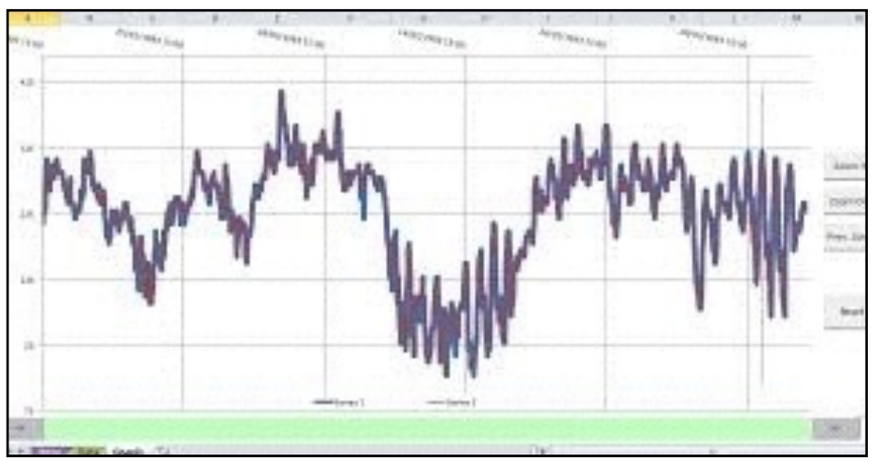
(Focus on **correlated data**)

Correlated Data

User information
in social networks



Physical Activity
Monitoring



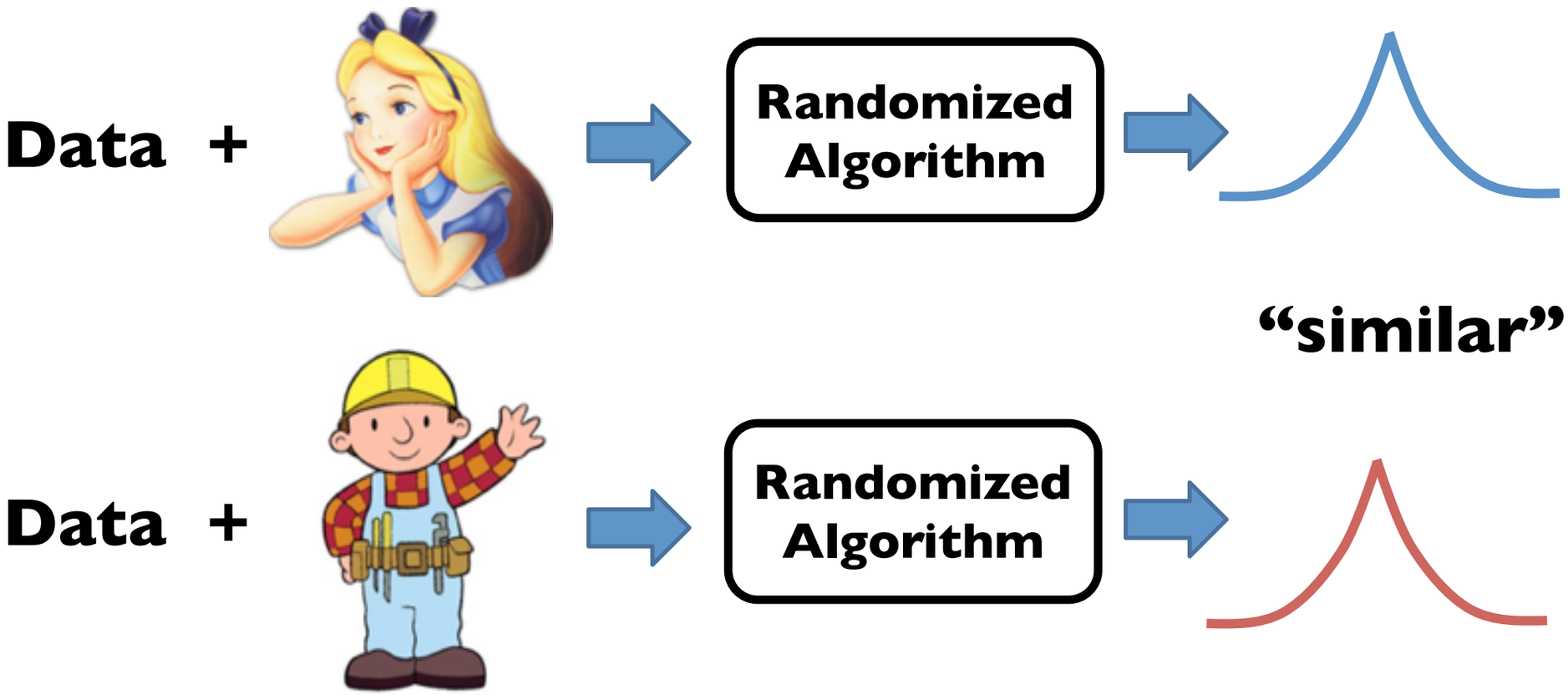
Why is Privacy Hard for Correlated Data?

Because neighbor's information leaks information on user

Talk Agenda:

1. Privacy for Correlated Data
   - How to define privacy (for uncorrelated data)

Differential Privacy [DMNS06]



Participation of a single person does not change output

Differential Privacy: Attacker's View

Prior Knowledge + **Algorithm** Output on Data &  = **Conclusion** on 

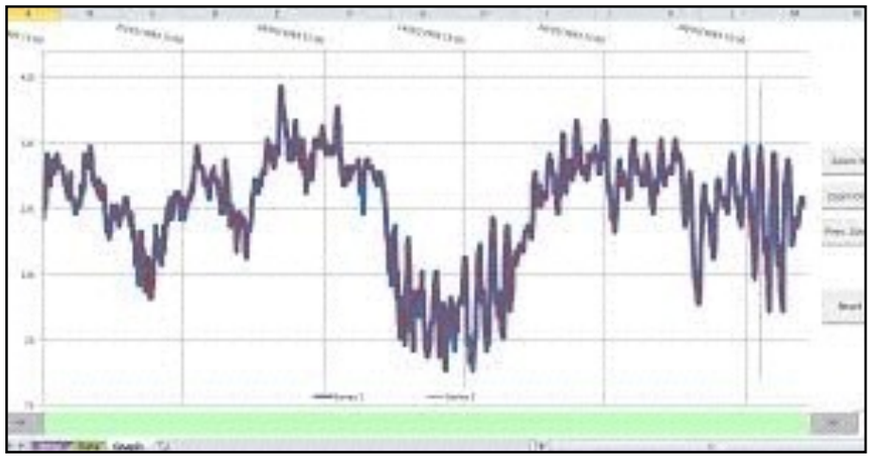
Prior Knowledge + **Algorithm** Output on Data &  = **Conclusion** on 

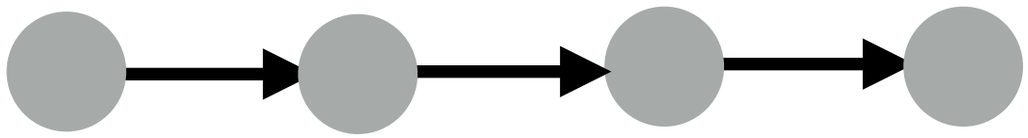
Note: a. Algorithm could draw **personal conclusions** about Alice

b. Alice has the **agency** to participate or not

What happens with correlated data?

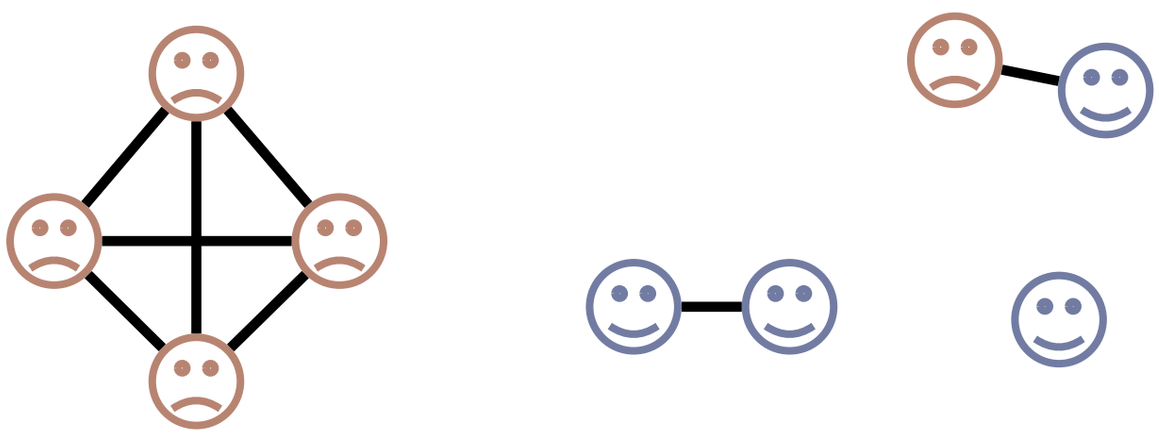
Example 1: Activity Monitoring



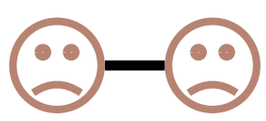


Goal: Share aggregate data on physical activity with doctor, while hiding activity at each specific time. Agency is at the individual level.

Example 2: Spread of Flu in Network



Interaction Network

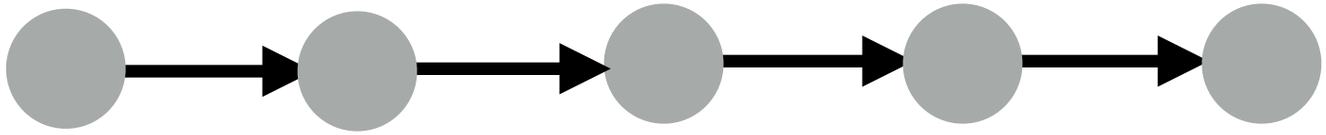


Goal: Publish aggregate statistics over a set of schools, prevent adversary from knowing who has flu. Agency at school level.

Why does Correlated data require a different notion of privacy?

Example: Activity Monitoring

$D = (x_1, .., x_T)$, x_t = activity at time t

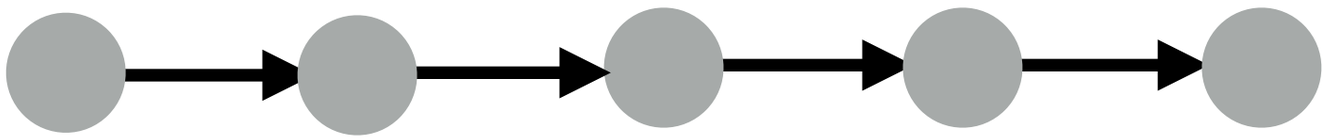


Correlation Network

Goal: (1) Publish activity histogram
(2) Prevent adversary from knowing activity at t

Example: Activity Monitoring

$D = (x_1, .., x_T)$, x_t = activity at time t



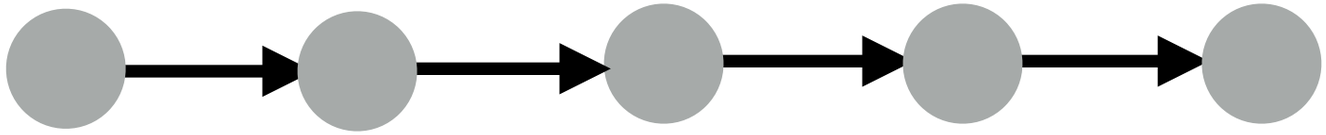
Correlation Network

Goal: (1) Publish activity histogram
(2) Prevent adversary from knowing activity at t

Agency is at individual level, not time entry level

Example: Activity Monitoring

$D = (x_1, .., x_T)$, x_t = activity at time t



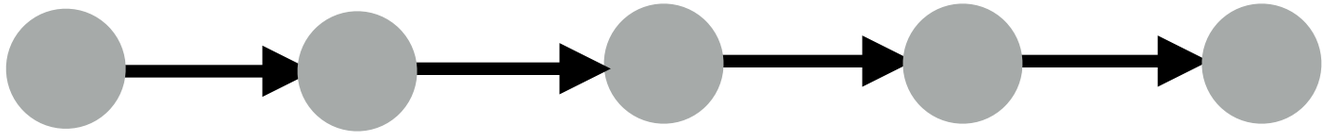
Correlation Network

1-DP: Output histogram of activities + noise with stdev T

Too much noise - no utility!

Example: Activity Monitoring

$D = (x_1, .., x_T)$, x_t = activity at time t



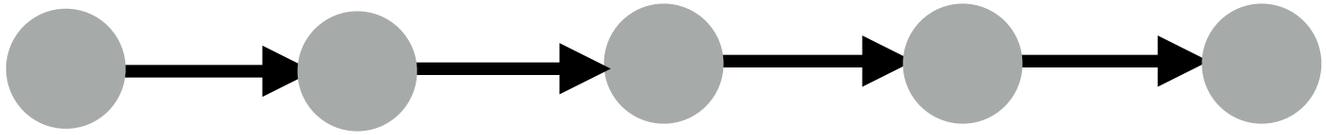
Correlation Network

1-entry-DP: Output histogram of activities + noise with stdev 1

Not enough - activities across time are correlated!

Example: Activity Monitoring

$D = (x_1, .., x_T)$, x_t = activity at time t



Correlation Network

1-Entry-Group DP: Output histogram of activities + noise with stdev T

Too much noise - no utility!

Pufferfish Privacy [KM12]

Secret Set S

S: Information to be protected

e.g: Alice's age is 25, Bob has a disease

Pufferfish Privacy [KM12]

Secret Set S

Secret Pairs Set Q

Q: Pairs of secrets we want to be indistinguishable

e.g: (Alice's age is 25, Alice's age is 40)

(Bob is in dataset, Bob is not in dataset)

Pufferfish Privacy [KM12]

Secret Set S | **Secret Pairs Set Q** | **Distribution Class Θ**

Θ: A set of distributions that plausibly generate the data

e.g: (connection graph G, disease transmits w.p [0.1, 0.5])

(Markov Chain with transition matrix in set **P**)

May be used to model correlation in data

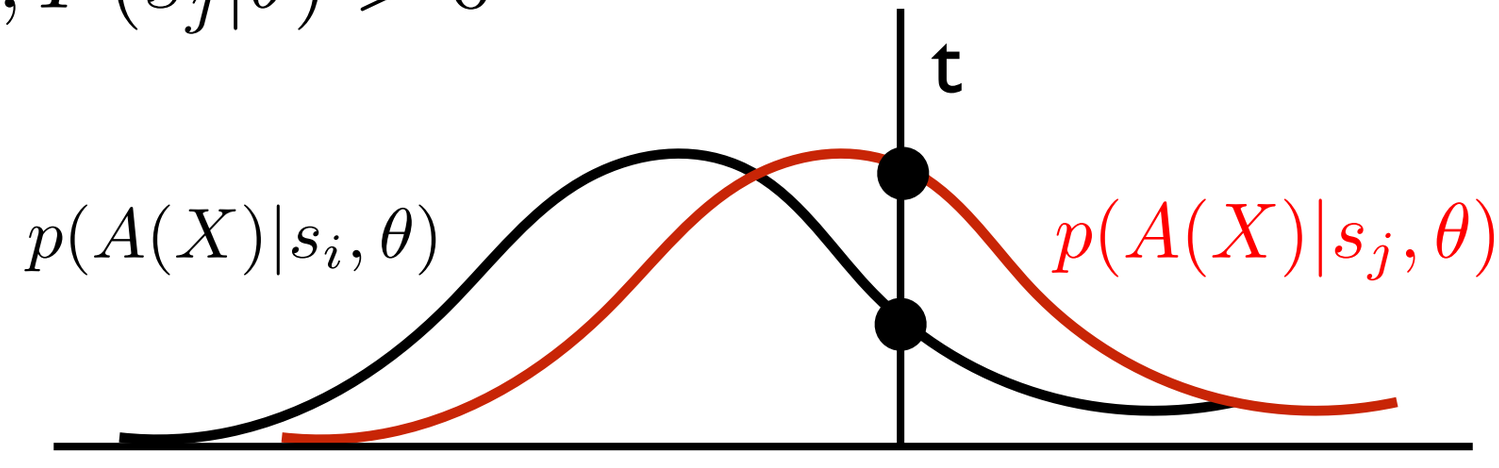
Pufferfish Privacy [KM12]

Secret Set S	Secret Pairs Set Q	Distribution Class Θ

An algorithm A is ϵ-Pufferfish private with parameters (S, Q, Θ) if for all (s_i, s_j) in Q, for all $\theta \in \Theta$, $X \sim \theta$, all t,

$$p_{\theta,A}(A(X) = t | s_i, \theta) \leq e^{\epsilon} \cdot p_{\theta,A}(A(X) = t | s_j, \theta)$$

whenever $P(s_i|\theta), P(s_j|\theta) > 0$



Pufferfish "Includes" DP [KM12]

Theorem: Pufferfish = Differential Privacy when:

S = { $s_{i,a}$:= Person i has value a, for all i, all a in domain X }

Q = { ($s_{i,a}$ $s_{i,b}$), for all i and (a, b) pairs in X x X }

Θ = { Distributions where each person i is independent }

Pufferfish "Includes" DP [KM12]

Theorem: Pufferfish = Differential Privacy when:

S = { $s_{i,a}$:= Person i has value a, for all i, all a in domain X }

Q = { ($s_{i,a}$ $s_{i,b}$), for all i and (a, b) pairs in X x X }

Θ = { Distributions where each person i is independent }

Theorem: No utility possible when:

Θ = { All possible distributions }

Talk Agenda:

1. Privacy for Correlated Data
    - How to define privacy (for uncorrelated data)
    - How to define privacy (for correlated data)
2. Privacy Mechanisms
    - A General Pufferfish Mechanism

How to get Pufferfish privacy?

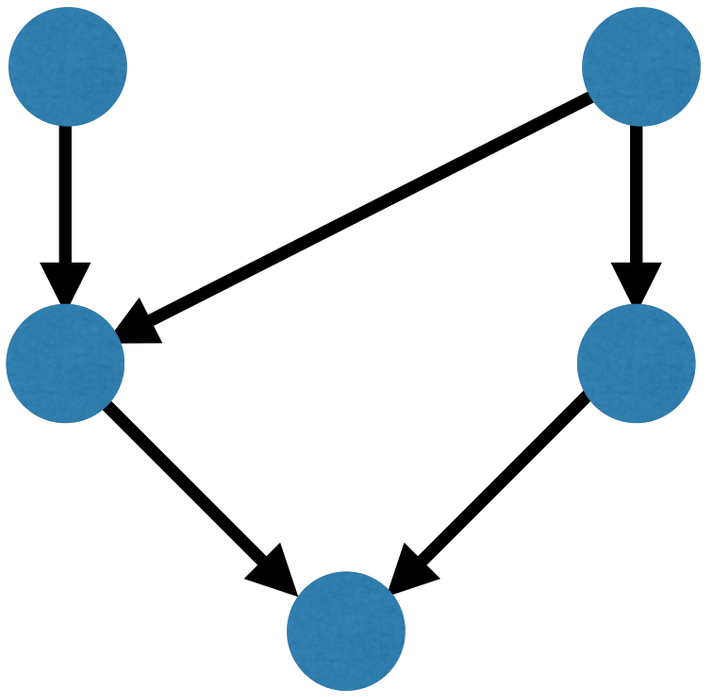
Special case mechanisms [KM12, HMD12]

Is there a **more general** Pufferfish mechanism for a large class of correlated data?

Our work: Yes, two - a. Wasserstein Mechanism
b. Markov Quilt Mechanism

(Also concurrent work [GK16])

Correlation Measure: Bayesian Networks



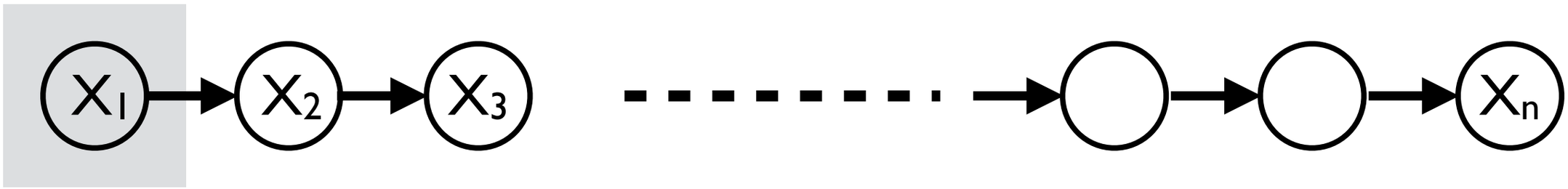
Node: variable

Directed Acyclic Graph

Joint distribution of variables:

$$\Pr(X_1, X_2, \ldots, X_n) = \prod_i \Pr(X_i | \text{parents}(X_i))$$

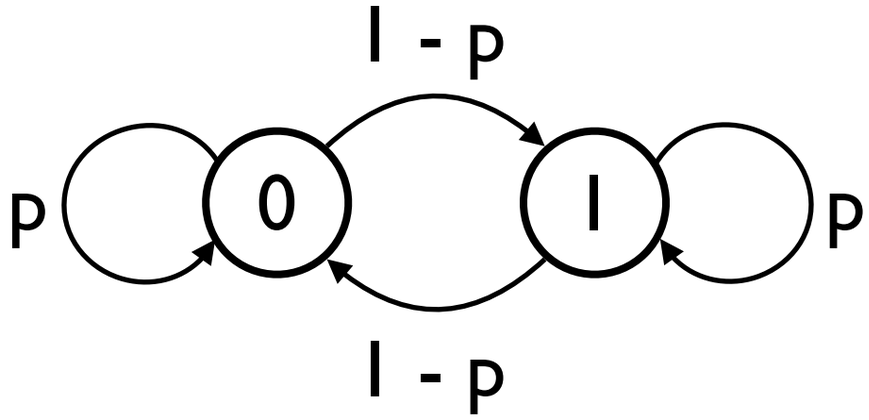
A Simple Example



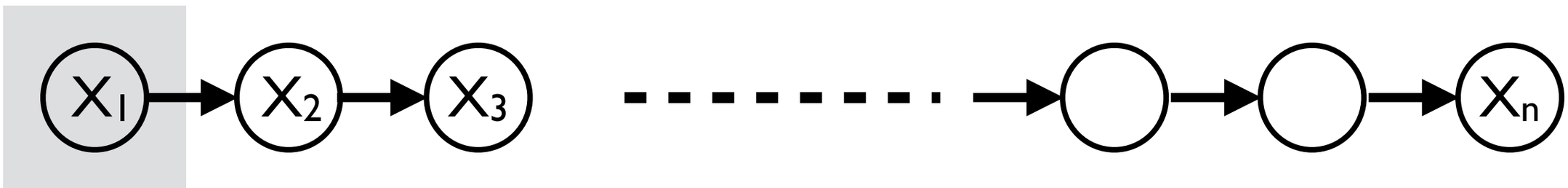
Model:

X_i in {0, 1}

State Transition Probabilities:



A Simple Example



Model:

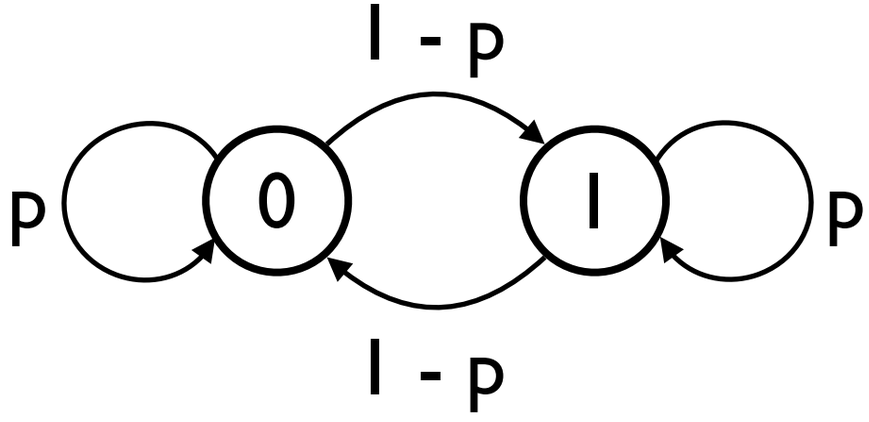
X_i in {0, 1}

$\Pr(X_2 = 0 | X_1 = 0) = p$

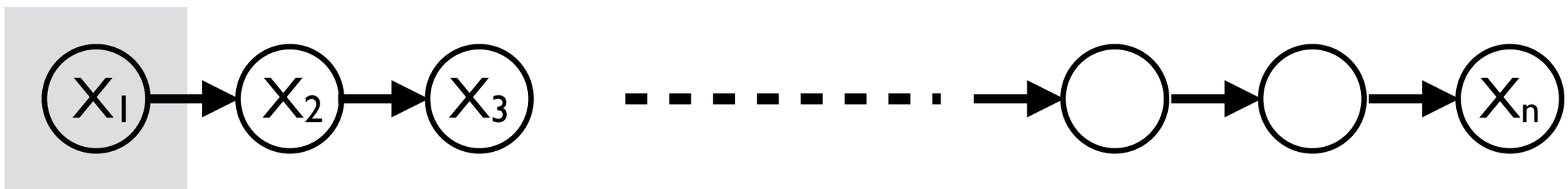
$\Pr(X_2 = 0 | X_1 = 1) = 1 - p$

….

State Transition Probabilities:



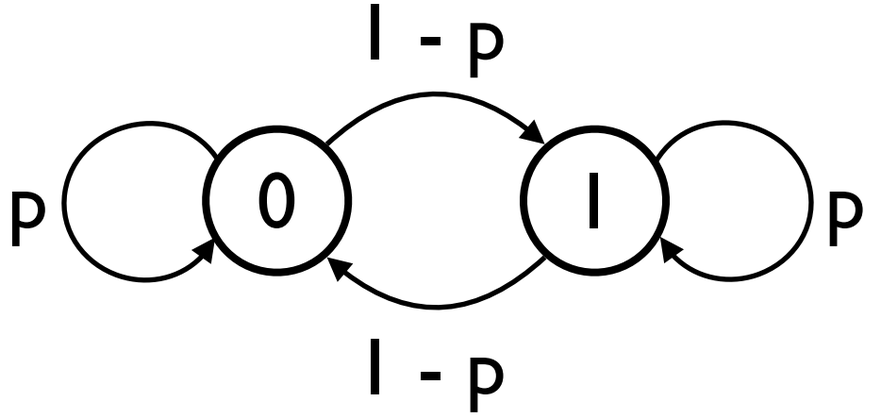
A Simple Example



Model:

X_i in {0, 1}

State Transition Probabilities:



$\Pr(X_2 = 0 | X_1 = 0) = p$

$\Pr(X_2 = 0 | X_1 = 1) = 1 - p$

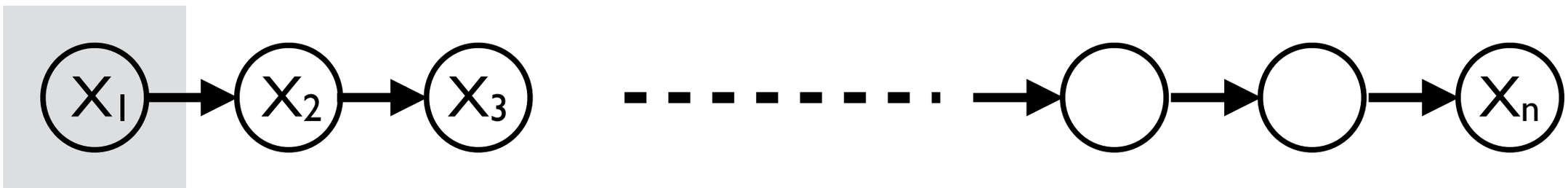
....

$\Pr(X_i = 0 | X_1 = 0) = \frac{1}{2} + \frac{1}{2}(2p-1)^{i-1}$

$\Pr(X_i = 0 | X_1 = 1) = \frac{1}{2} - \frac{1}{2}(2p-1)^{i-1}$

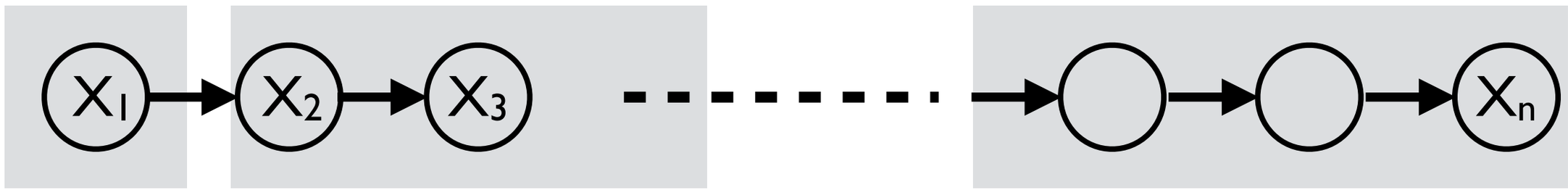
Influence of X_1 diminishes with distance

Algorithm: Main Idea



Goal: Protect X_1

Algorithm: Main Idea

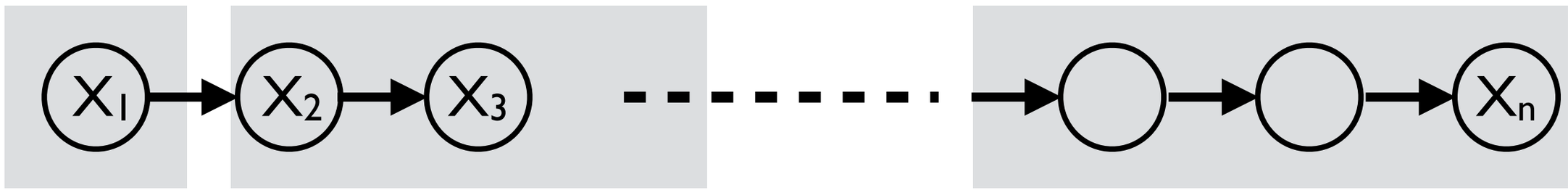


Local nodes (high correlation)

Rest (almost independent)

Goal: Protect X_1

Algorithm: Main Idea



Local nodes (high correlation)

Rest (almost independent)

Goal: Protect X_1

Add noise to hide local nodes **+** Small correction for rest

Measuring "Independence"

Max-influence of X_i on a set of nodes X_R:

$$e(X_R|X_i) = \max_{a,b} \sup_{\theta \in \Theta} \max_{x_R} \log \frac{\Pr(X_R = x_R | X_i = a, \theta)}{\Pr(X_R = x_R | X_i = b, \theta)}$$

Low $e(X_R|X_i)$ means X_R is almost independent of X_i

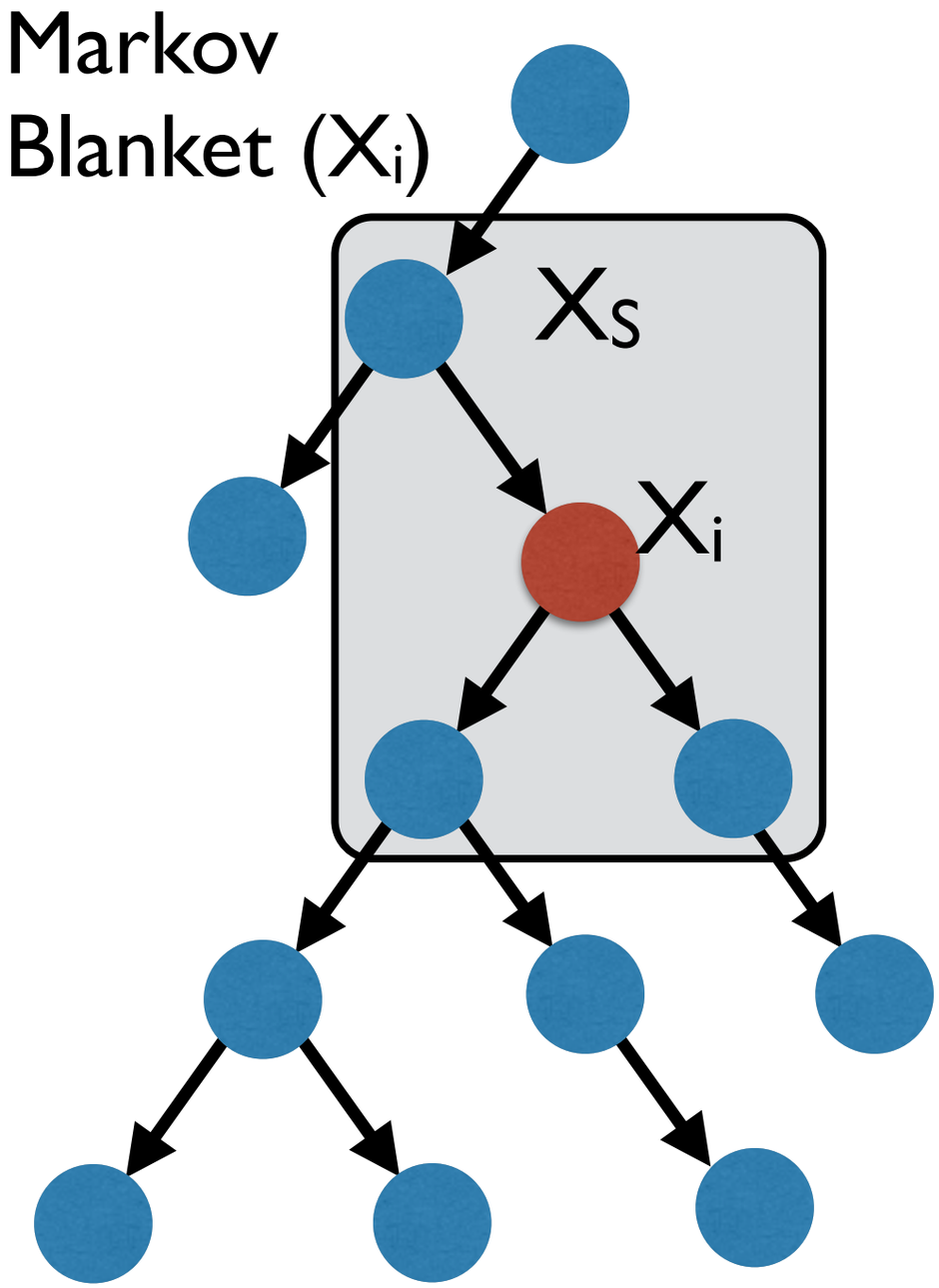
To protect X_i, correction term needed for X_R is $\exp(e(X_R|X_i))$

How to find large “almost independent” sets

Brute force search is expensive

Use structural properties of the Bayesian network

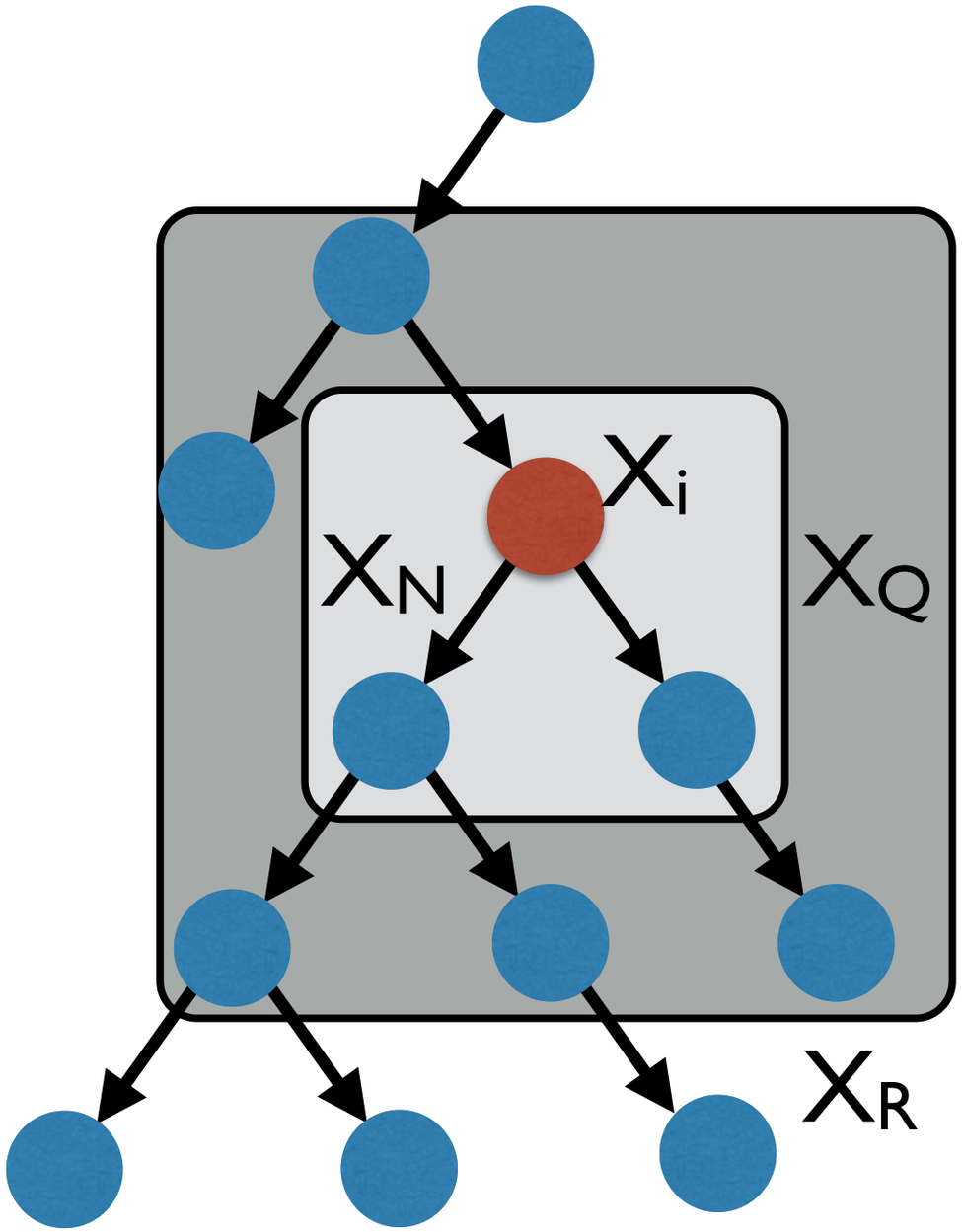
Markov Blanket



Markov Blanket(X_i) =
Set of nodes X_S s.t Xi is **independent of** $X \setminus (X_i \cup X_S)$ given X_S

(usually, parents, children, other parents of children)

Define: Markov Quilt

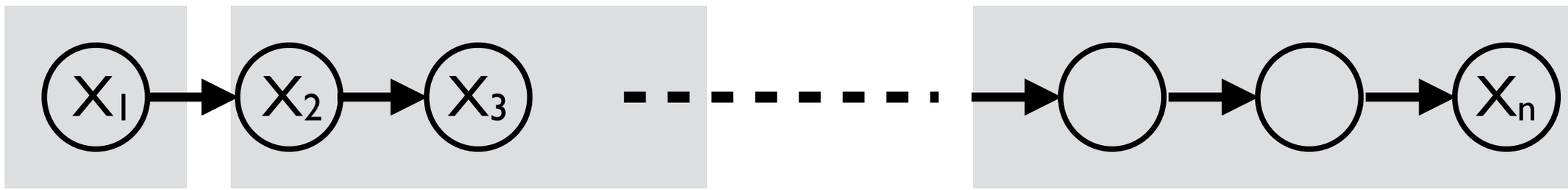


X_Q is a Markov Quilt of X_i if:

1. Deleting X_Q breaks graph into X_N and X_R
2. X_i lies in X_N
3. X_R is independent of X_i given X_Q

(For Markov Blanket $X_N = X_i$)

Recall: Algorithm



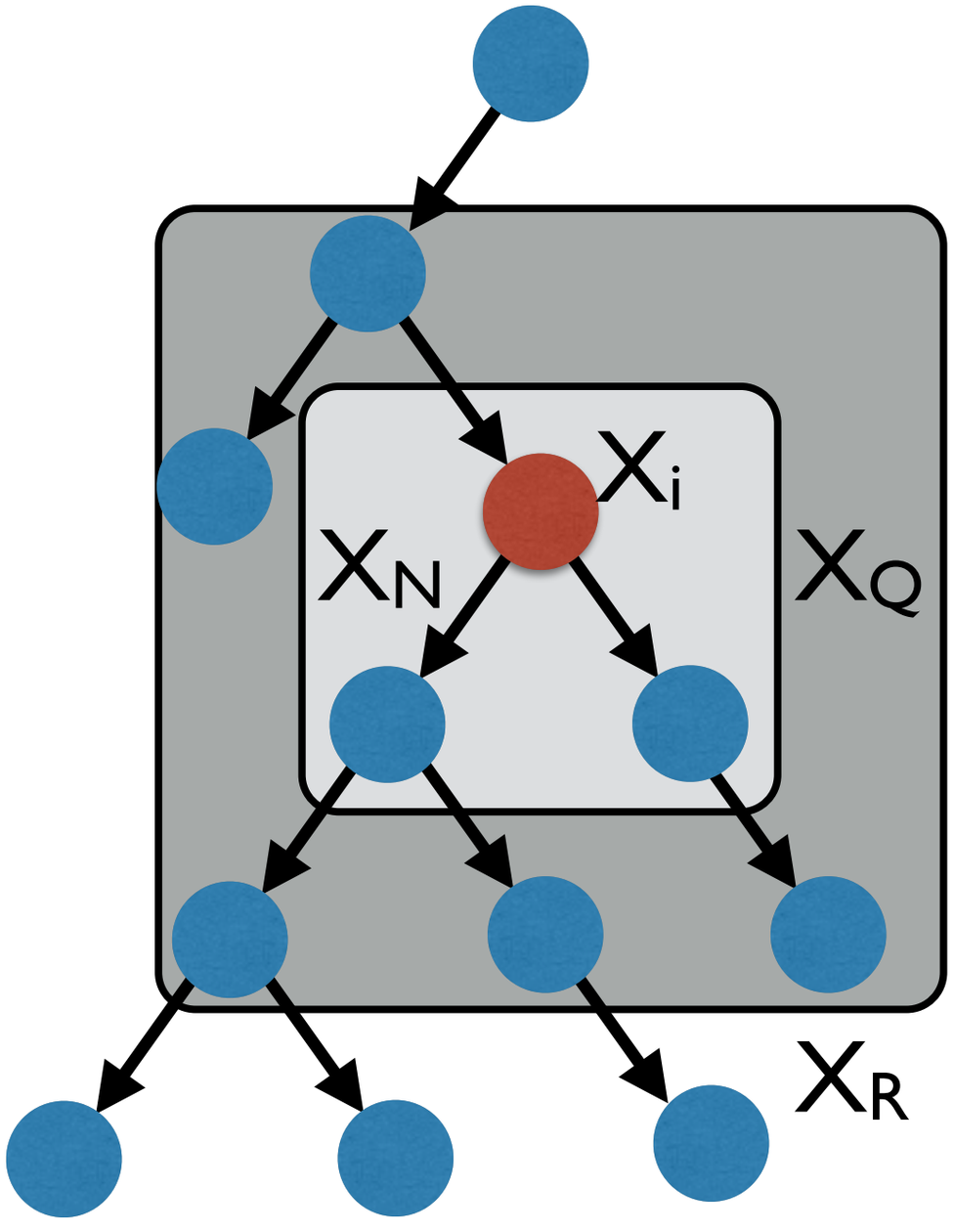
Local nodes (high correlation)

Rest (almost independent)

Goal: Protect X_1

Add noise to hide local nodes **+** Small correction for rest

Why do we need Markov Quilts?

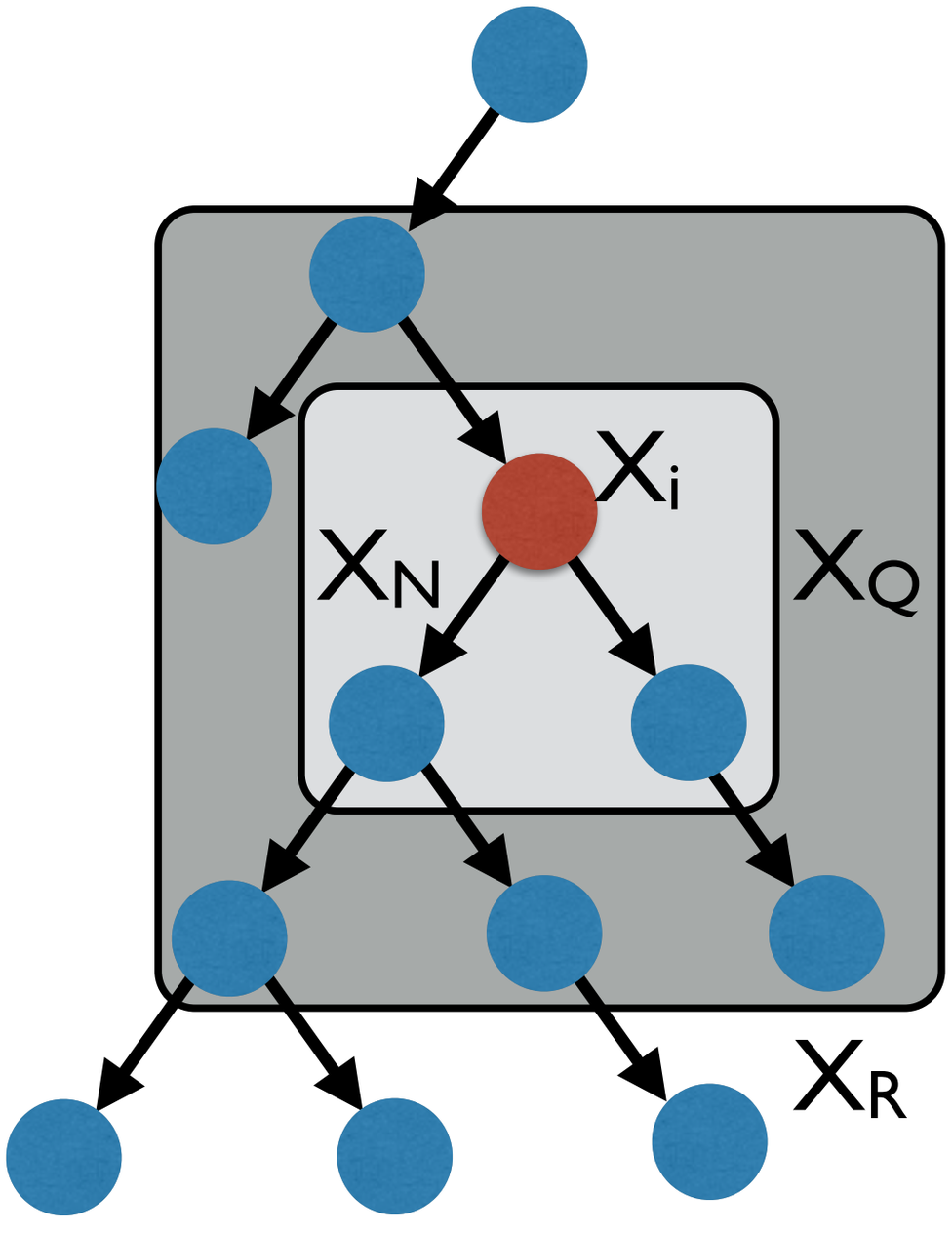


Given a Markov Quilt,

X_N = local nodes for X_i

$X_Q \cup X_R$ = rest

Why do we need Markov Quilts?



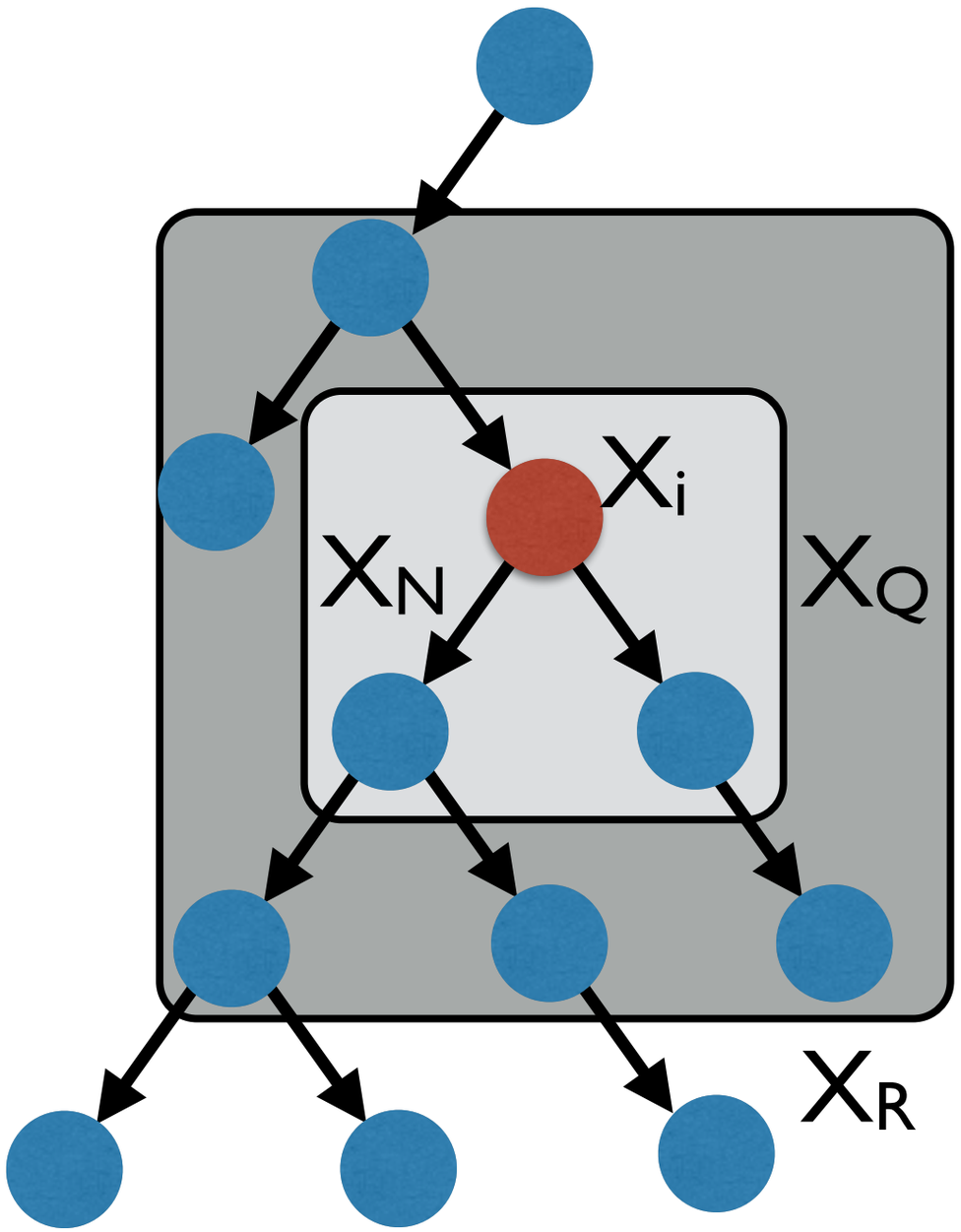
Given a Markov Quilt,

X_N = local nodes for X_i

X_Q U X_R = rest

Need to search over Markov Quilts X_Q to find the one which needs optimal amount of noise

From Markov Quilts to Amount of Noise



Let X_Q = Markov Quilt for X_i

Stdev of noise to protect X_i:

Noise due to X_N

$$\text{Score}(X_Q) = \frac{card(X_N)}{\epsilon - e(X_Q|X_i)}$$

Correction for $X_Q \cup X_R$

The Markov Quilt Mechanism

For each X_i

Find the Markov Quilt X_Q for X_i with minimum score s_i

Output F(D) + ($\max_i s_i$) Z where $Z \sim Lap(1)$

The Markov Quilt Mechanism

For each X_i

Find the Markov Quilt X_Q for X_i with minimum score s_i

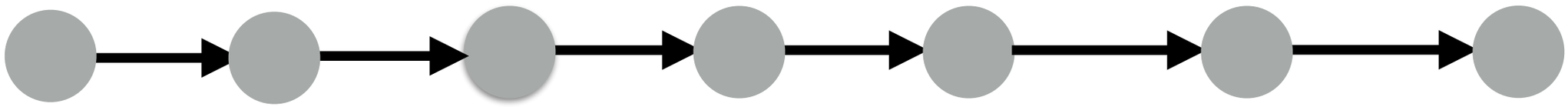
Output F(D) + ($\max_i s_i$) Z where $Z \sim Lap(1)$

Theorem: This preserves ϵ **-Pufferfish privacy**

Advantage: Poly-time in special cases.

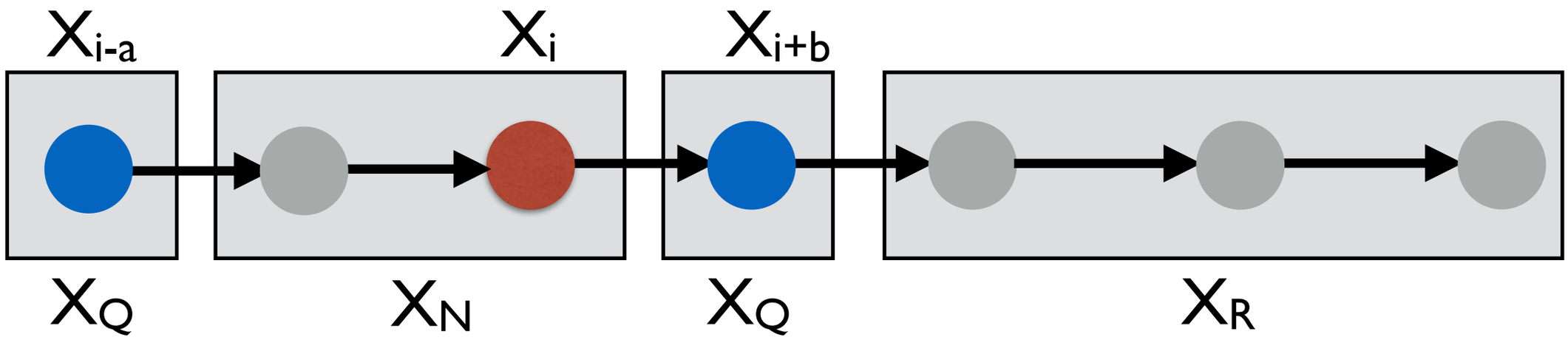
Example: Activity Monitoring

$D = (x_1, .., x_T)$, x_t = activity at time t



Example: Activity Monitoring

$D = (x_1, .., x_T)$, x_t = activity at time t



(Minimal) Markov Quilts for X_i have form $\{X_{i-a}, X_{i+b}\}$

Efficiently searchable

Example: Activity Monitoring

$\mathcal{X}$: set of states

P_θ : transition matrix describing each $\theta \in \Theta$

Example: Activity Monitoring

$\mathcal{X}$: set of states

P_θ : transition matrix describing each $\theta \in \Theta$

Under some assumptions, relevant parameters are:

$\pi_\Theta = \min_{x \in \mathcal{X}, \theta \in \Theta} \pi_\theta(x)$ (min prob of x under stationary distr.)

$g_\Theta = \min_{\theta \in \Theta} \min\{1 - |\lambda| : P_\theta x = \lambda x, \lambda < 1\}$ (min eigengap of any P_θ)

Example: Activity Monitoring

$\mathcal{X}$: set of states

P_θ : transition matrix describing each $\theta \in \Theta$

Under some assumptions, relevant parameters are:

$\pi_\Theta = \min_{x \in \mathcal{X}, \theta \in \Theta} \pi_\theta(x)$ (min prob of x under stationary distr.)

$g_\Theta = \min_{\theta \in \Theta} \min\{1 - |\lambda| : P_\theta x = \lambda x, \lambda < 1\}$ (min eigengap of any P_θ)

Max-influence of $X_Q = \{X_{i-a}, X_{i+b}\}$ for X_i

$$e(X_Q|X_i) \leq \log\left(\frac{\pi_\Theta + \exp(-g_\Theta b)}{\pi_\Theta - \exp(-g_\Theta b)}\right) + 2\log\left(\frac{\pi_\Theta + \exp(-g_\Theta a)}{\pi_\Theta - \exp(-g_\Theta a)}\right)$$

Score(X_Q) = $\dfrac{a+b-1}{\epsilon - e(X_Q|X_i)}$

Markov Quilt Mechanism for Activity Monitoring

For each X_i

Find Markov Quilt $X_Q = \{X_{i-a}, X_{i+b}\}$ with minimum score s_i

Output F(D) + ($\max_i s_i$) Z where $Z \sim Lap(1)$

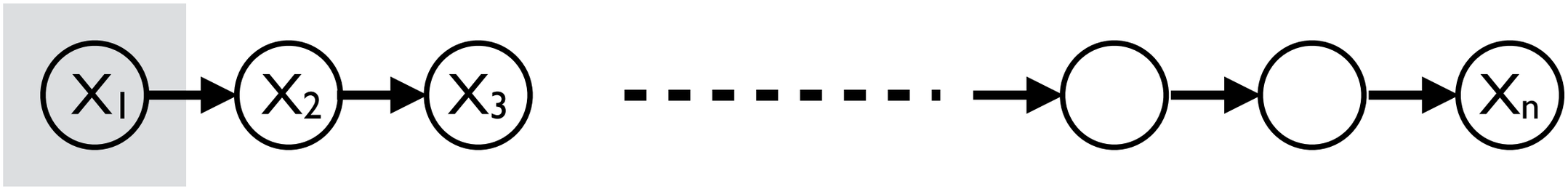
Running Time: $O(T^3)$ (can be made $O(T^2)$)

Advantage 1: Consistency

Advantage 2: Composition (for chains)

Experiments

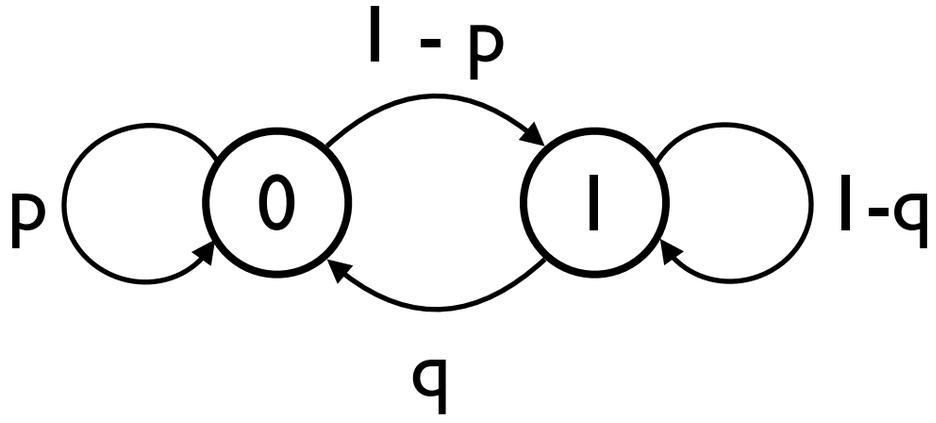
Simulations - Task



Model:

X_i in {0, 1}

State Transition Probabilities:



Model Class:

$$\Theta = [\ell, 1 - \ell]$$

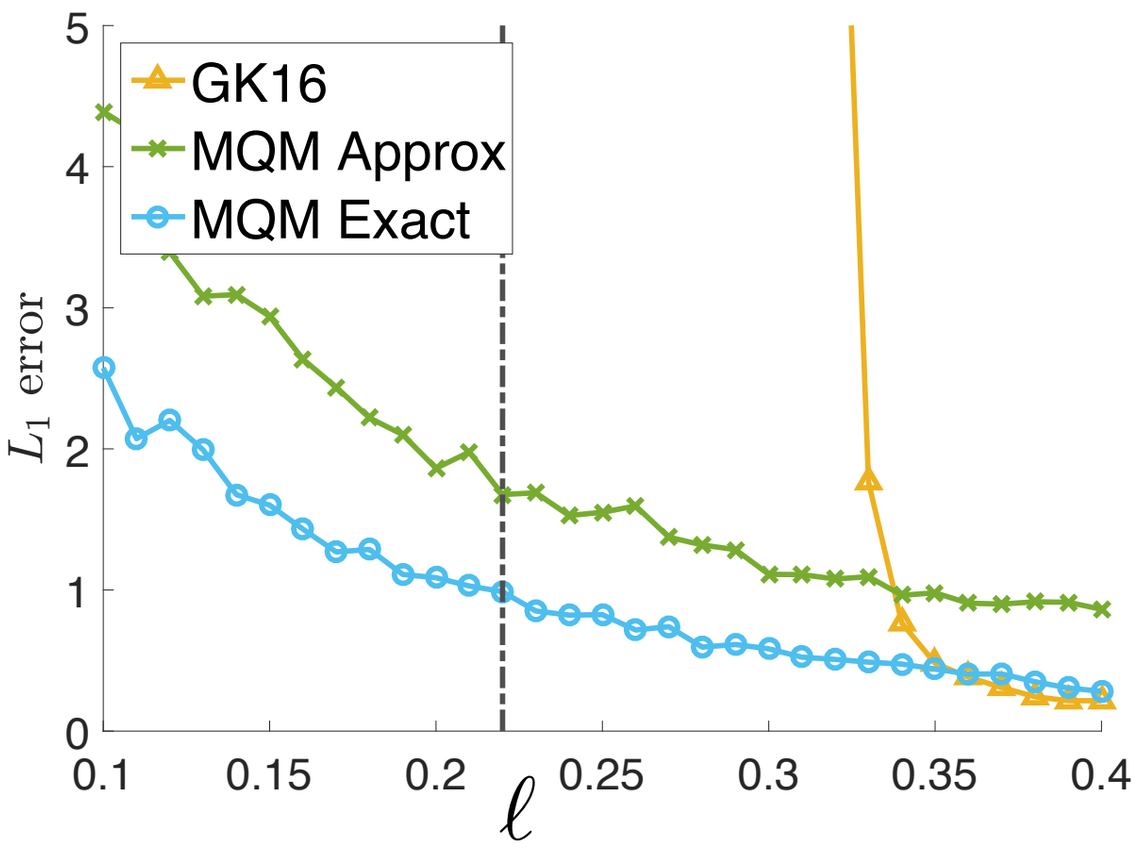
(implies p and q can lie anywhere in Θ)

Sequence length = 100

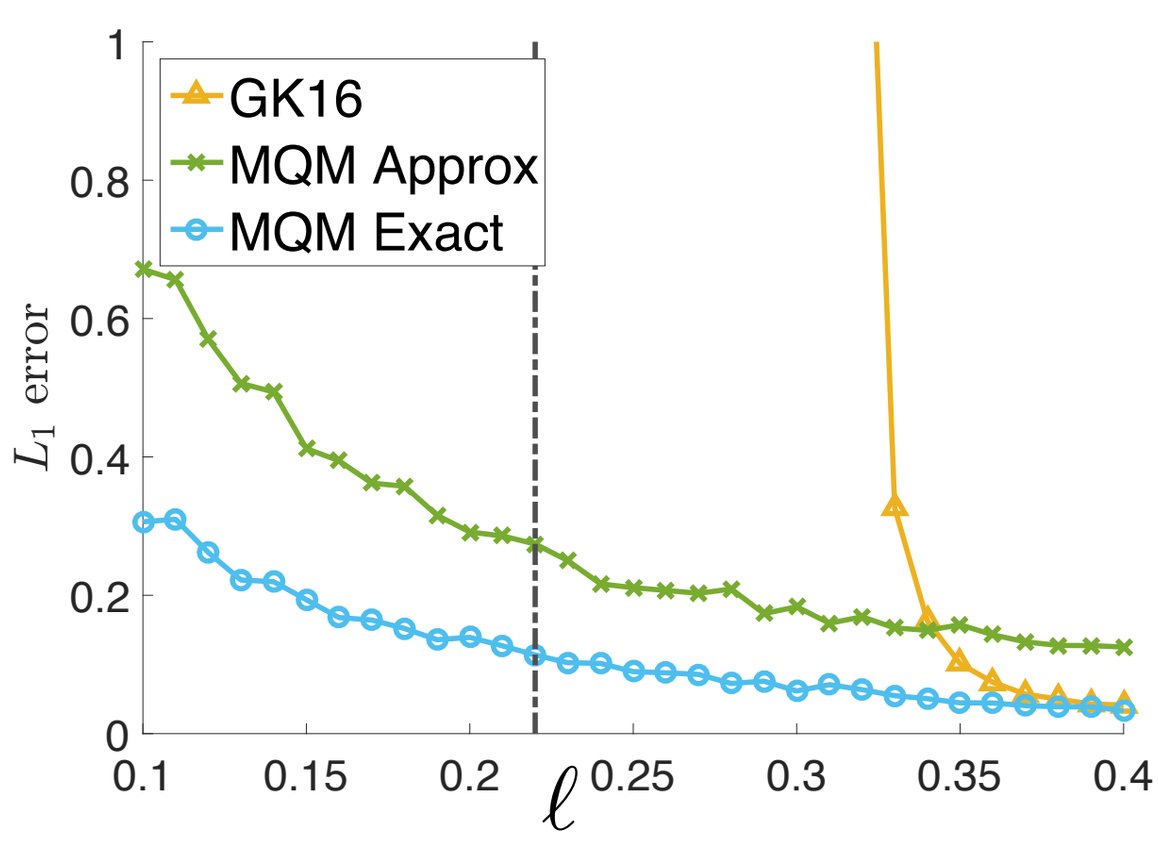
Simulations - Results

Methods:

- Two versions of Markov Quilt Mechanism (MQMExact, MQMApprox)
- GK16



epsilon=0.2



epsilon=1

Real Data - Activity Measurement

Dataset on physical activity by three groups of subjects:
40 cyclists, 16 older women and 36 overweight women

4 states (active, standing still, standing moving, sedentary)

Over 9,000 observations per subject

Θ = { Empirical data generating distribution }

Methods:

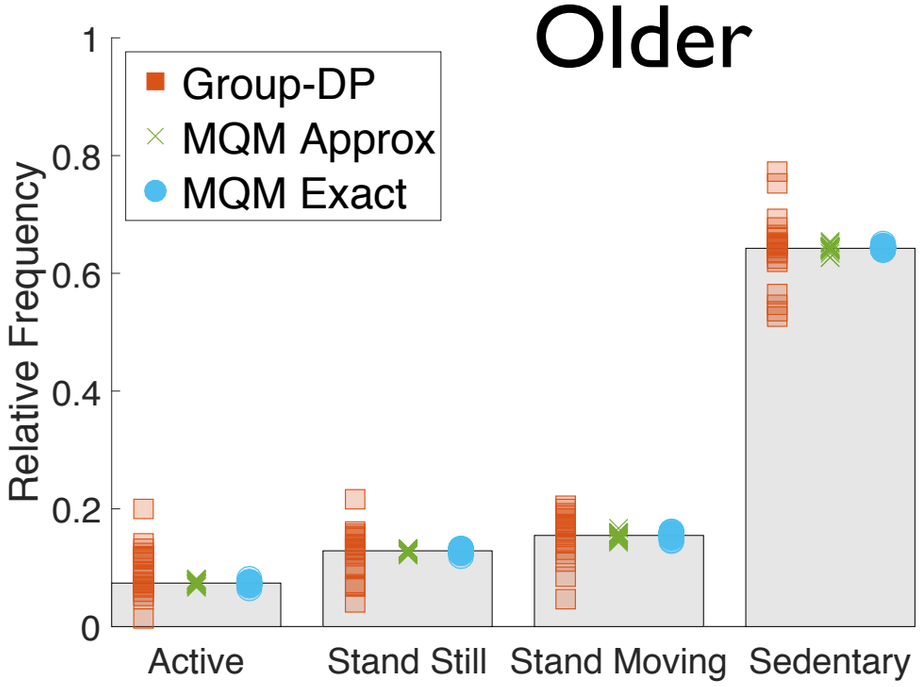
MQMExact and MQMApprox

GroupDP

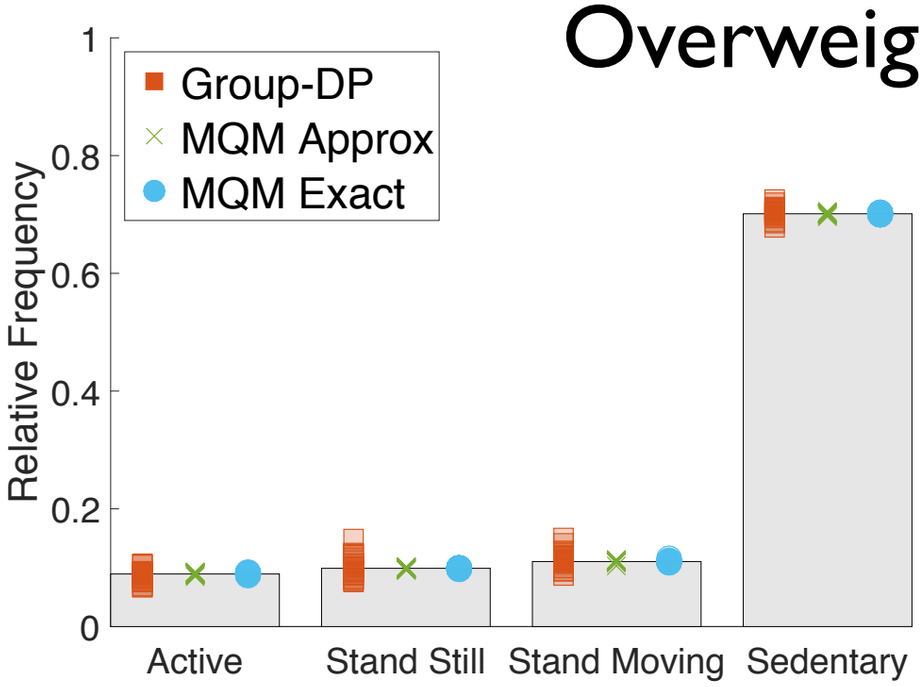
GK16 does not apply

Real Data - Activity Measurement

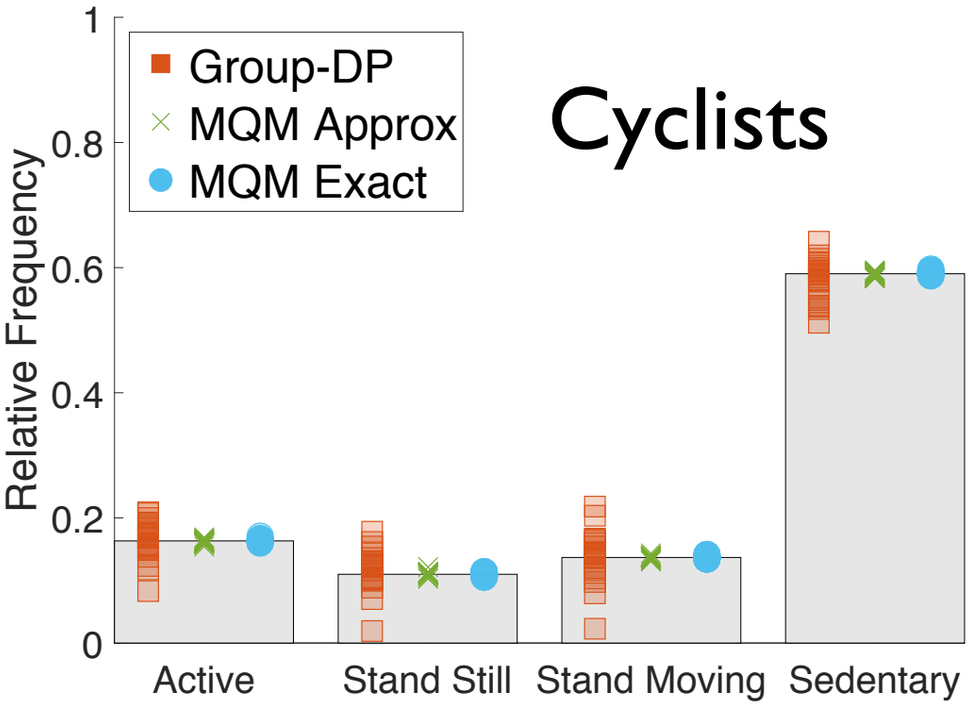
Older



Overweight



Cyclists



Aggregated results (over groups)

epsilon=1

Real Data - Power Consumption

Dataset on power consumption in a single household

Power consumption discretized to 51 levels (51 states)

Over 1 million observations

Θ = { Empirical data generating distribution }

Methods:

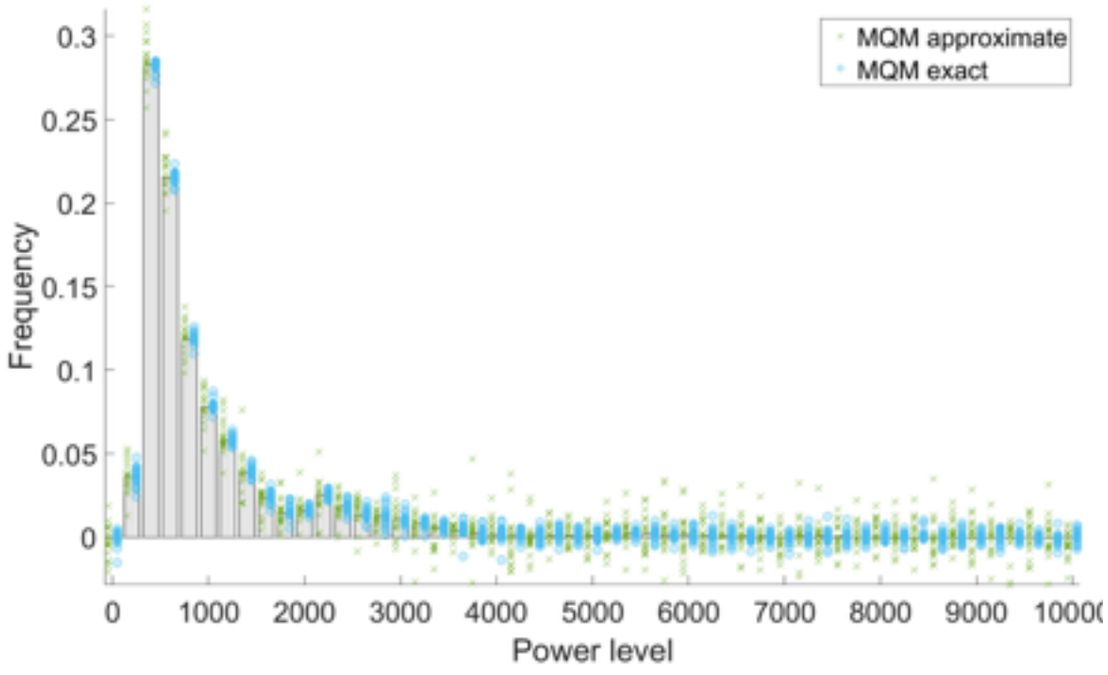
MQMExact vs. MQMApprox

GK16 does not apply

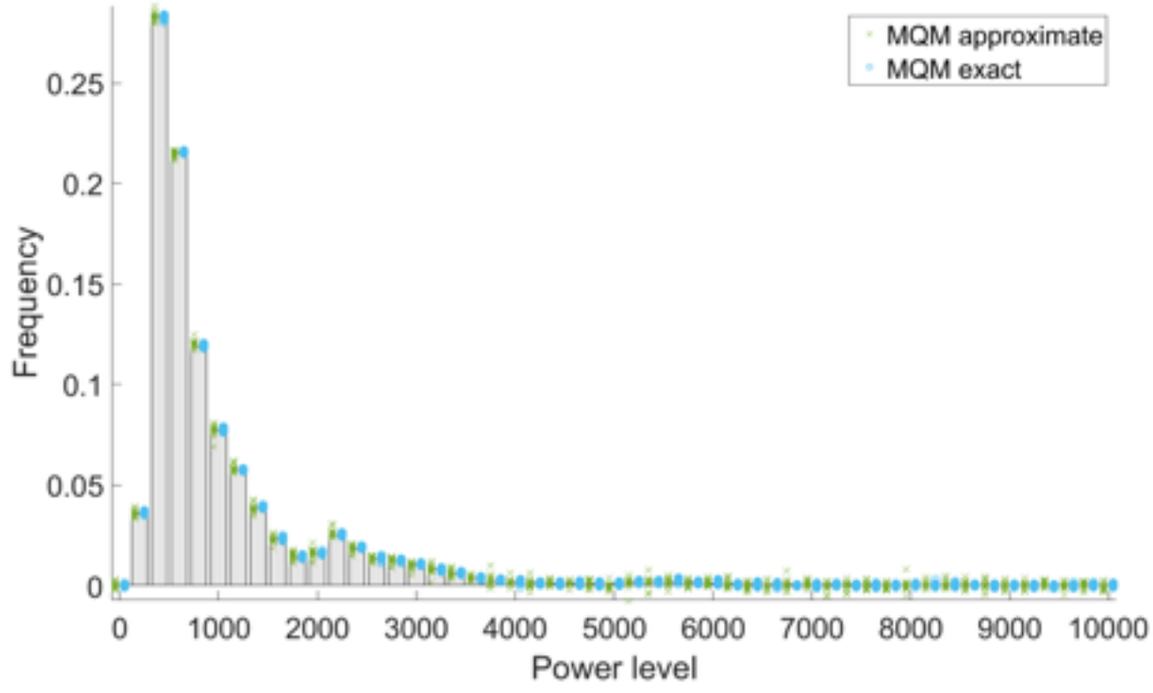
GroupDP has too little utility

Real Data - Power Consumption

Methods:
Two versions of Markov Quilt Mechanism (MQMExact, MQMApprox)



epsilon=0.2



epsilon=1

Conclusion

Problem:

privacy of correlated data - time series, social networks

Contributions:

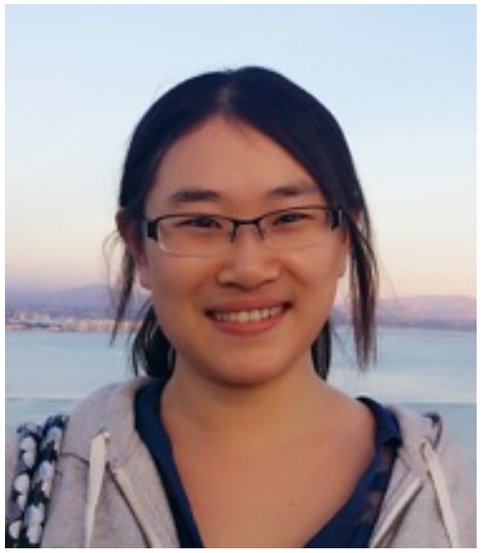
Two new mechanisms - a fully general mechanism, and a more efficient mechanism

Established composition of the Markov Quilt Mechanism

Future Work:

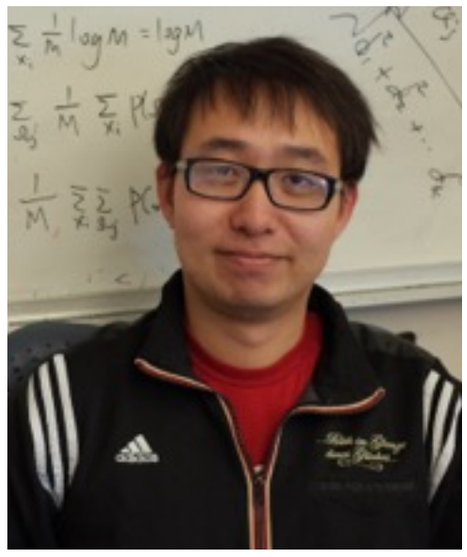
More efficient mechanisms, more detailed composition properties

Acknowledgements



Shuang Song

Mani Srivastava



Yizhen Wang

Questions?