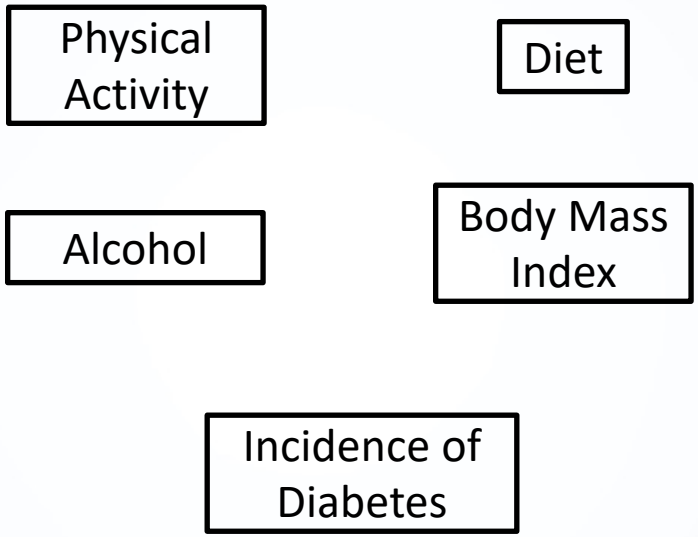
# Collaborative Causal Discovery with Atomic Interventions

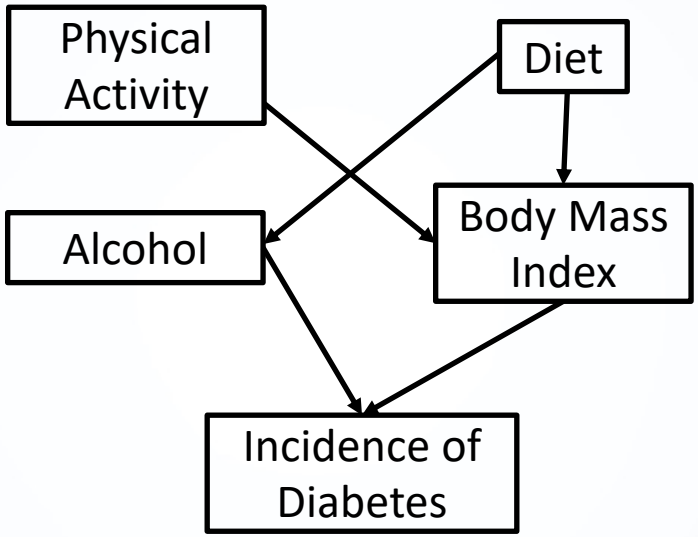
Shiva Kasiviswanathan
Amazon

Joint work with Raghavendra Addanki (UMass Amherst)

# Causal Discovery



# Causal Discovery

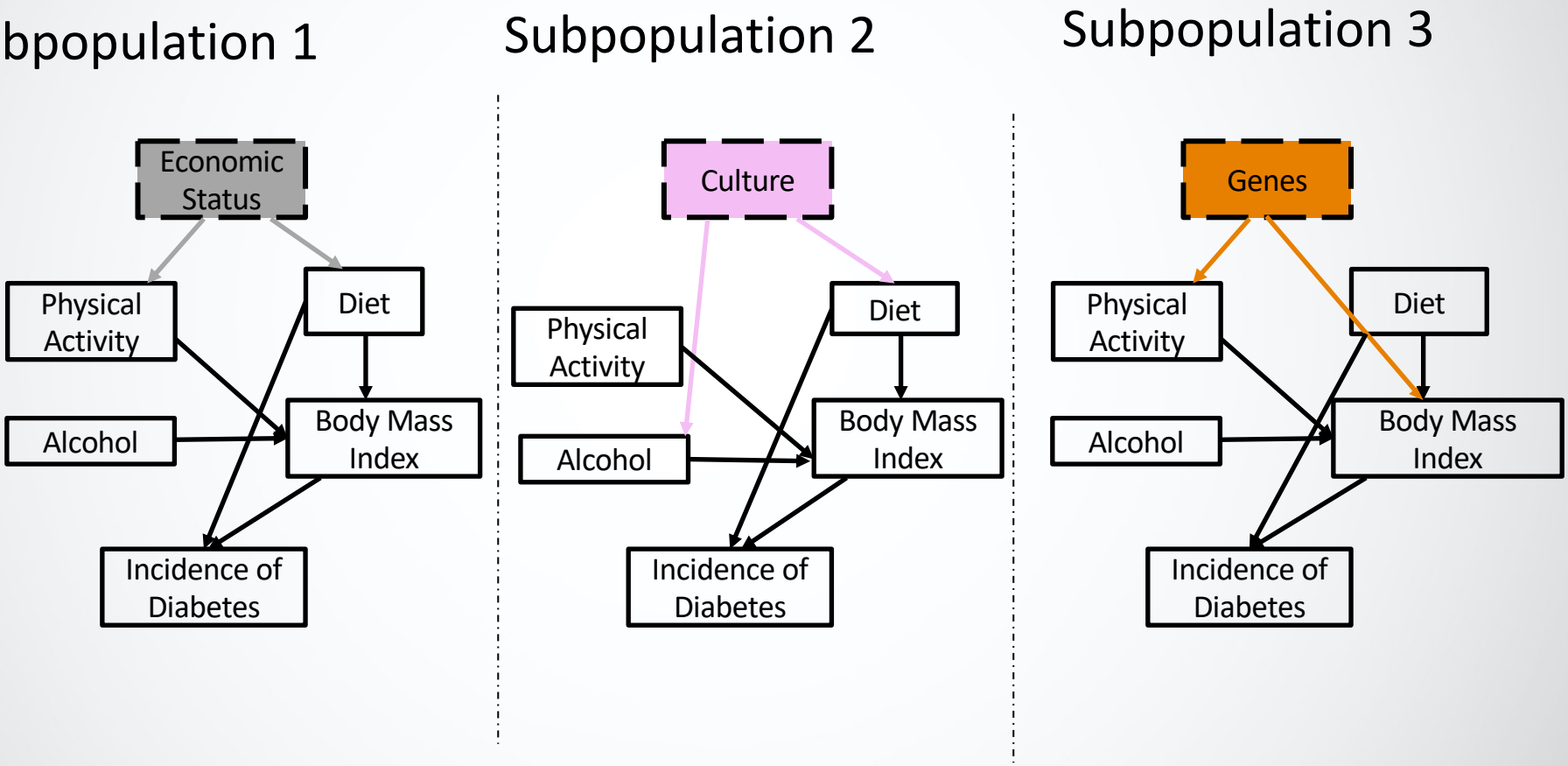



Causal Discovery: Learn the graph that describes the ``causal relationship" between these variables

But is there a **unique** causal graph?

In many scenarios, the answer is no

# An Example*

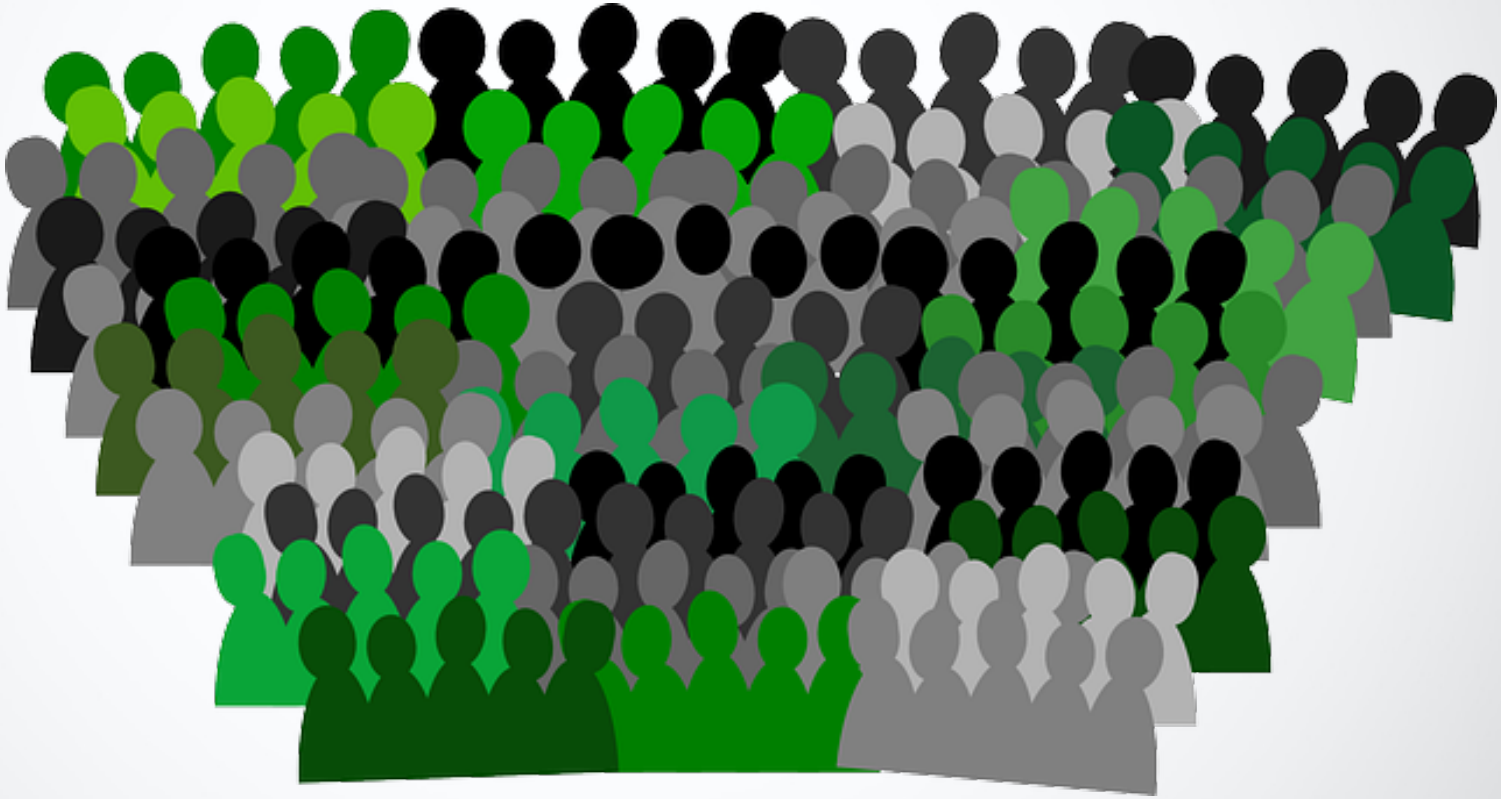



* Motivated from: Michael Joffe, Manoj Gambhir, Marc Chadeau-Hyam, and Paolo Vineis. Causal diagrams in systems epidemiology, 2012.


Goal: To learn all these causal graphs

# Say we have $N$ entities

Entities could be individuals (or collection of individuals)

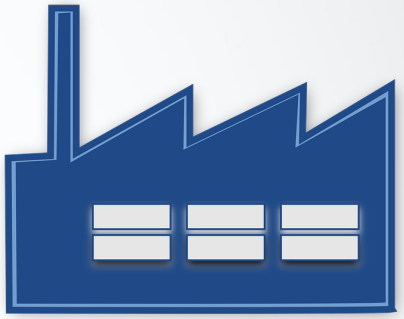


Each with their own causal graph

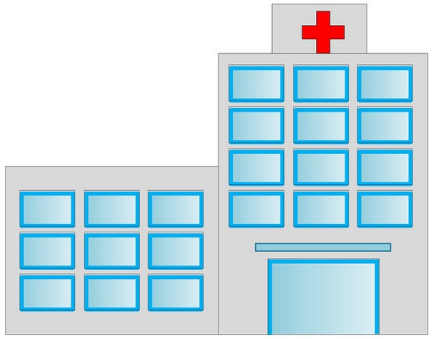
# Say we have $N$ entities

Entities could be businesses

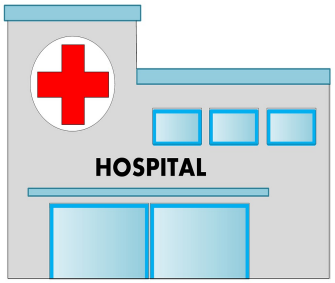
.....



Entities could be medical agencies



.....




Each with their own causal graph

**Goal:** To learn the causal graphs of all the entities, assuming access to their individual data

Assumptions on entities:

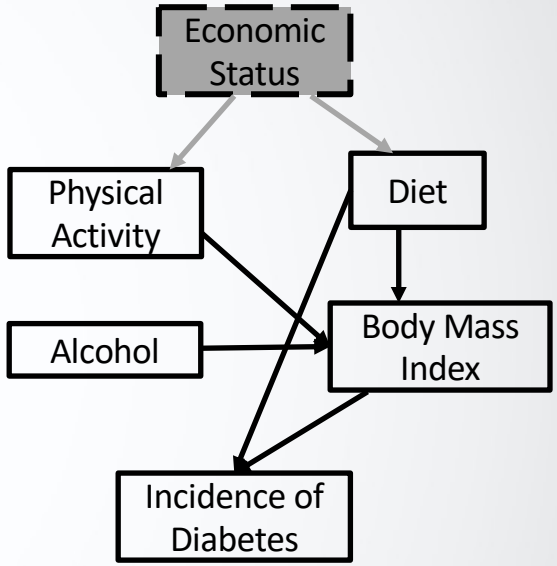
1) They are independent
2) The set of observed variables is the same for all the entities
3) Each of them generates their own data

# Detour: How do we learn a Causal Graph?

- Observational data is not sufficient to learn the exact causal relations
- We cannot distinguish graphs in a Markov equivalence class
- To distinguish in the equivalence class, we need additional mechanism such as ability to perform "interventions"

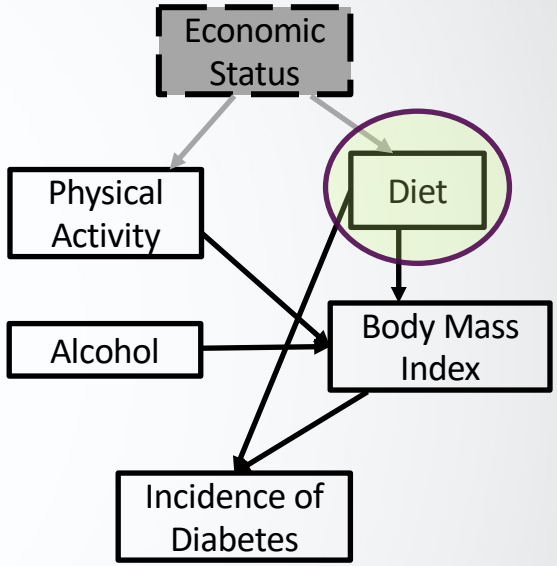
# Detour: How do we learn a Causal Graph?

- Observational data is not sufficient to learn the exact causal relations
- We cannot distinguish graphs in a Markov equivalence class
- To distinguish in the equivalence class, we need additional mechanism such as ability to perform "interventions"



# Detour: How do we learn a Causal Graph?

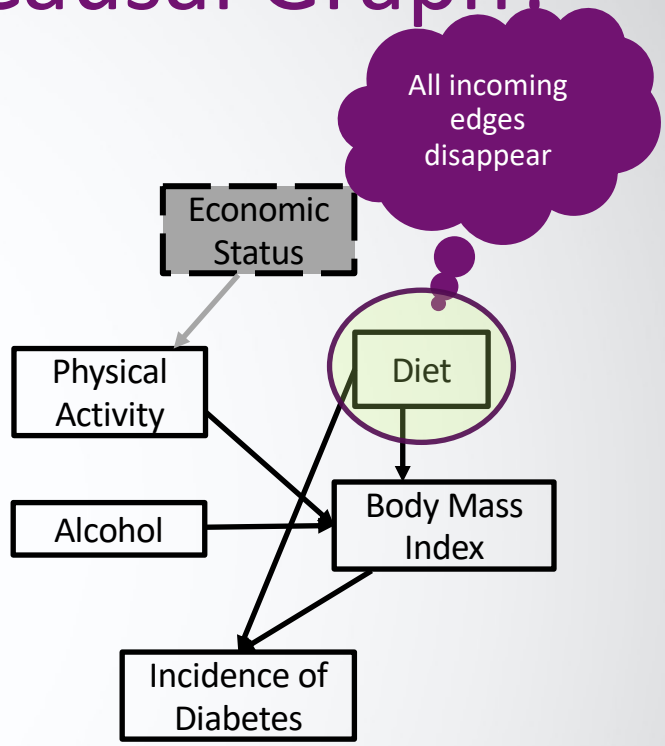
- Observational data is not sufficient to learn the exact causal relations
- We cannot distinguish graphs in a Markov equivalence class
- To distinguish in the equivalence class, we need additional mechanism such as ability to perform "interventions"




E.g.: An intervention on "*Diet*", involves fixing the *Diet* level of the individual to some value (either low or high).

# Detour: How do we learn a Causal Graph?

- Observational data is not sufficient to learn the exact causal relations
- We cannot distinguish graphs in a Markov equivalence class
- To distinguish in the equivalence class, we need additional mechanism such as ability to perform "interventions"




E.g.: An intervention on "*Diet*", involves fixing the *Diet* level of the individual to some value (either low or high).

# Detour: How do we learn a Causal Graph?

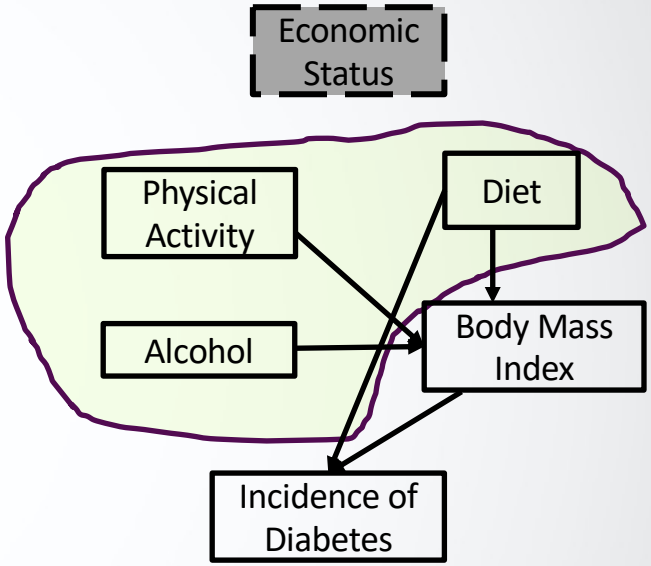
- Much of the recent work on causal discovery with minimum interventions uses “large” intervention sets.

- Moreover, such large interventions are *necessary* [AKMM 2020].

- In practice, however large intervention sets are infeasible.

# Detour: How do we learn a Causal Graph?

- Much of the recent work on causal discovery with minimum interventions uses "large" intervention sets.
- Moreover, such large interventions are *necessary* [A**K**MM 2020].
- In practice, however large intervention sets are infeasible.




E.g.: An **simultaneous** intervention on "*Diet*", "*Physical Activity*" and "*Alcohol*" might be hard for the individual.

Here we only focus on atomic interventions

Here we only focus on atomic interventions

Learning the exact causal DAG not possible

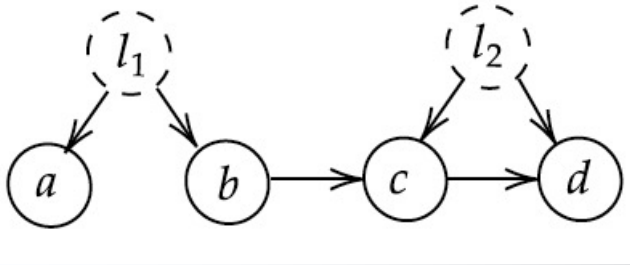
# Here we only focus on atomic interventions

Learning the exact causal DAG not possible

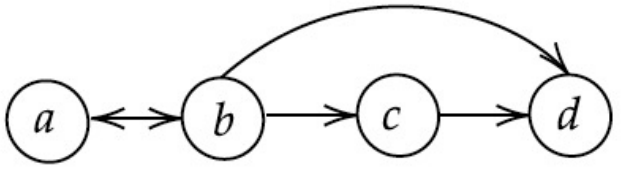
So what else can we learn?

# Maximal Ancestral Graphs (MAGs)

- MAGs are a type of causal graphs which encode higher order causal relations (such as “paths”)
- MAGs encode the presence of latents using bidirected edges
- We show that MAGs can be recovered under atomic interventions




DAG $\mathcal{D}$




MAG corresponding to the DAG $\mathcal{D}$

# Meaning of the Edges of a MAG

Directed edges $a \to b$ means:

*i.* $a$ is an ancestor of $b$.

*ii.* $b$ is not an ancestor of $a$.

iii. This does not rule out possible latent confounding between $a$ and $b$.

Bidirected edges as $a \leftrightarrow b$ means:

*i.* $a$ is not an ancestor of $b$.

*ii.* $b$ is not an ancestor of $a$.

*iii.* $a$ and $b$ are confounded.

Every DAG with latents can be transformed into an unique MAG over the observed variables [Richardson and Sprites, 2002]

Without latents (i.e., Causal Sufficiency)

MAG = DAG

# Maximal Ancestral Graphs (MAGs)

Long line of work investigating MAGs*:

- introduced by Richardson and Sprites in 2002
- at most one edge between each pair of vertices
- closed under marginalization and conditioning
- causal reasoning with MAGs well-understood [Zhang 2008]
- Markov equivalence class is represented by Partial Ancestral Graph (PAG)
- PAGs can be recovered from observational data

  e.g., FCI algorithm [Sprites,Glymour,Scheines 2000] and variants

*Lots of other great results are skipped

## Revised Goal: To learn the **MAGs** of all the entities, assuming access to their individual data

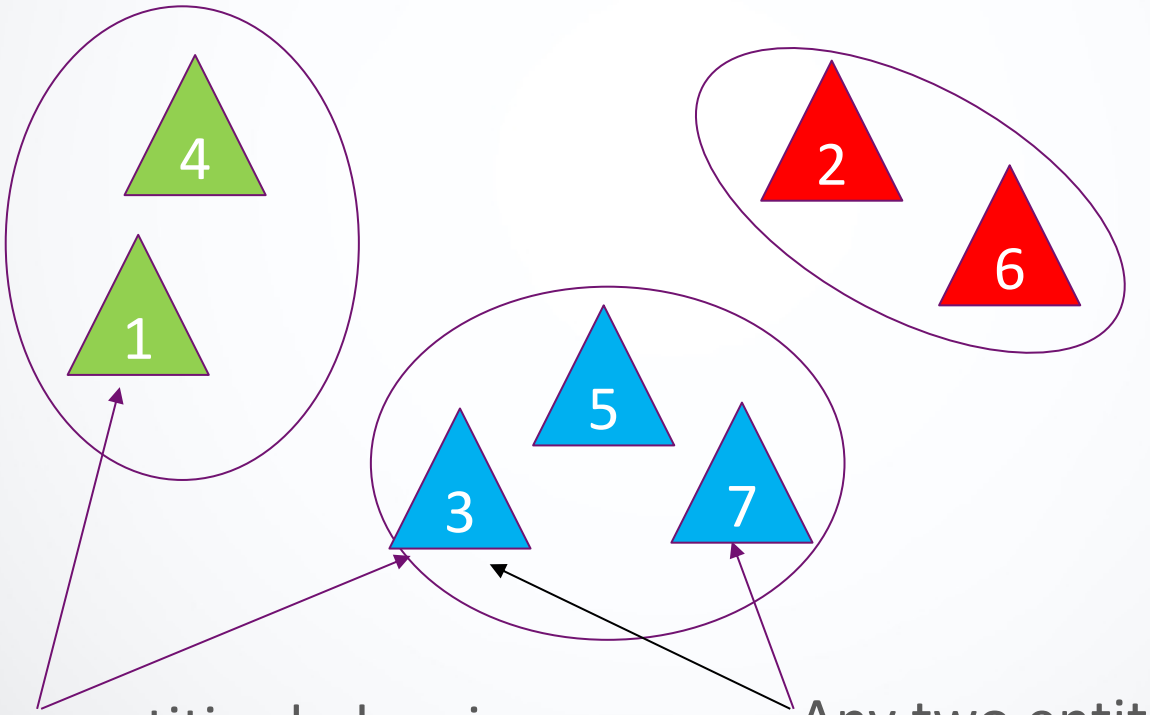
Allow atomic interventional access on each entity

Objective is to minimize the maximum number of atomic interventions per entity


Collaboration
"Working together is success"

We assume (some **unknown**) underlying clustering of entities based on their MAGs

Here # entities $N = 7$

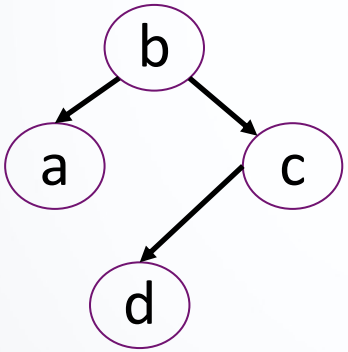



Any two entities belonging to different clusters have ”far apart” MAGs

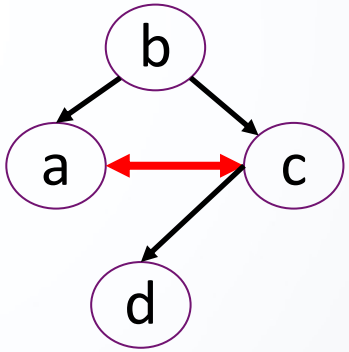
Any two entities within the same cluster have ”close by” MAGs

# Model

- Let $\mathcal{M}_i$ denote the MAG associated with entity $i \in [N]$
- Let $n$ denote the number of observables (same for all entities)
- Distance $d(\mathcal{M}_i, \mathcal{M}_j)$ between two MAGs $\mathcal{M}_i, \mathcal{M}_j$ is the number of nodes with different neighborhoods




$\mathcal{M}_i$




$\mathcal{M}_j$

$\mathcal{M}_i, \mathcal{M}_j$ differ in the neighborhoods on $\{a, c\}$

$$\Longrightarrow d(\mathcal{M}_i, \mathcal{M}_j) = 2$$

# Model

- Let $\mathcal{M}_i$ denote the MAG associated with entity $i \in [N]$
- Let $n$ denote the number of observables (same for all entities)
- Distance $d(\mathcal{M}_i, \mathcal{M}_j)$ between two MAGs $\mathcal{M}_i, \mathcal{M}_j$ is the number of nodes with different neighborhoods

Defn. $(\alpha, \beta)$-clustering of the entities:

$\mathcal{M}_1, \mathcal{M}_2, \ldots \mathcal{M}_N$ denote MAGs associated with the entities and belonging to clusters $C_1^*, C_2^*, \ldots, C_k^*$. The entities satisfy $(\alpha, \beta)$-clustering property if $\forall\, i, j \in [N]$:

- Entities $i, j$ in same cluster then, $d(\mathcal{M}_i, \mathcal{M}_j) \leq \beta n$
- Entities $i, j$ in different cluster then, $d(\mathcal{M}_i, \mathcal{M}_j) \geq \alpha n$ (with $\beta < \alpha$)

Defn. $(\alpha, \beta)$-clustering of the entities:

$\mathcal{M}_1, \mathcal{M}_2, \ldots \mathcal{M}_N$ denote MAGs associated with the entities and belonging to clusters $C_1^*, C_2^*, \ldots, C_k^*$. The entities satisfy $(\alpha, \beta)$-clustering property if $\forall\, i, j \in [N]$:

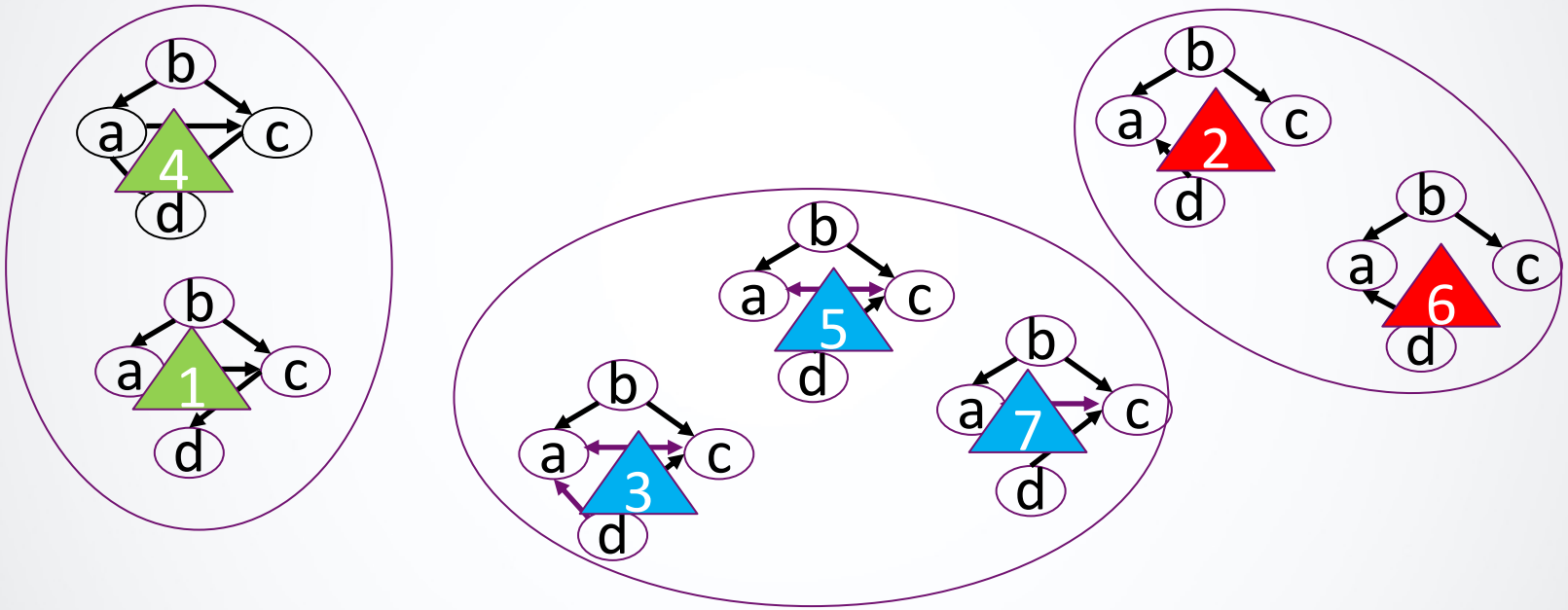
- Entities $i, j$ in same cluster then, $d(\mathcal{M}_i, \mathcal{M}_j) \leq \beta n$
- Entities $i, j$ in different cluster then, $d(\mathcal{M}_i, \mathcal{M}_j) \geq \alpha n$ (with $\beta < \alpha$ )

Quick Points:

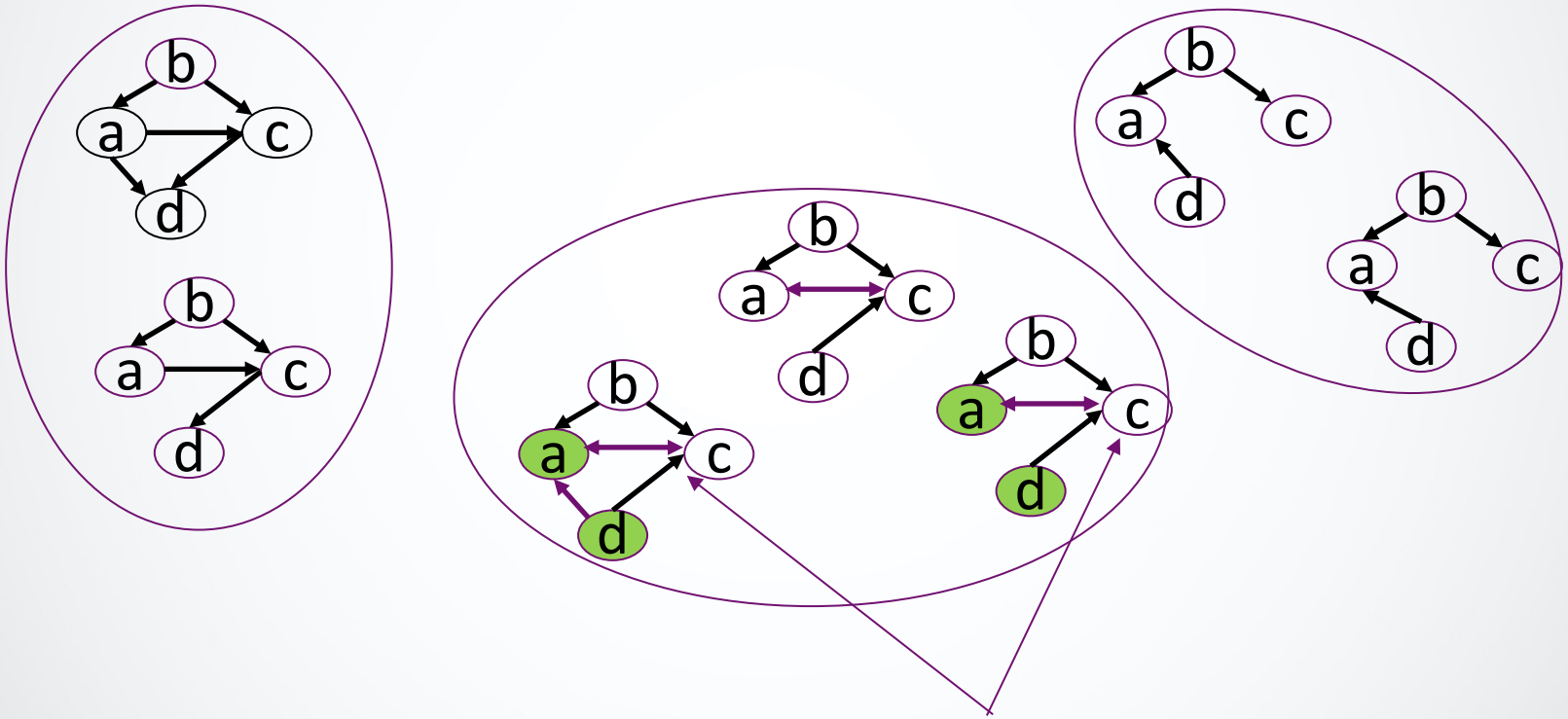
- No restriction on $k$
- Any set of $N$ MAGs will be captured by this definition as

i.e., we can put each entity in a different cluster

# MAGs with $(\alpha = 0.75, \beta = 0.5)$ clustering

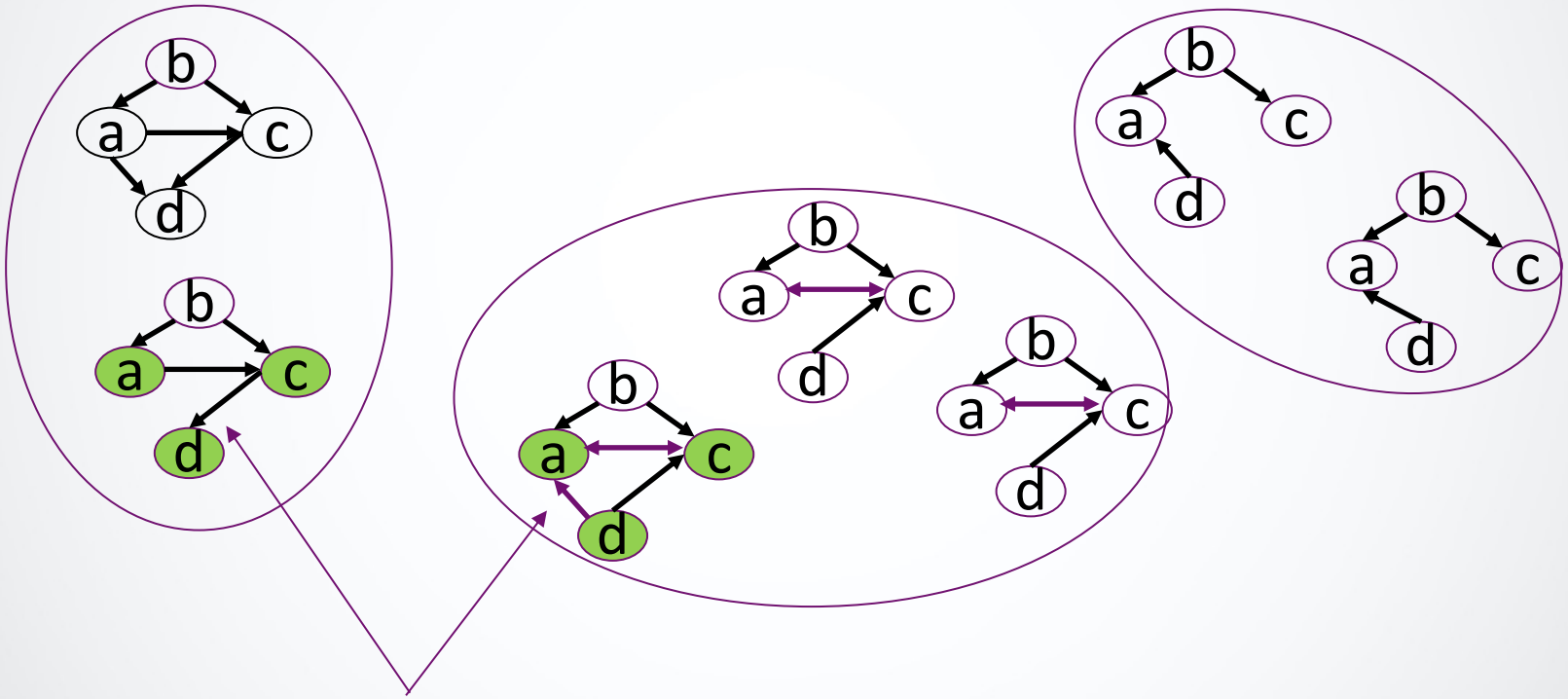


# MAGs with $(\alpha = 0.75, \beta = 0.5)$ clustering




Take these two, $\beta = \frac{2}{4} = 0.5$

# MAGs with $(\alpha = 0.75, \beta = 0.5)$ clustering




Take these two, $\alpha = \frac{3}{4} = 0.75$

# Our Theoretical Guarantees

- Let $\Delta$ be the maximum degree among all the MAGs.
- Assume $\min_{\{i \in [k]\}} |C_k^*| \geq n$ (min true cluster size)

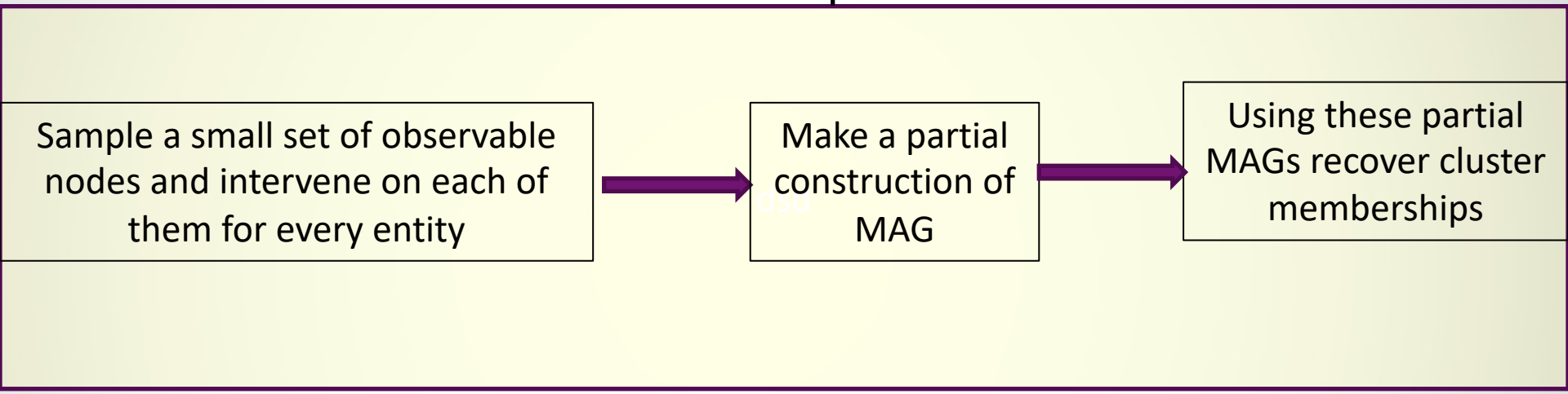
| **Assumption** | Learning each MAG Independently | Our Collaborative Algorithms | Lower Bound |
|---|---|---|---|
| $(\alpha, \beta)$ | $\Theta(n)$ | $O(\Delta/(\alpha - \beta)^2 \log N)$** | $\Omega(1/\alpha)$ |
| $(\alpha, 0)$ | $\Theta(n)$ | $\min\{O(\frac{1}{\alpha}\log N + k^2), O(\frac{\Delta}{\alpha}\log N)\}$ | $\Omega(1/\alpha)$ |

In the $(\alpha, 0)$ case each cluster represents the same MAG

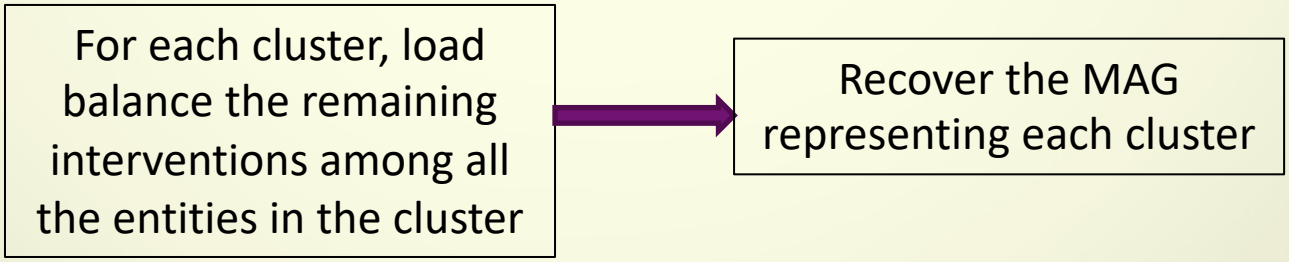
**: Requires additional assumptions

# Overview of Our Approach

## Phase 1: Recover Cluster Memberships




## Phase 2: Clusters to MAGs



# Phase 1: Learning the Clustering

Undirected skeleton (more generally Partial Ancestral Graph) is easy to obtain

1) Let $V$ denote the set of observable nodes
2) Sample $S$, a set of $O\left(\frac{\log N}{\alpha}\right)$ nodes from $V$ at random
3) For every entity $i \in [N]$
   a) Intervene on the nodes in $S$ and their (undirected) neighbors
   b) Construct a partial MAG for the entity $i$

   Idea: if an undirected edge $(a, b)$ exists for the entity $i$:
   i. If $a \not\perp\!\!\!\perp b \mid do(a)$ then $a \to b$
   ii. If $a \perp\!\!\!\perp b \mid do(a)$ and $a \perp\!\!\!\perp b \mid do(b)$ then $a \leftrightarrow$ b

4) Group entities with same partial MAG into the same cluster

Thm: Under $(\alpha, 0)$ clustering assumption, with high probability, we can recover the exact cluster membership for all the entities with at most $O(\frac{\Delta \log N}{\alpha})$ atomic interventions per entity.

1) Can remove the dependence on maximum degree $\Delta$
2) Idea also extends to the more general $(\alpha, \beta)$-clustering

# A Word about the Lower Bound

We show that $\Omega(1/\alpha)$ interventions per entity are needed

We construct a distribution $\mu$ over MAGs and show that every (deterministic) algorithm requires requires $\Omega(1/\alpha)$ interventions for distinguishing a pair of MAGs drawn from $\mu$.

Yao's minimax theorem [Yao 1977]

Worst case lower bound for any randomized algorithm

# Experimental Evaluation

Recovering cluster membership under $(\alpha, 0)$-clustering

| Causal Network | Our Algorithm (Accuracy) | # interventions per entity |
|---|---|---|
| *Earthquake* (n=5) | 100% | 3 |
| *Survey* (n=6) | 89% | 4 |
| *Asia* (n=8) | 89% | 4 |
| *Sachs* (n=11) | 79% | 5 |
| *Erdős–Rényi* (n=10) | 100% | 5 |

# Conclusion

- We introduce a new causal discovery model to capture practical scenarios involving multiple causal graphs.
- Under natural clustering assumption(s), we showed that the graphs can be learnt with far fewer interventions

## Open Directions

- Develop non-adaptive algorithms that can be run in parallel
- Using interventional equivalence classes [Kocaoglu et al. 2019, Jaber et al. 2020].

Thanks for your attention!