

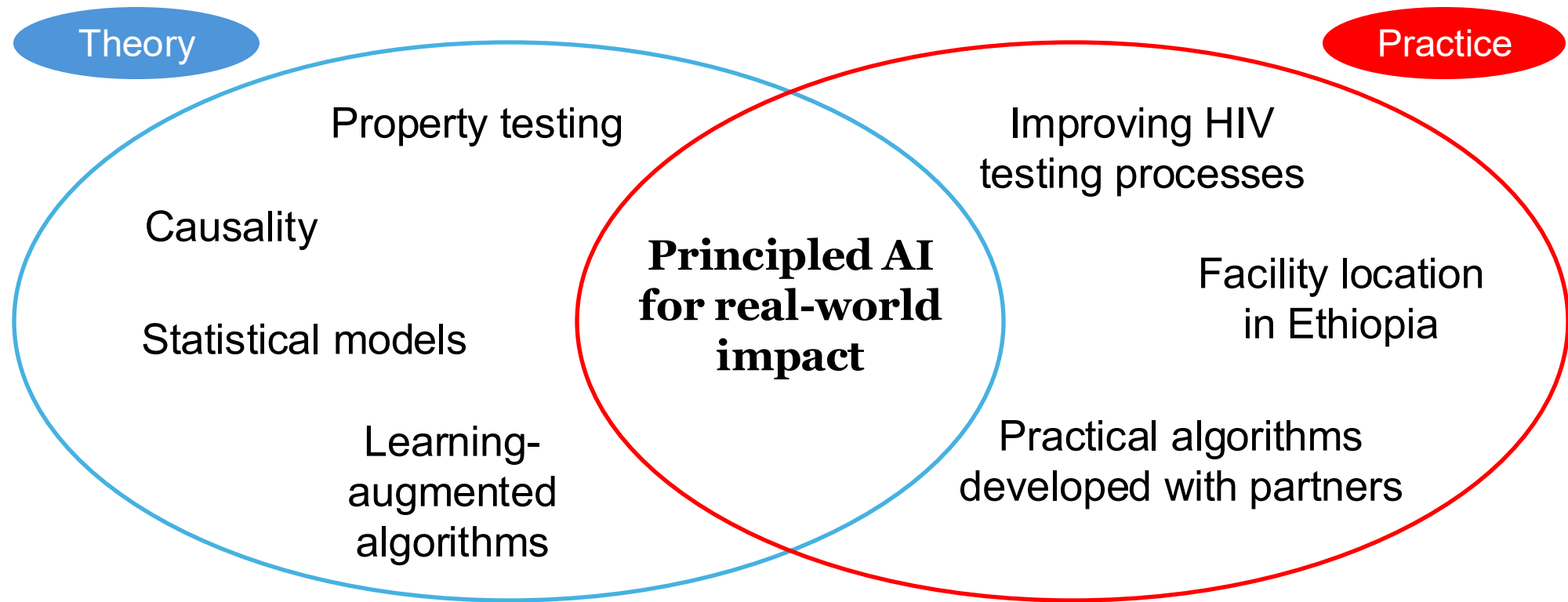
Towards theory-guided, impact-driven AI

Davin Choo

Postdoctoral Fellow @ Harvard SEAS

A bit about myself

>20 papers in top-tier AI/ML venues such as NeurIPS, ICML, AAAI, AISTATS, UAI



PhD @ NUS
Masters @ ETH Zürich

Postdoc @ Harvard

Research vision, in 3 research thrusts

RT1: Foundational AI / ML Research

RT2: Algorithms with Imperfect Advice

RT3: AI + X



Research vision, in 3 research thrusts

RT1: Foundational AI / ML Research

- Many fundamental real-world problems involve partial, noisy, or structure-rich settings
- *How do we develop statistical and computational efficient methods to tackle these challenges?*

RT2: Algorithms with Imperfect Advice

RT3: AI + X



Research vision, in 3 research thrusts

RT1: Foundational AI / ML Research

- Many fundamental real-world problems involve partial, noisy, or structure-rich settings
- *How do we develop statistical and computational efficient methods to tackle these challenges?*

RT2: Algorithms with Imperfect Advice

- Many real-world instances are often not worst-case. We should adapt accordingly.
- *How can algorithms benefit from imperfect instance-specific advice, without blindly trusting it?*

RT3: AI + X



Research vision, in 3 research thrusts

RT1: Foundational AI / ML Research

- Many fundamental real-world problems involve partial, noisy, or structure-rich settings
- *How do we develop statistical and computational efficient methods to tackle these challenges?*

RT2: Algorithms with Imperfect Advice

- Many real-world instances are often not worst-case. We should adapt accordingly.
- *How can algorithms benefit from imperfect instance-specific advice, without blindly trusting it?*

RT3: AI + X

- Greatest potential of AI methods lies in domains outside of classic CS/Eng/OR/Stats settings
- *How can we improve upon existing processes after principled and useful formulations?*





**Some of my
past and
current
efforts**

RT1: Foundational AI / ML Research

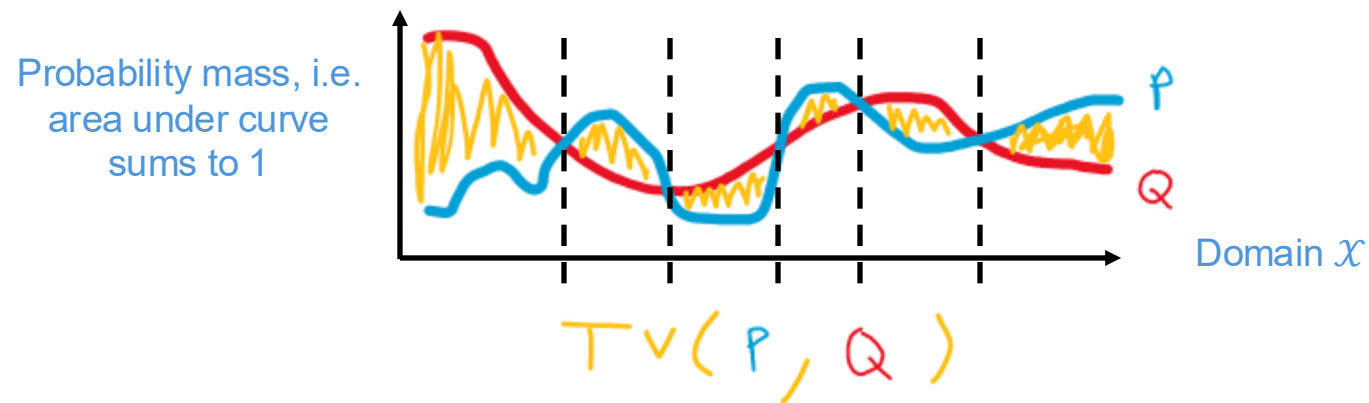
RT2: Algorithms with Imperfect Advice

RT3: AI + X

RT1: Foundational AI / ML Research

Background on some statistical problems

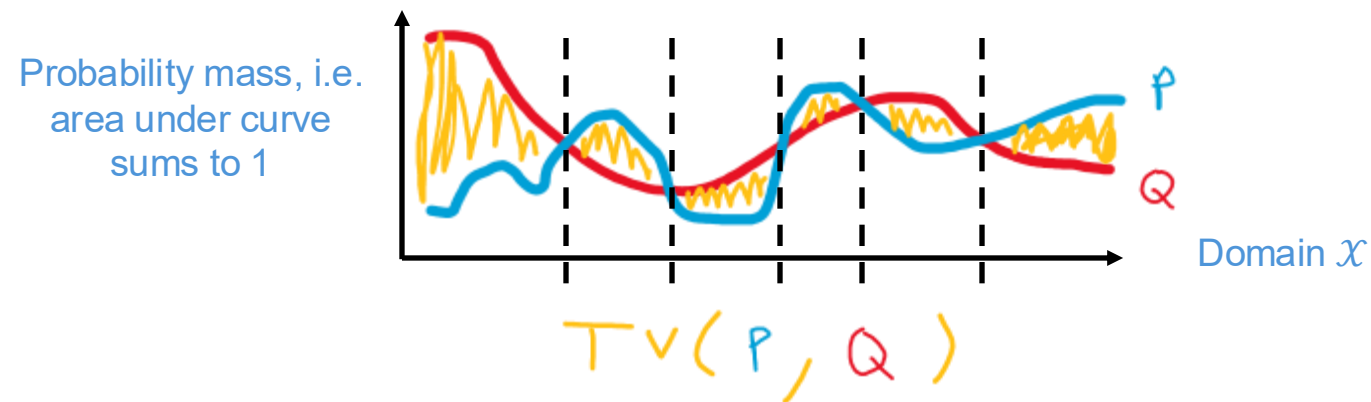
- Classic results in statistics show asymptotic convergence of estimators in the limit of large data
- Probably Approximately Correct (PAC) learning model [Val84]
 - Given sample access to distribution \mathcal{P} , produce $\hat{\mathcal{P}}$ such that $\text{TV}(\mathcal{P}, \hat{\mathcal{P}}) \leq \varepsilon$ w.p. $\geq 1 - \delta$



RT1: Foundational AI / ML Research

Background on some statistical problems

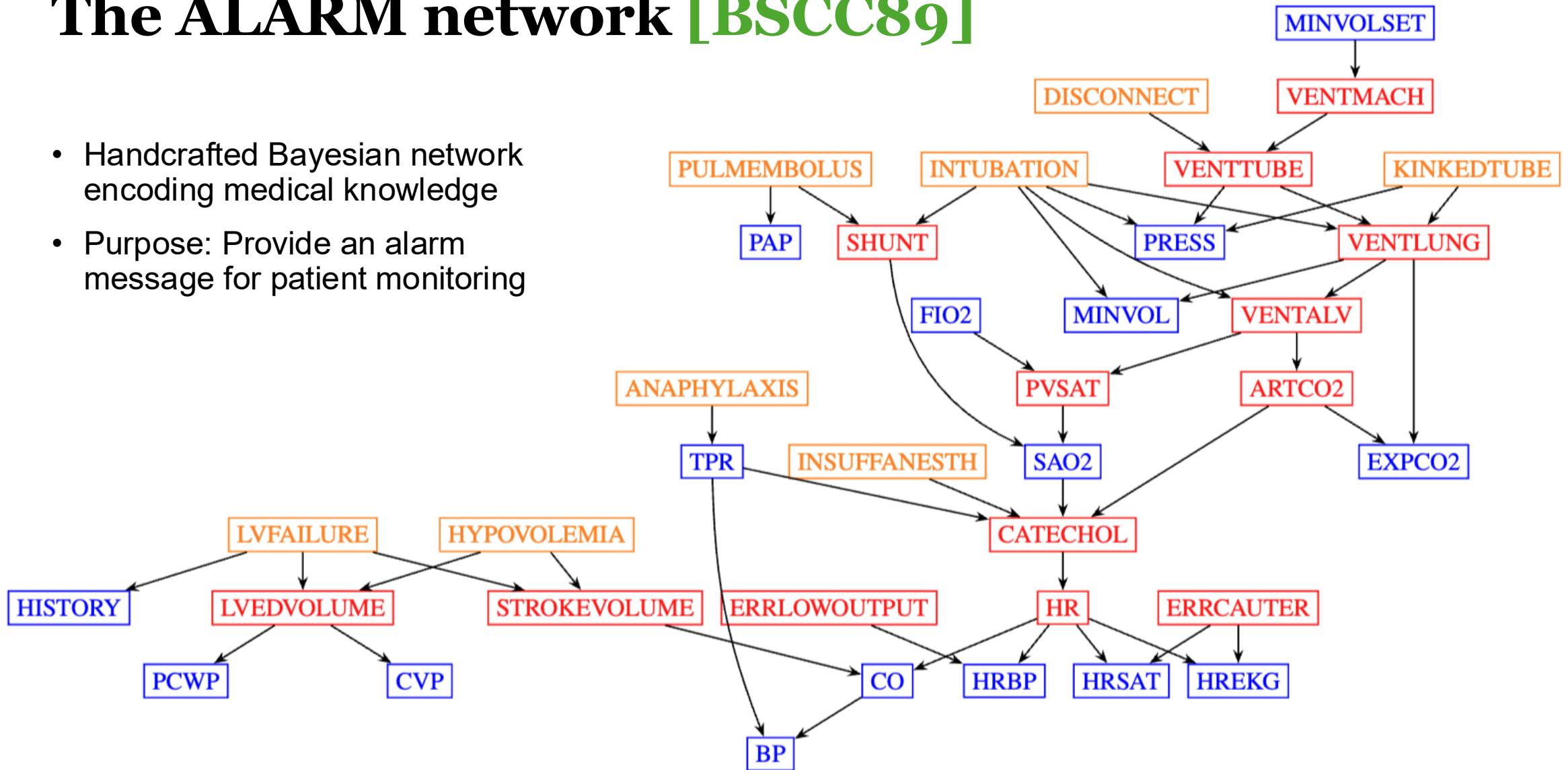
- Classic results in statistics show asymptotic convergence of estimators in the limit of large data
- Probably Approximately Correct (PAC) learning model [Val84]
 - Given sample access to distribution \mathcal{P} , produce $\hat{\mathcal{P}}$ such that $\text{TV}(\mathcal{P}, \hat{\mathcal{P}}) \leq \varepsilon$ w.p. $\geq 1 - \delta$



- Bayesian networks [Pea88] are a probabilistic graphical model commonly used to model beliefs
 - \mathcal{P} on n variables \Leftrightarrow Directed acyclic graph (DAG) over n vertices + conditional distributions at vertices

The ALARM network [BSCC89]

- Handcrafted Bayesian network encoding medical knowledge
- Purpose: Provide an alarm message for patient monitoring



RT1: Foundational AI / ML Research

Some of my results on statistical learning problems

- Suppose distribution \mathcal{P} is described by Bayesian network
 - Knowledge is encoded in *absence* of edges since these encode conditional independences
 - A clique can represent any distribution, but sparse representations are more interpretable

RT1: Foundational AI / ML Research

Some of my results on statistical learning problems

- Suppose distribution \mathcal{P} is described by Bayesian network
 - Knowledge is encoded in *absence* of edges since these encode conditional independences
 - A clique can represent any distribution, but sparse representations are more interpretable
- NP-hard to find “score maximizing” DAG from data [Chi96] and to decide whether \mathcal{P} can be described by a DAG with p parameters [CHM04]
- Even under the promise that \mathcal{P} can be described by a DAG with p parameters, it is NP-hard to find such a parameter-bounded DAG [BCGM25]
- We also have some PAC-style finite sample results in learning the structure and parameters of Bayesian network for \mathcal{P} [BCG+22, DDKC23, CYBC24]
- **Insight: If network’s in-degree is bounded, we can use less samples**

(Remark: Some papers are in alphabetical ordering of last names, and my PhD advisor is Arnab Bhattacharrya)

[Chi96] David Maxwell Chickering. *Learning Bayesian networks is NP-complete*. Lecture Notes in Statistics, vol 112, 1996

[CHM04] Max Chickering, David Heckerman, and Chris Meek. *Large-sample learning of Bayesian networks is NP-hard*. Journal of Machine Learning Research (JMLR), 2004

[BCGM25] Arnab Bhattacharyya, Davin Choo, Sutanu Gayen, Dimitrios Myrisiotis. *Learnability of Parameter-Bounded Bayes Nets*. AAAI Conference on Artificial Intelligence (AAAI), 2025

[BCG+22] Arnab Bhattacharyya, Davin Choo, Rishikesh Gajjala, Sutanu Gayen, Yuhao Wang. *Learning Sparse Fixed-Structure Gaussian Bayesian Networks*. International Conference on Artificial Intelligence and Statistics (AISTATS), 2022

[DDKC23] Yuval Dagan, Constantinos Daskalakis, Anthimos-Vardis Kandiros, Davin Choo. *Learning and Testing Latent-Tree Ising Models Efficiently*. Conference on Learning Theory (COLT), 2023

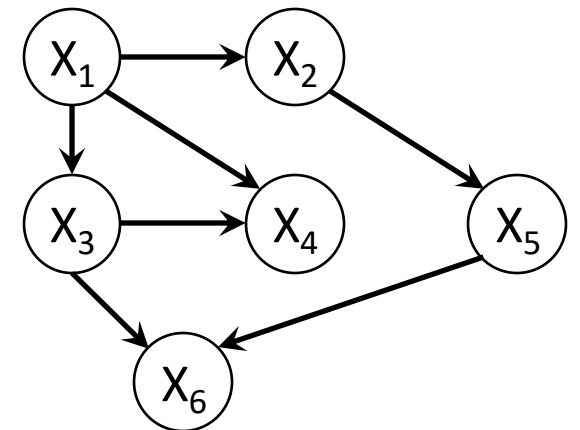
A glimpse of [BCG+22]

Insight: If network's in-degree is bounded, we can use less samples

- Consider i.i.d. samples from \mathcal{P} described by this Bayesian network
- Under the *linear Gaussian* setting, this is *equivalent* to drawing i.i.d. samples from some underlying n -dimensional Gaussian

$$X^{(1)}, \dots, X^{(m)} \in \mathbb{R}^n, \text{ where } n = 6 \text{ in this example}$$

- Question: How many samples to produce a candidate distribution $\hat{\mathcal{P}}$ so that $\text{TV}(\mathcal{P}, \hat{\mathcal{P}}) \leq \varepsilon$?



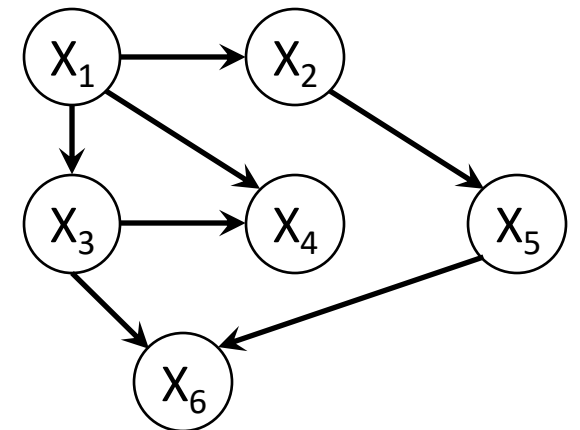
A glimpse of [BCG+22]

Insight: If network's in-degree is bounded, we can use less samples

- Consider i.i.d. samples from \mathcal{P} described by this Bayesian network
- Under the *linear Gaussian* setting, this is *equivalent* to drawing i.i.d. samples from some underlying n -dimensional Gaussian

$$X^{(1)}, \dots, X^{(m)} \in \mathbb{R}^n, \text{ where } n = 6 \text{ in this example}$$

- Question: How many samples to produce a candidate distribution $\hat{\mathcal{P}}$ so that $\text{TV}(\mathcal{P}, \hat{\mathcal{P}}) \leq \varepsilon$?
- In general, $m \in \tilde{\Omega}\left(\frac{n^2}{\varepsilon^2}\right)$ samples necessary
 - Intuition: All n^2 entries of covariance matrix “matters”
- [BCG+22]: $m \in \tilde{\mathcal{O}}\left(\frac{nd}{\varepsilon^2}\right)$ samples suffices
 - In this example, maximum in-degree $d = 2$
 - Idea: Least squares estimation at each node + careful analysis
 - Least squares to estimate linear coefficients in the linear Gaussian setting



RT1: Foundational AI / ML Research

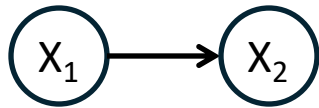
Correlation does not imply causation



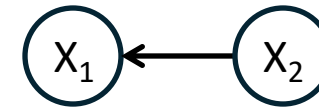
RT1: Foundational AI / ML Research

Two equivalent statistical models could have generated this data

X_1	-0.27	0.29	0.37	-0.09	0.34	0.33	0.30	-1.34	0.68
X_2	-0.10	1.65	0.47	1.92	2.04	1.67	0.11	-3.58	1.97



- $\varepsilon_1 \sim N(0, 1)$
- $\varepsilon_2 \sim N(0, 1)$
- $X_1 = \varepsilon_1 \sim N(0, 1)$
- $X_2 = X_1 + \varepsilon_2 \sim N(0, 2)$



- $\varepsilon_3 \sim N(0, 2)$
- $\varepsilon_4 \sim N(0, 1/2)$
- $X_2 = \varepsilon_3 \sim N(0, 2)$
- $X_1 = \frac{1}{2} \cdot X_2 + \varepsilon_4 \sim N(0, 1)$

For statistical purposes, both models are fine and equally good.
However, from a causal standpoint, they represent very different underlying systems.

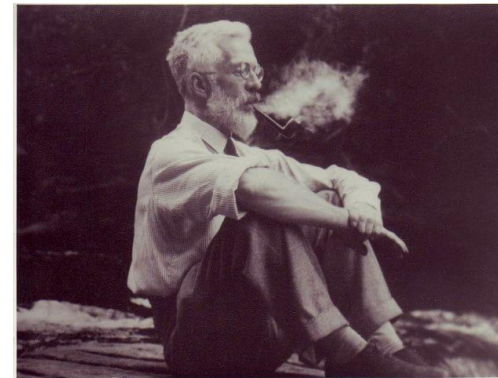
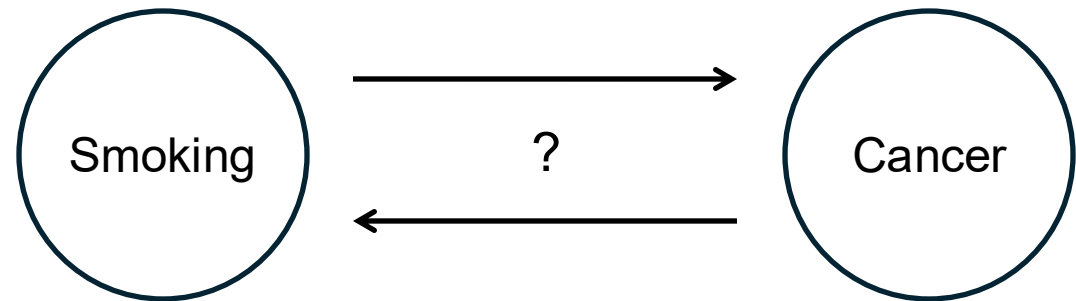
RT1: Foundational AI / ML Research

Figuring out the right causal direction has important real-world ramifications

Smoking	Y	N	N	N	Y	Y	N	N	Y
Cancer	Y	Y	N	N	N	Y	N	N	Y

Fisher's letter to Nature, 1958

“The curious associations with lung cancer found in relation to smoking habits do not, in the minds of some of us, lend themselves easily to the simple conclusion that the products of combustion reaching the surface of the bronchus induce, though after a long interval, the development of a cancer... *Such results suggest that an error has been made, of an old kind, in arguing from correlation to causation...*”



RT1: Foundational AI / ML Research

Key problems in causal inference

- Causal representation learning [[CSB22](#), [CS23a](#), [CGB23](#), [CS23b](#), [CS23c](#), [CSU24](#)]
 - How are things in the system related? What affects what?
 - Examples: DAGs, CPDAGs, ancestral graphs, etc.
- Causal effect estimation [[CSBS24](#)]
 - How much does X affect Y?
 - Examples: Average Treatment Effect (ATE), Conditional ATE (CATE), etc.

Solving these problems enable better predictive and prescriptive decisions

- Predictive: What will happen if we change something?
- Prescriptive: What should we change if we want to achieve something?

[[CSB22](#)] Davin Choo, Kirankumar Shiragur, Arnab Bhattacharyya. *Verification and search algorithms for causal DAGs*. Conference on Neural Information Processing Systems (NeurIPS), 2022.

[[CS23a](#)] Davin Choo, Kirankumar Shiragur. *Subset verification and search algorithms for causal DAGs*. International Conference on Artificial Intelligence and Statistics (AISTATS), 2023.

[[CGB23](#)] Davin Choo, Themistoklis Gouleakis, Arnab Bhattacharyya. *Active causal structure learning with advice*. International Conference on Machine Learning (ICML), 2023.

[[CS23b](#)] Davin Choo, Kirankumar Shiragur. *New metrics and search algorithms for weighted causal DAGs*. International Conference on Machine Learning (ICML), 2023.

[[CS23c](#)] Davin Choo, Kirankumar Shiragur. *Adaptivity Complexity for Causal Graph Discovery*. Conference on Uncertainty in Artificial Intelligence (UAI), 2023.

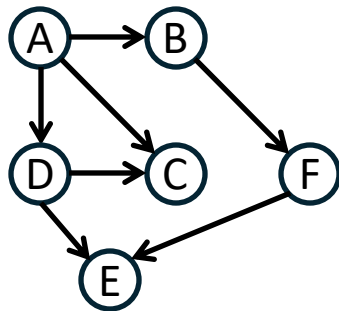
[[CSU24](#)] Davin Choo, Kirankumar Shiragur, Caroline Uhler. *Causal discovery under off-target interventions*. International Conference on Artificial Intelligence and Statistics (AISTATS), 2024.

[[CSBS24](#)] Davin Choo, Chandler Squires, Arnab Bhattacharyya, and David Sontag. *Probably approximately correct high-dimensional causal effect estimation given a valid adjustment set*. Conference on Causal Learning and Reasoning (CLeaR), 2024.

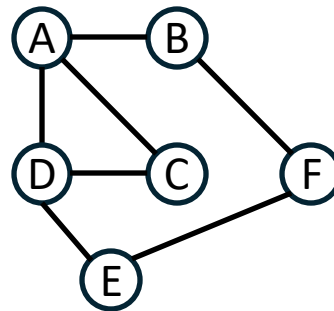
A glimpse of [CSB22]

With just observational data, there are statistically indistinguishable models

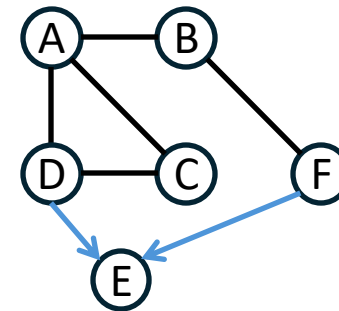
- Even with infinite observational data, can only determine causal graph up to some equivalence class where all conditional independence relations agree. E.g., $X_1 \rightarrow X_2$ vs. $X_1 \leftarrow X_2$ earlier
- Graphically, equivalence class of a DAG (a.k.a., essential graph) is **partially oriented** version of the DAG on the same variables and edge set (i.e., same skeleton)



True underlying DAG



Skeleton



Essential graph

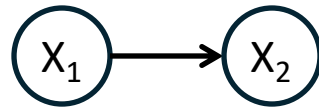
A glimpse of [CSB22]

With just observational data, there are statistically indistinguishable models

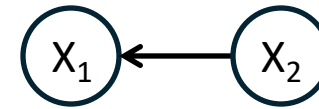
- Even with infinite observational data, can only determine causal graph up to some equivalence class where all conditional independence relations agree. E.g., $X_1 \rightarrow X_2$ vs. $X_1 \leftarrow X_2$ earlier

Interventions can help distinguish between causal models

- Randomized Controlled Trials (RCTs) are the gold standard



- $X_1 = \varepsilon_1 \sim N(0, 1)$
- $X_2 = X_1 + \varepsilon_2 \sim N(0, 2)$



- $X_2 = \varepsilon_3 \sim N(0, 2)$
- $X_1 = \frac{1}{2} \cdot X_2 + \varepsilon_4 \sim N(0, 1)$

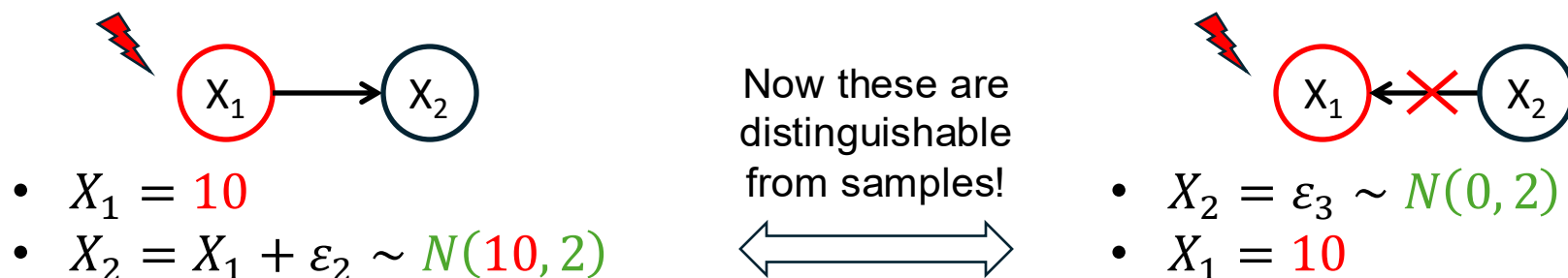
A glimpse of [CSB22]

With just observational data, there are statistically indistinguishable models

- Even with infinite observational data, can only determine causal graph up to some equivalence class where all conditional independence relations agree. E.g., $X_1 \rightarrow X_2$ vs. $X_1 \leftarrow X_2$ earlier

Interventions can help distinguish between causal models

- Randomized Controlled Trials (RCTs) are the gold standard



- **Intuitively, interventions recover orientation of incident edges**
(There are additional details, and something called Meek's rules, but we will ignore for this talk)

A glimpse of [CSB22]

With just observational data, there are statistically indistinguishable models

Interventions can help distinguish between causal models

Verifying and searching causal DAGs using *adaptive* interventions

1. Many possible DAGs agree with observational data
 2. Perform interventions and gather interventional data
 3. Repeat step 2 until we can identify the true underlying DAG generating the data
- Our decision can depend on revealed information

Search

How many *adaptive* interventions do we need to figure out the true underlying DAG?

Verification

If someone gives us a proposed DAG, how many interventions to determine TRUE or FALSE?

A glimpse of [CSB22]

With just observational data, there are statistically indistinguishable models

Interventions can help distinguish between causal models

Verifying and searching causal DAGs using *adaptive* interventions

- Let MVC represent “minimum vertex cover”
- Trivial search solution: Intervene on MVC of unoriented edges

A glimpse of [CSB22]

With just observational data, there are statistically indistinguishable models

Interventions can help distinguish between causal models

Verifying and searching causal DAGs using *adaptive* interventions

- Let MVC represent “minimum vertex cover” and $\nu(G^*)$ be “verification number”
 - Note that $\nu(G^*)$ is a function of the ground truth DAG G^*
- Trivial search solution: Intervene on MVC of unoriented edges
- [CSB22]: **Exact characterization** of verification number $\nu(G^*)$ via MVC of covered edges
 - [Chi95]: $u - v$ is *covered edge* if and only if $\text{Pa}(u) \setminus \{v\} = \text{Pa}(v) \setminus \{u\}$ in the underlying DAG
 - **Implication:** Prior bounds on $\nu(G^*)$ are loose + covered edges perspective enables nice re-interpretation of known facts about interventions for causal graph discovery

A glimpse of [CSB22]

With just observational data, there are statistically indistinguishable models

Interventions can help distinguish between causal models

Verifying and searching causal DAGs using *adaptive* interventions

- Let MVC represent “minimum vertex cover” and $\nu(G^*)$ be “verification number”
 - Note that $\nu(G^*)$ is a function of the ground truth DAG G^*
- Trivial search solution: Intervene on MVC of unoriented edges
- [CSB22]: **Exact characterization** of verification number $\nu(G^*)$ via MVC of covered edges
 - [Chi95]: $u - v$ is *covered edge* if and only if $\text{Pa}(u) \setminus \{v\} = \text{Pa}(v) \setminus \{u\}$ in the underlying DAG
 - **Implication**: Prior bounds on $\nu(G^*)$ are loose + covered edges perspective enables nice re-interpretation of known facts about interventions for causal graph discovery
- [CSB22]: $\mathcal{O}(\nu(G^*) \cdot \log n)$ *adaptive* interventions suffice for searching. Worst case necessary.
 - Since the search algorithm does not know the ground truth DAG G^* , it does not know $\nu(G^*)$
 - Yet, we can guarantee at most an $\mathcal{O}(\log n)$ overhead compared to the easier verification problem

RT1: Foundational AI / ML Research

One final thing in RT1: revisit computational subtasks of modern ML pipelines

- Many existing subroutines in modern ML pipelines are heuristically chosen because “it works well in practice” or “it beats the previous baseline on benchmark X”

My claim: There are potential untapped gains if we principally approach these tasks

- [LTCLB25]: New principally designed tokenization scheme which empirically beats the commonly used BPE in both compression ratio and downstream tasks, in a plug-and-play fashion



**Some of my
past and
current
efforts**

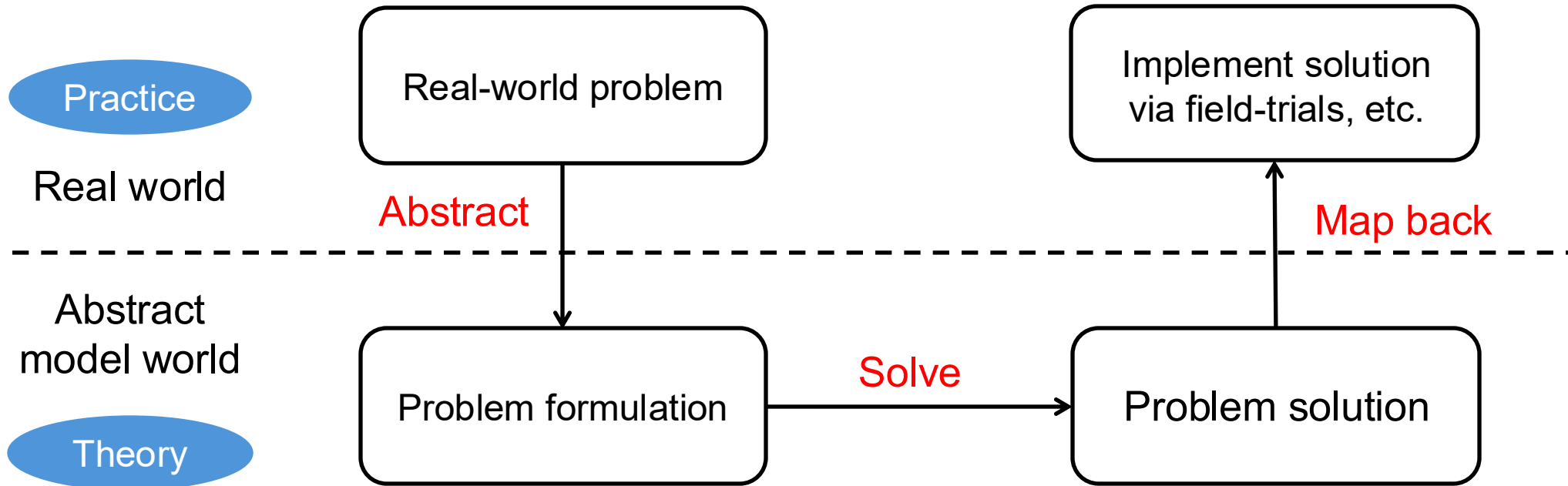
RT1: Foundational AI / ML Research

- Statistical and causal problems
- Improving subtasks in modern ML pipelines

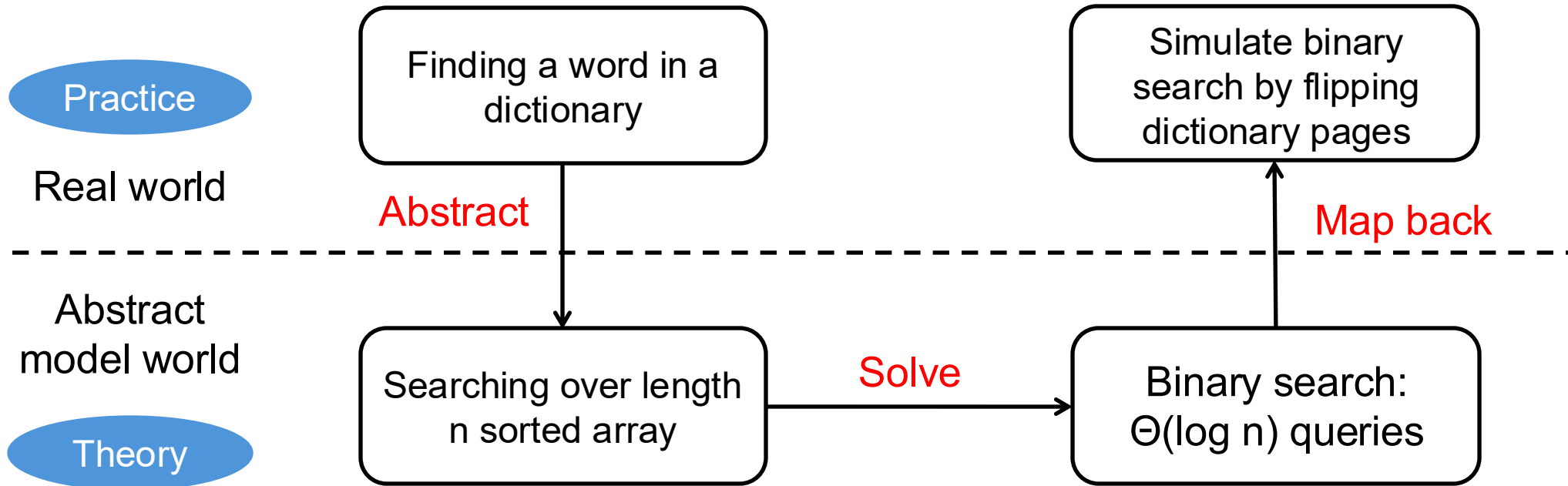
RT2: Algorithms with Imperfect Advice

RT3: AI + X

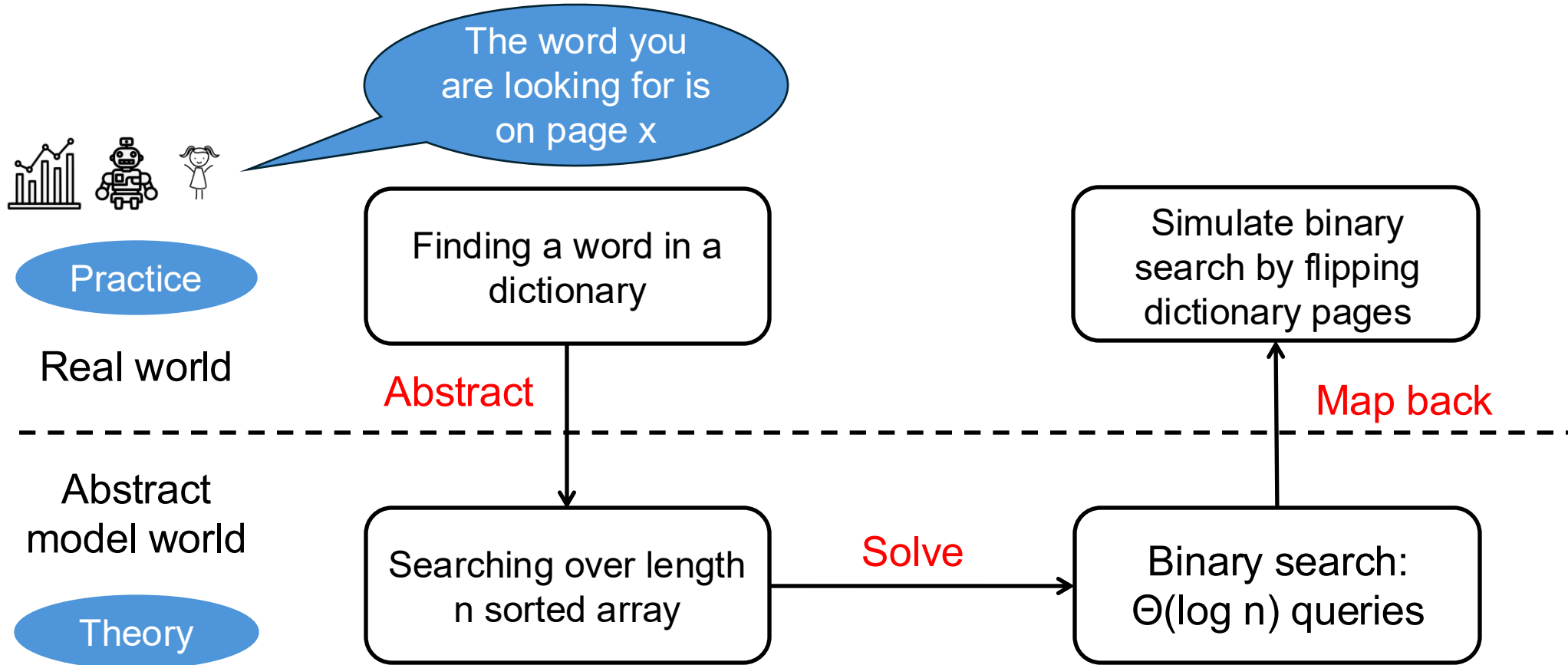
Learning-augmented algorithms are a way to harness imperfect instance-specific information



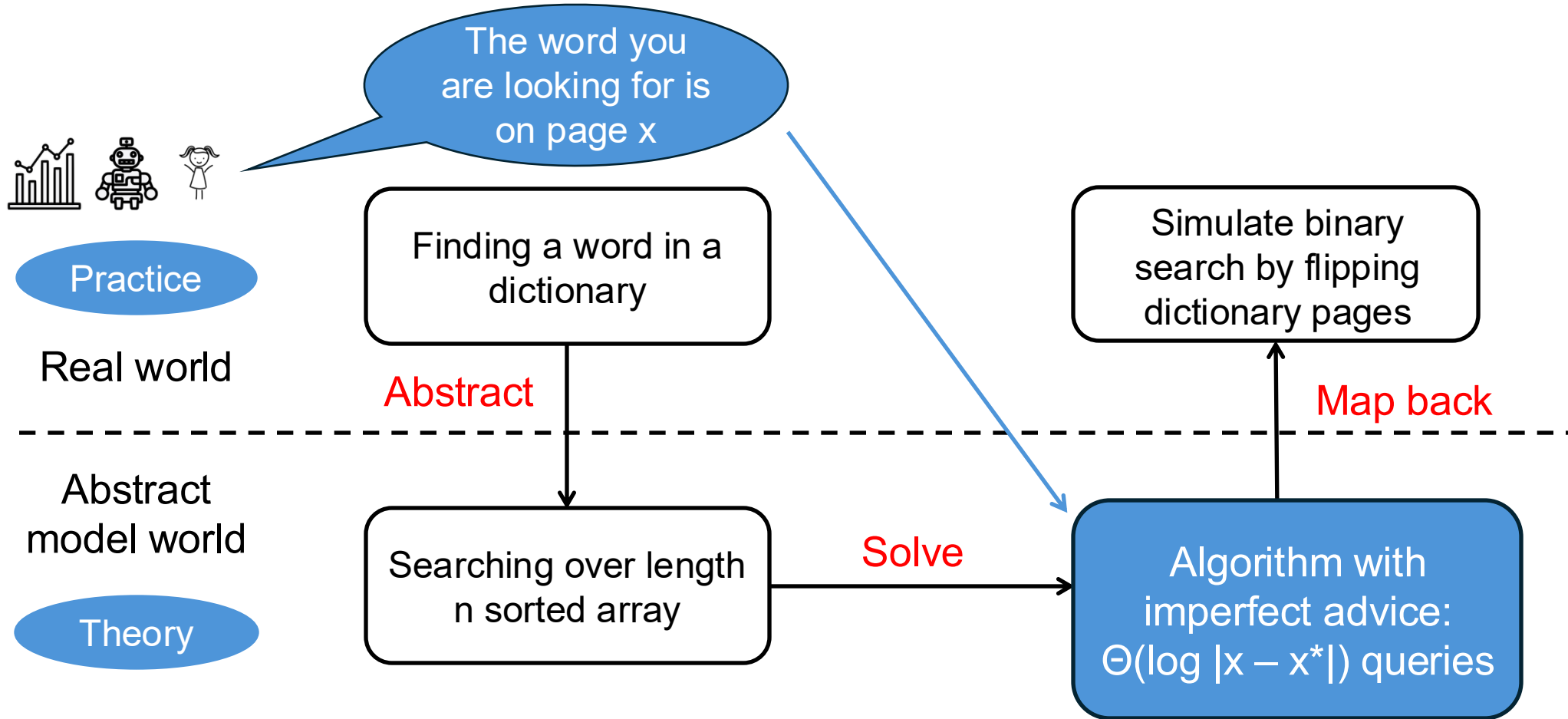
Learning-augmented algorithms are a way to harness imperfect instance-specific information



Learning-augmented algorithms are a way to harness imperfect instance-specific information



Learning-augmented algorithms are a way to harness imperfect instance-specific information



RT2: Algorithms with Imperfect Advice

The Test-and-Act framework

- **Insight: “Testing *can* be easier than learning”**
- Developed as part of my PhD, where we leverage on results from property testing literature
- [CGB23] [TestAndSubsetSearch](#): Reduce number of interventions for causal graph discovery
- [CGLB24] [TestAndMatch](#): Improve competitive ratio of online bipartite matching
- [BCGG24] [TestAndOptimize](#): Improve sample complexity of learning multivariate Gaussians
- [BCGG25] [TestAndOptimize](#): Improve sample complexity of learning product distributions

Other results in learning-augmented algorithms

- [CJS25] We design new learning-augmented algorithms for fractional online bipartite matching that trade-offs consistency and robustness --- two standard metrics in the literature

[CGB23] Davin Choo, Themistoklis Gouleakis, Arnab Bhattacharyya. *Active causal structure learning with advice*. International Conference on Machine Learning (ICML), 2023.

[CGLB24] Davin Choo, Themistoklis Gouleakis, Chun Kai Ling, and Arnab Bhattacharyya. *Online bipartite matching with imperfect advice*. International Conference on Machine Learning (ICML), 2024.

[BCGG24] Arnab Bhattacharyya, Davin Choo, Philips George John, and Themistoklis Gouleakis. *Learning multivariate Gaussians with imperfect advice*. International Conference on Machine Learning (ICML), 2024.

[BCGG25] Arnab Bhattacharyya, Davin Choo, Philips George John, and Themistoklis Gouleakis. *Product Distribution Learning with Imperfect Advice*. Conference on Neural Information Processing Systems (NeurIPS), 2025.

[CJS25] Davin Choo, Billy Jin, and Yongho Shin. *Learning-Augmented Online Bipartite Fractional Matching*. Conference on Neural Information Processing Systems (NeurIPS), 2025.

A glimpse of [BCGG24]

Producing a candidate distribution $\hat{\mathcal{P}}$ when drawing i.i.d. samples from $\mathcal{P} = N(\mu, I)$

- PAC-learning: We want $TV(\mathcal{P}, \hat{\mathcal{P}})$ to be small, with “good chance”, using few samples
- For Gaussians with identity covariance, we just need good estimation $\hat{\mu}$ of the mean $\mu \in \mathbb{R}^d$

A glimpse of [BCGG24]

Producing a candidate distribution $\hat{\mathcal{P}}$ when drawing i.i.d. samples from $\mathcal{P} = N(\mu, I)$

- PAC-learning: We want $TV(\mathcal{P}, \hat{\mathcal{P}})$ to be small, with “good chance”, using few samples
- For Gaussians with identity covariance, we just need good estimation $\hat{\mu}$ of the mean $\mu \in \mathbb{R}^d$

Learning Gaussians: $\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to produce “ ε -good” $\hat{\mu}$

- Idea: Empirical mean works

Quadratic gap!

Tolerant testing Gaussians: $\tilde{\Theta}\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to “ ε -tolerant-test” μ

- If $\|\mu\|_2 \leq \varepsilon$, output Accept
- If $\|\mu\|_2 \geq 2\varepsilon$, output Reject
- Otherwise, decide arbitrarily
- Idea: Define statistic Z and threshold τ . Output “Accept” if $Z > \tau$, else output “Reject”

A glimpse of [BCGG24]

Learning Gaussians: $\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to produce “ ε -good” $\hat{\mu}$

Tolerant testing Gaussians: $\tilde{\Theta}\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to “ ε -tolerant-test” μ

[BCGG24]: Efficient algorithm for improving sample complexity given advice $\tilde{\mu} \in \mathbb{R}^d$

- If $\ell_2(\mu, \tilde{\mu})$ is small enough, just output $\tilde{\mu}$ with zero samples
- Unfortunately, we do not know the advice quality

A glimpse of [BCGG24]

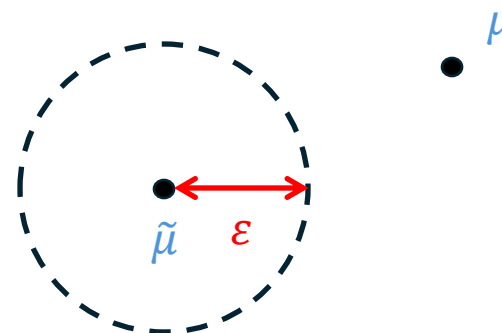
Learning Gaussians: $\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to produce “ ε -good” $\hat{\mu}$

Tolerant testing Gaussians: $\tilde{\Theta}\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to “ ε -tolerant-test” μ

[BCGG24]: Efficient algorithm for improving sample complexity given advice $\tilde{\mu} \in \mathbb{R}^d$

- If $\ell_2(\mu, \tilde{\mu})$ is small enough, just output $\tilde{\mu}$ with zero samples
- Unfortunately, we do not know the advice quality
- Idea: Use tolerant testing to first estimate advice quality of $\tilde{\mu}$

“Guess and double”



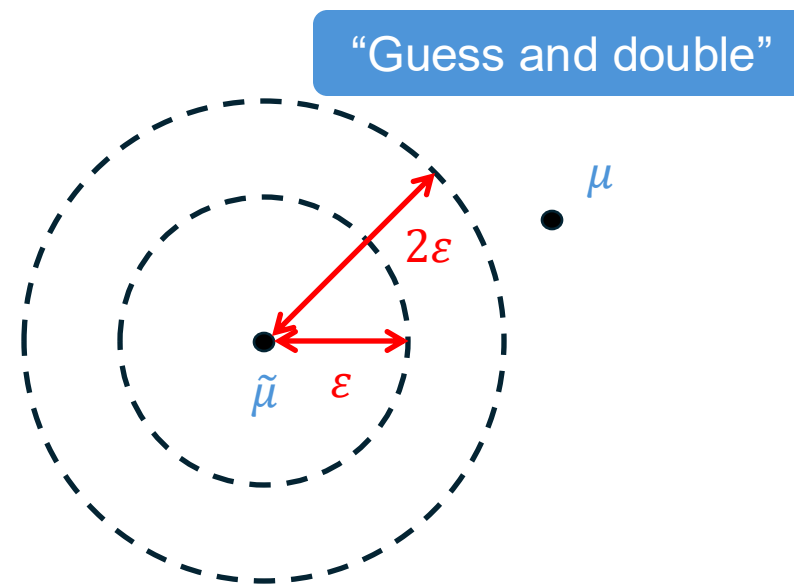
A glimpse of [BCGG24]

Learning Gaussians: $\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to produce “ ε -good” $\hat{\mu}$

Tolerant testing Gaussians: $\tilde{\Theta}\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to “ ε -tolerant-test” μ

[BCGG24]: Efficient algorithm for improving sample complexity given advice $\tilde{\mu} \in \mathbb{R}^d$

- If $\ell_2(\mu, \tilde{\mu})$ is small enough, just output $\tilde{\mu}$ with zero samples
- Unfortunately, we do not know the advice quality
- Idea: Use tolerant testing to first estimate advice quality of $\tilde{\mu}$



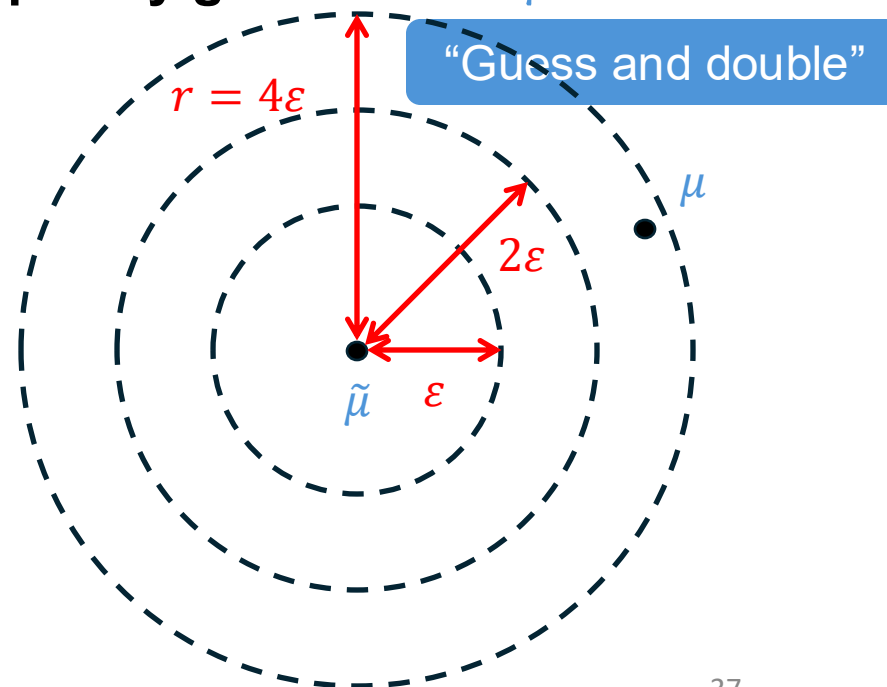
A glimpse of [BCGG24]

Learning Gaussians: $\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to produce “ ε -good” $\hat{\mu}$

Tolerant testing Gaussians: $\tilde{\Theta}\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to “ ε -tolerant-test” μ

[BCGG24]: Efficient algorithm for improving sample complexity given advice $\tilde{\mu} \in \mathbb{R}^d$

- If $\ell_2(\mu, \tilde{\mu})$ is small enough, just output $\tilde{\mu}$ with zero samples
- Unfortunately, we do not know the advice quality
- Idea: Use tolerant testing to first estimate advice quality of $\tilde{\mu}$



A glimpse of [BCGG24]

Learning Gaussians: $\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to produce “ ε -good” $\hat{\mu}$

Tolerant testing Gaussians: $\tilde{\Theta}\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to “ ε -tolerant-test” μ

[BCGG24]: Efficient algorithm for improving sample complexity given advice $\tilde{\mu} \in \mathbb{R}^d$

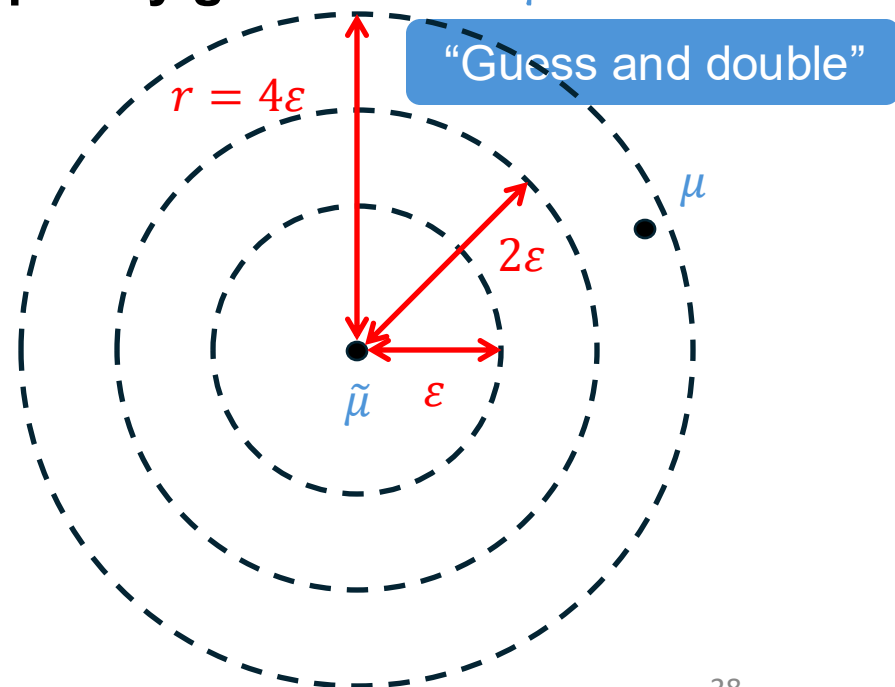
- If $\ell_2(\mu, \tilde{\mu})$ is small enough, just output $\tilde{\mu}$ with zero samples
- Unfortunately, we do not know the advice quality
- Idea: Use tolerant testing to first estimate advice quality of $\tilde{\mu}$, then run LASSO-style optimization to produce $\hat{\mu}$

$$\hat{\mu} = \operatorname{argmin}_{\|\beta\|_1 \leq r} \sum_{i=1}^n \|x^{(i)} - \beta\|_2^2$$

samples

- Polynomial time with overall sample complexity

$$\tilde{\Theta}\left(\frac{d}{\varepsilon^2} \left(\min \left\{ 1, \frac{\|\mu - \tilde{\mu}\|_1^2}{\varepsilon^2 d} \right\} \right)\right)$$



A glimpse of [BCGG24]

Learning Gaussians: $\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to produce “ ε -good” $\hat{\mu}$

Tolerant testing Gaussians: $\tilde{\Theta}\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to “ ε -tolerant-test” μ

[BCGG24]: Efficient algorithm for improving sample complexity given advice $\tilde{\mu} \in \mathbb{R}^d$

- If $\ell_2(\mu, \tilde{\mu})$ is small enough, just output $\tilde{\mu}$ with zero samples
- Unfortunately, we do not know the advice quality
- Idea: Use tolerant testing to first estimate advice quality of $\tilde{\mu}$, then run LASSO-style optimization to produce $\hat{\mu}$

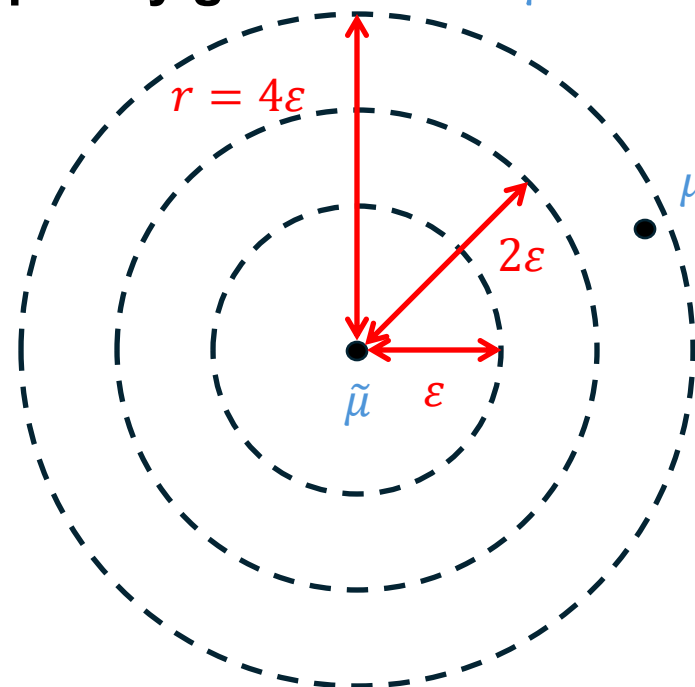
These ℓ_1 are not typos. Happy to discuss after the talk

$$\hat{\mu} = \operatorname{argmin}_{\|\beta\|_1 \leq r} \sum_{i=1}^n \|x^{(i)} - \beta\|_2^2$$

samples

- Polynomial time with overall sample complexity

$$\tilde{\Theta}\left(\frac{d}{\varepsilon^2} \left(\min \left\{ 1, \frac{\|\mu - \tilde{\mu}\|_1^2}{\varepsilon^2 d} \right\} \right)\right)$$



A glimpse of [BCGG24]

Learning Gaussians: $\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to produce “ ε -good” $\hat{\mu}$

Tolerant testing Gaussians: $\tilde{\Theta}\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$ to “ ε -tolerant-test” μ

[BCGG24]: Efficient algorithm for improving sample complexity given advice $\tilde{\mu} \in \mathbb{R}^d$

- Polynomial time with overall sample complexity

$$\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\left(\min\left\{1, \frac{\|\mu - \tilde{\mu}\|_1^2}{\varepsilon^2 d}\right\}\right)\right)$$

- When $\|\mu - \tilde{\mu}\|_1 \ll \varepsilon\sqrt{d}$, sample complexity is *sublinear in d*
- $\Omega(d)$ samples unavoidable when $\|\mu - \tilde{\mu}\|_1 \in \Omega(\varepsilon\sqrt{d})$
- Bound can be slightly parameterized, and similar idea works for non-identity covariance (with SDP instead of LASSO; matching lower bound exists)



Some of my past and current efforts

RT1: Foundational AI / ML Research

- Statistical and causal problems
- Improving subtasks in modern ML pipelines

RT2: Algorithms with Imperfect Advice

- Test-and-Act framework

RT3: AI + X

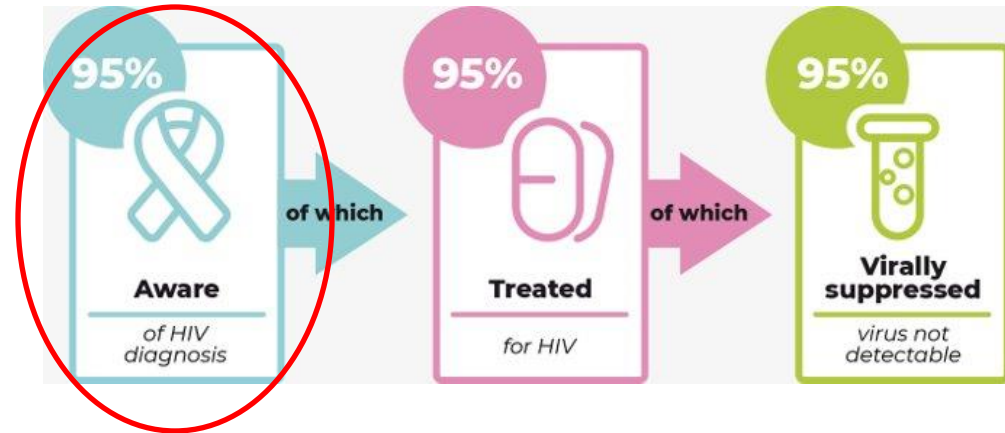
Targeting the first 95% in UN's 95-95-95 initiative

Global initiative by UNAIDS to control the HIV epidemic

- Undetectable = Untransmittable (U = U)
- UN Sustainable Development Goal 3.3

Motivation

- The faster we detect positive cases, the faster we can start treatment cascade
- Individuals who learn about being HIV+ undergo behavioral changes that help limit spread
- Individuals with low-risk characteristics may be interacting with high-risk individuals



Targeting the first 95% in UN's 95-95-95 initiative

Global initiative by UNAIDS to control the HIV epidemic

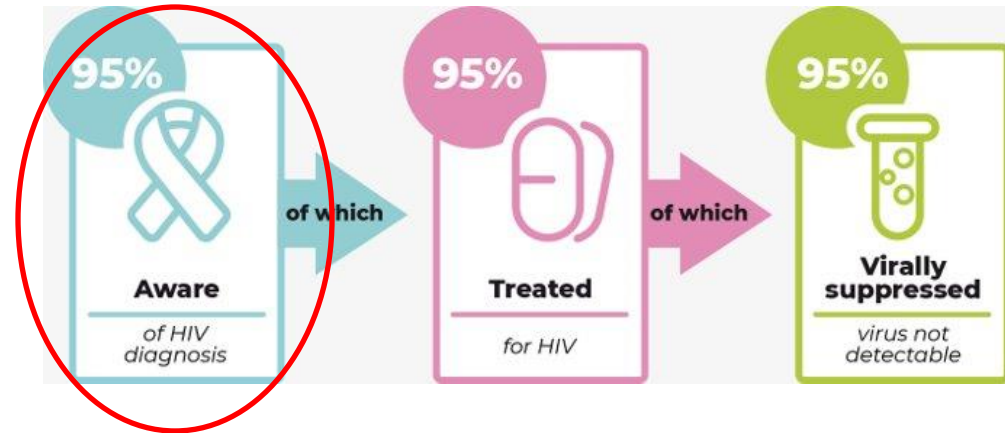
- Undetectable = Untransmittable (U = U)
- UN Sustainable Development Goal 3.3

Motivation

- The faster we detect positive cases, the faster we can start treatment cascade
- Individuals who learn about being HIV+ undergo behavioral changes that help limit spread
- Individuals with low-risk characteristics may be interacting with high-risk individuals

Challenges

- In 2024, WHO estimates that 1 in 7 HIV positive individuals do not know they are infected
- Current testing prioritizations do not effectively exploit the underlying disease transmission network or social network amongst the population
- Resource limitations due to funding cuts, e.g. to USAID and WHO



Re-imagining a new testing approach

Objective

- Maximize efficiency of testing resources in detecting HIV+ cases as quickly as possible
 - Resource can be # test kits, or where to focus efforts of human workers in recruiting people for testing

Re-imagining a new testing approach

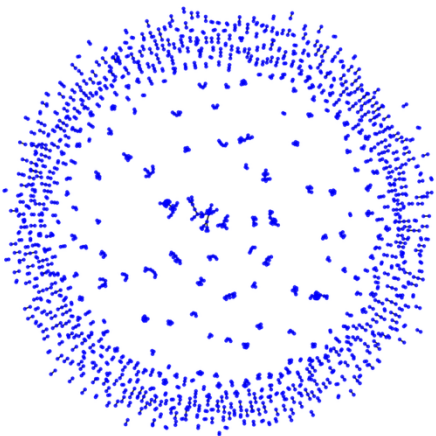
Objective

- Maximize efficiency of testing resources in detecting HIV+ cases as quickly as possible
 - Resource can be # test kits, or where to focus efforts of human workers in recruiting people for testing

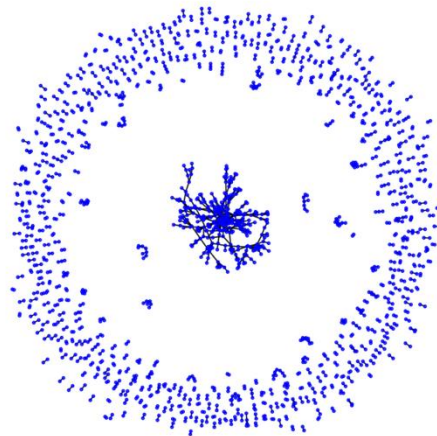
Constraints and structural properties of our problem

- Sexually transmitted diseases do not spread like flu
 - The transmission graph \mathcal{G} is sparse and tree-like
 - Note: We only ever observe a subset of the true graph

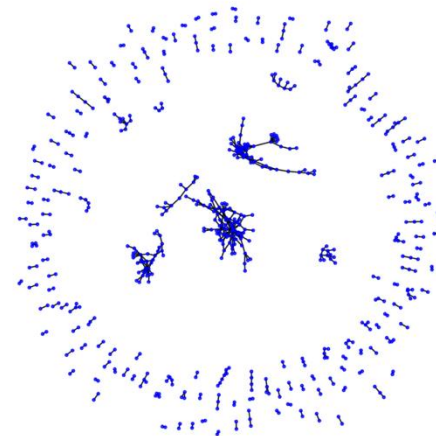
Gonorrhea



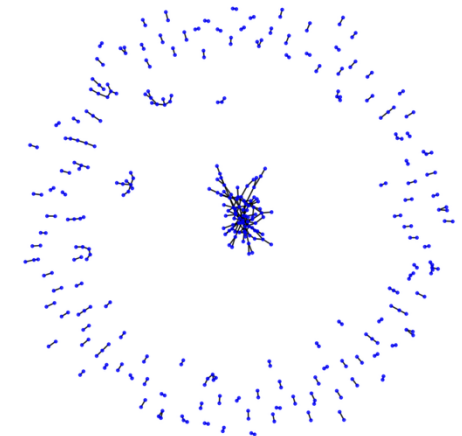
Hepatitis



HIV



Syphilis



Re-imagining a new testing approach

Objective

- Maximize efficiency of testing resources in detecting HIV+ cases as quickly as possible
 - Resource can be # test kits, or where to focus efforts of human workers in recruiting people for testing

Constraints and structural properties of our problem

- Sexually transmitted diseases do not spread like flu
 - The transmission graph \mathcal{G} is sparse and tree-like
 - Note: We only ever observe a subset of the true graph
- Policy-wise, it is preferable to test individuals whose neighbors (in \mathcal{G}) have been tested
 - This is because it is more informative than randomly picking individuals to test next
 - This maps to a kind of “frontier exploration” constraint on the graph
 - First tested person (the root) of each component is chosen via some domain policy consideration



Re-imagining a new testing approach

Objective

- Maximize efficiency of testing resources in detecting HIV+ cases as quickly as possible
 - Resource can be # test kits, or where to focus efforts of human workers in recruiting people for testing

Constraints and structural properties of our problem

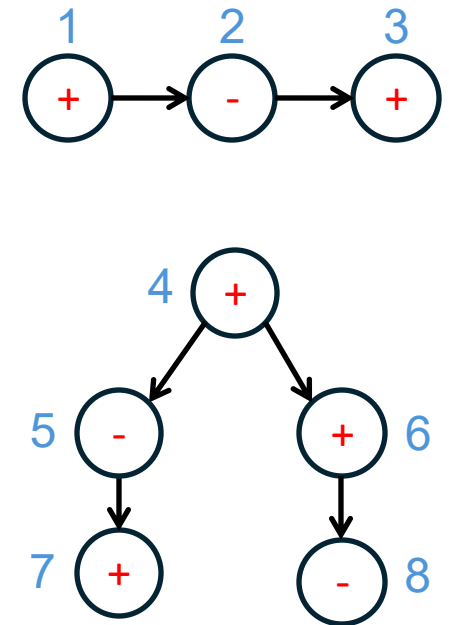
- Sexually transmitted diseases do not spread like flu
 - The transmission graph \mathcal{G} is sparse and tree-like
 - Note: We only ever observe a subset of the true graph
- Policy-wise, it is preferable to test individuals whose neighbors (in \mathcal{G}) have been tested
 - This is because it is more informative than randomly picking individuals to test next
 - This maps to a kind of “frontier exploration” constraint on the graph
 - First tested person (the root) of each component is chosen via some domain policy consideration
- Mathematically speaking, there is some underlying transmission probability \mathcal{P} that is Markov with respect to \mathcal{G}
 - Every individual has an unobserved discrete label of $+$ or $-$, that is revealed upon testing
 - Markov: $\mathcal{P}(\text{person} = + \mid \text{revealed statuses}) = \mathcal{P}(\text{person} = + \mid \text{revealed statuses of neighbors})$

Designing a reward metric for optimization

Illustrative example

- Suppose we know the underlying labels (disease status) on a forest
- How good is this sequence of testing?
 - Early detection → early intervention; Also, we may have sudden budget cuts

Testing budget	1	2	3	4	5	6	7	8
# positive detected	1	1	2	3	3	4	5	5



Designing a reward metric for optimization

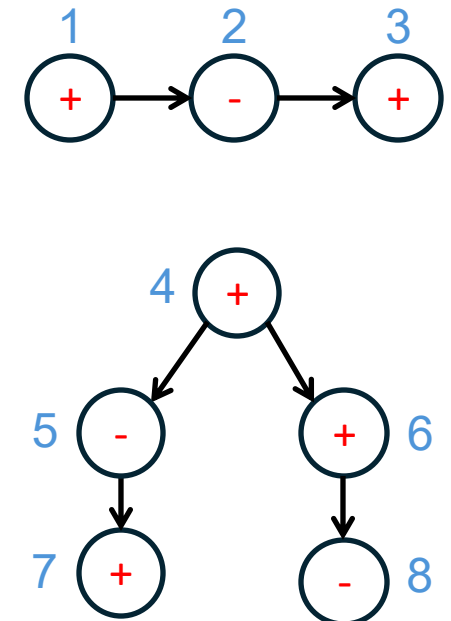
Illustrative example

- Suppose we know the underlying labels (disease status) on a forest
- How good is this sequence of testing?
 - Early detection \rightarrow early intervention; Also, we may have sudden budget cuts

Testing budget	1	2	3	4	5	6	7	8
# positive detected	1	1	2	3	3	4	5	5

Reward metric

- Use discount factor $\beta \in (0, 1)$ to favor early discovery of positive cases
 - This sequence : $\beta^0 * 1 + \beta^1 * 0 + \beta^2 * 1 + \beta^3 * 1 + \beta^4 * 0 + \beta^5 * 1 + \beta^6 * 1 + \beta^7 * 0$

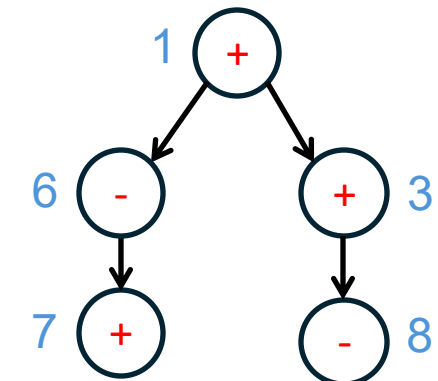
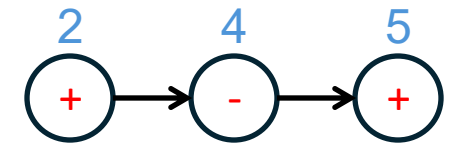


Designing a reward metric for optimization

Illustrative example

- Suppose we know the underlying labels (disease status) on a forest
- How good is this sequence of testing?
 - Early detection \rightarrow early intervention; Also, we may have sudden budget cuts

Testing budget	1	2	3	4	5	6	7	8
# positive detected	1	1	2	3	3	4	5	5
	1	2	3	3	4	4	5	5



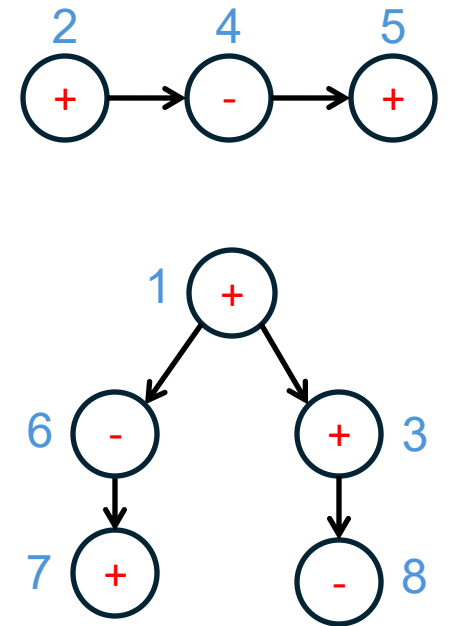
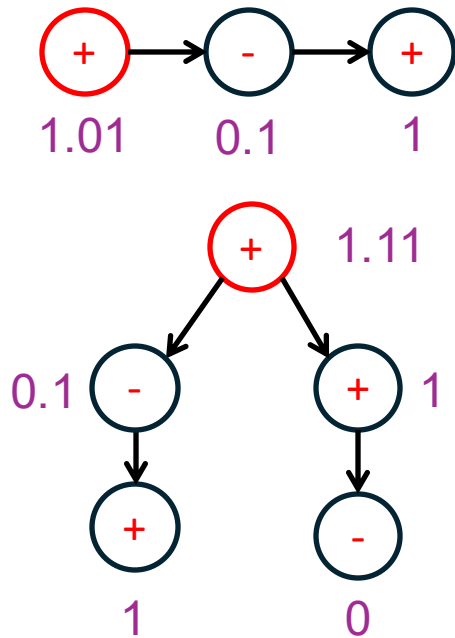
Reward metric

- Use discount factor $\beta \in (0, 1)$ to favor early discovery of positive cases
 - This sequence : $\beta^0 * 1 + \beta^1 * 0 + \beta^2 * 1 + \beta^3 * 1 + \beta^4 * 0 + \beta^5 * 1 + \beta^6 * 1 + \beta^7 * 0$
 - A better sequence: $\beta^0 * 1 + \beta^1 * 1 + \beta^2 * 1 + \beta^3 * 0 + \beta^4 * 1 + \beta^5 * 0 + \beta^6 * 1 + \beta^7 * 0$
- Finding a sequence to optimize this reward metric helps detect positive cases faster, for any fixed amount of testing budget

Gittins indices are optimal on rooted forests

Using the idea of Gittins indices for our example

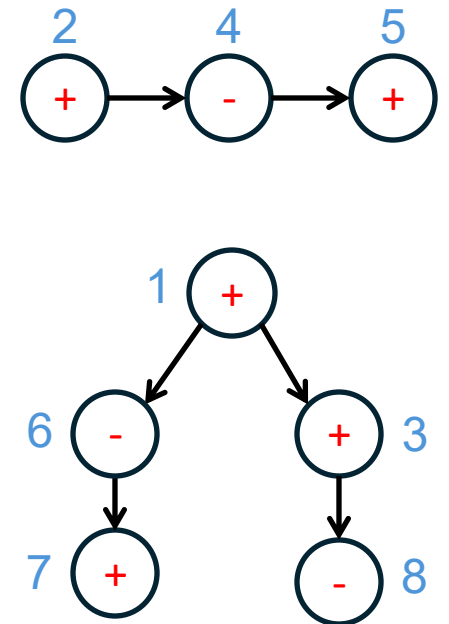
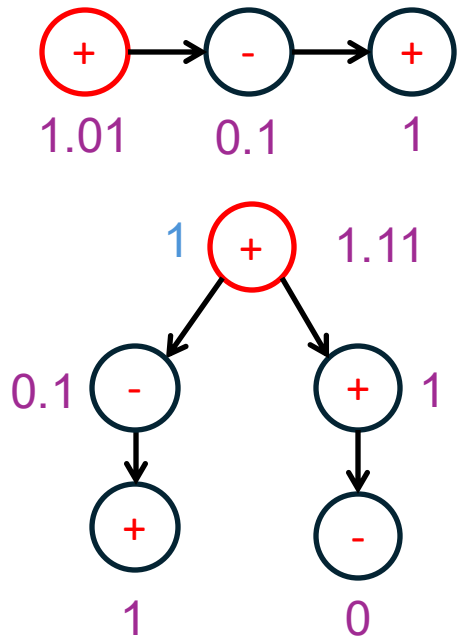
- Compute a score* for each node, then greedily select from the nodes in the frontier
- Known to produce an optimal sequence on our reward metric when the input graph is a forest



Gittins indices are optimal on rooted forests

Using the idea of Gittins indices for our example

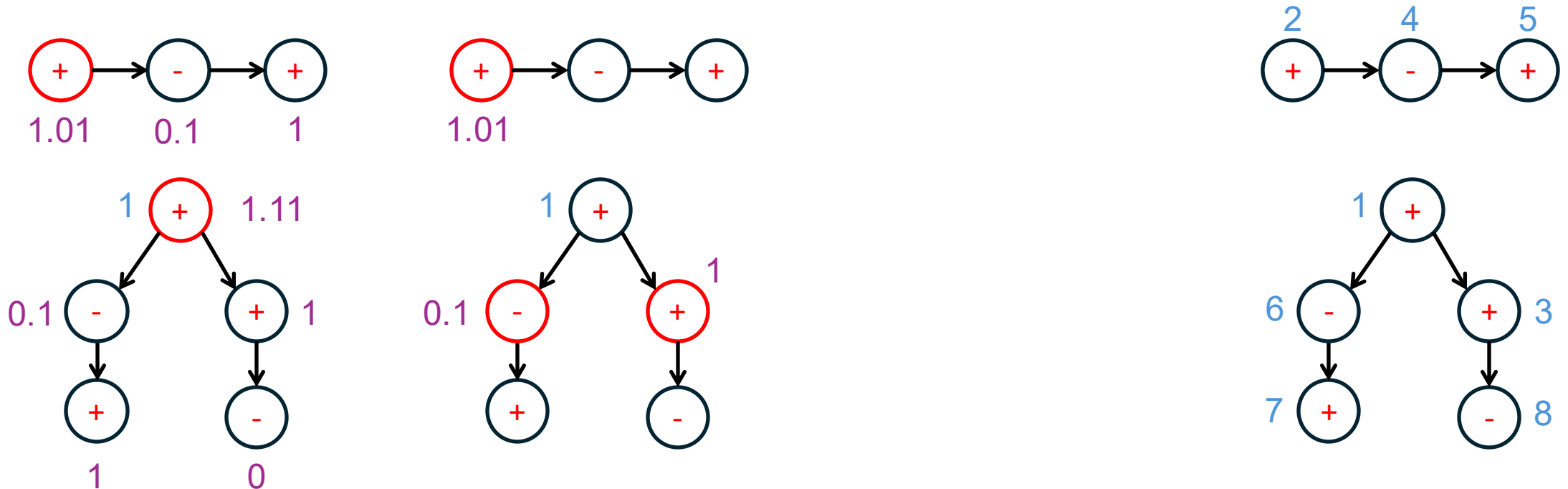
- Compute a score* for each node, then greedily select from the nodes in the frontier
- Known to produce an optimal sequence on our reward metric when the input graph is a forest



Gittins indices are optimal on rooted forests

Using the idea of Gittins indices for our example

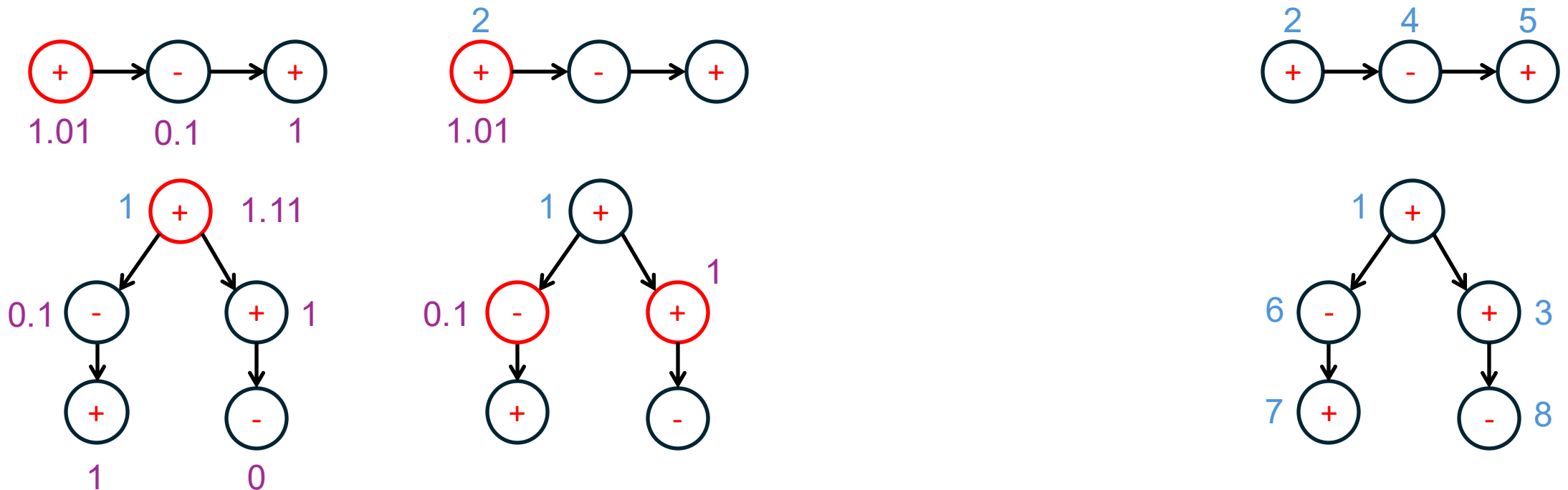
- Compute a score* for each node, then greedily select from the nodes in the frontier
- Known to produce an optimal sequence on our reward metric when the input graph is a forest



Gittins indices are optimal on rooted forests

Using the idea of Gittins indices for our example

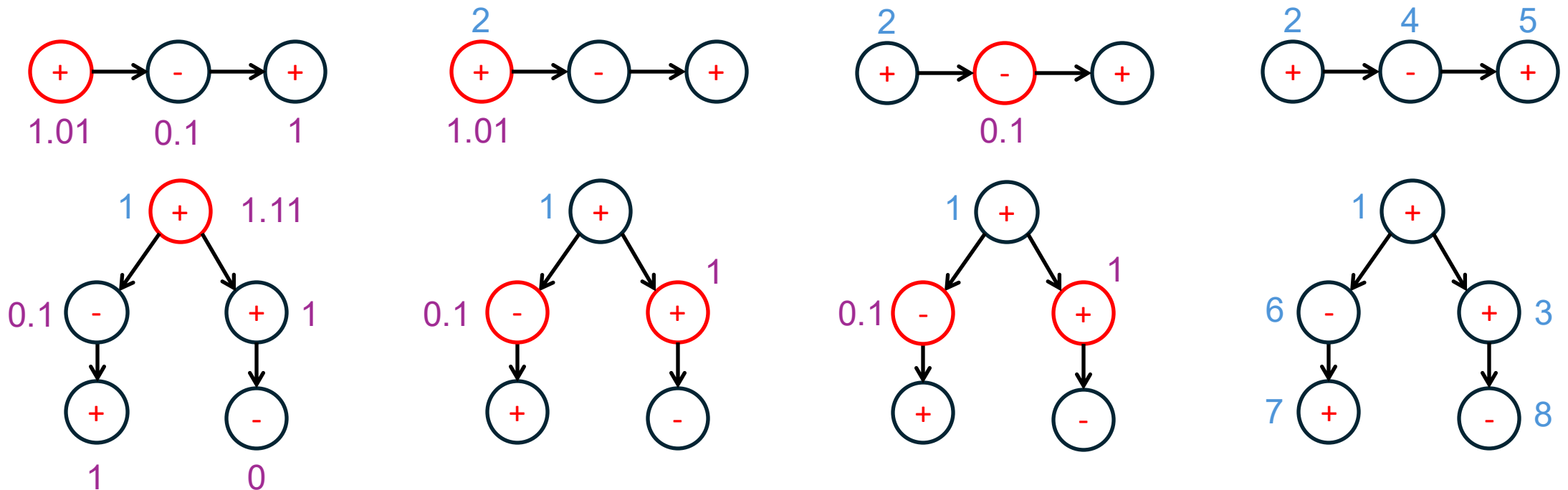
- Compute a score* for each node, then greedily select from the nodes in the frontier
- Known to produce an optimal sequence on our reward metric when the input graph is a forest



Gittins indices are optimal on rooted forests

Using the idea of Gittins indices for our example

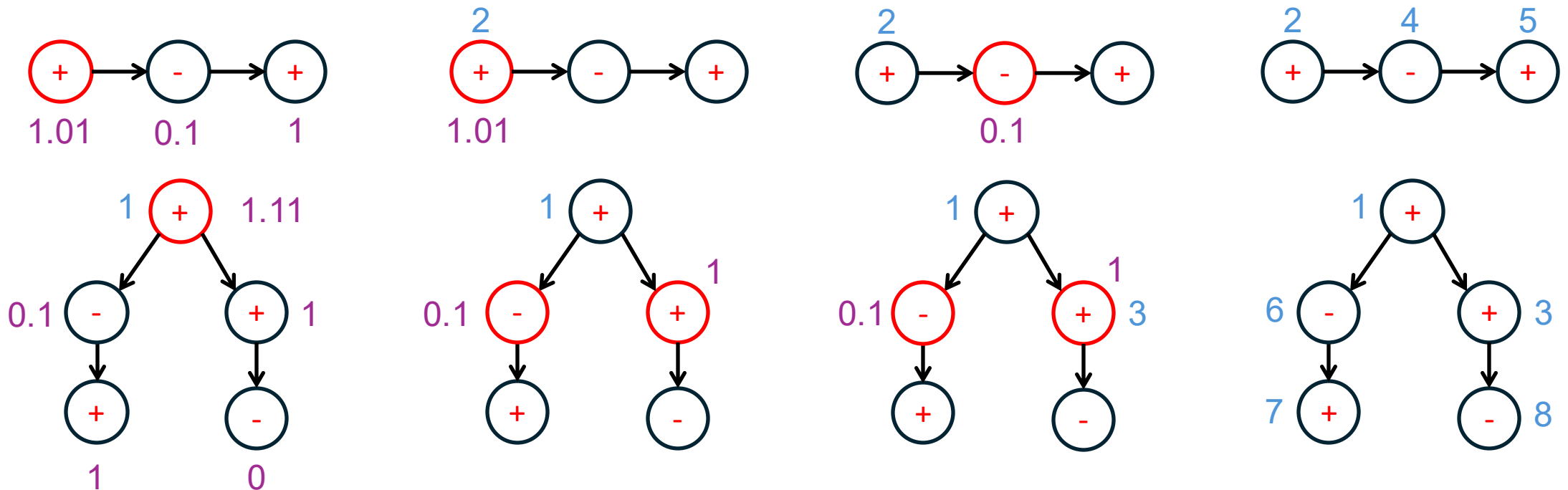
- Compute a score* for each node, then greedily select from the nodes in the frontier
- Known to produce an optimal sequence on our reward metric when the input graph is a forest



Gittins indices are optimal on rooted forests

Using the idea of Gittins indices for our example

- Compute a score* for each node, then greedily select from the nodes in the frontier
- Known to produce an optimal sequence on our reward metric when the input graph is a forest



Gittins indices are optimal on rooted forests

Using the idea of Gittins indices for our example

- Compute a score* for each node, then greedily select from the nodes in the frontier
- Known to produce an optimal sequence on our reward metric when the input graph is a forest

This approach generalizes to rooted forests and probabilistic labels

- Statuses can be a function of the revealed parent's status, based on a known distribution \mathcal{P}
- Now, we pre-compute a score for node X based on every possible parent status, then use the score that corresponds to revealed parent status when ranking node X in the frontier

Gittins indices are optimal on rooted forests

To define the Gittins index, let us first define two recursive functions ϕ and Φ , as per [KO03]. For any non-root node $X \in \mathbf{X}$, label $b \in \Omega$, and value $0 \leq m \leq \frac{\bar{r}}{1-\beta}$,

$$\phi_{X,b}(m) = \max \left\{ m, \sum_{v \in \Omega} \mathcal{P}(X = v \mid \text{Pa}(X) = b) \cdot [r(X, v) + \beta \cdot \Phi_{\text{Ch}(X),v}(m)] \right\} \quad (1)$$

If X is the root, we define $\phi_{X,\emptyset}(m) = \max \left\{ m, \sum_{v \in \Omega} \mathcal{P}(X = v) \cdot [r(X, v) + \beta \cdot \Phi_{\text{Ch}(X),v}(m)] \right\}$. For any subset of nodes $\mathbf{S} \in \mathbf{X}$, label $v \in \Omega$, and value $0 \leq m \leq \frac{\bar{r}}{1-\beta}$,

$$\Phi_{\mathbf{S},v}(m) = \begin{cases} \frac{\bar{r}}{1-\beta} - \int_m^{\frac{\bar{r}}{1-\beta}} \prod_{Y \in \mathbf{S}} \frac{\partial \phi_{Y,v}(k)}{\partial k} dk & \text{if } \mathbf{S} \neq \emptyset \\ m & \text{if } \mathbf{S} = \emptyset \end{cases} \quad (2)$$

Gittins index for node X when parent's realization is b is to be $g(X, b) = \min \left\{ m \in \left[0, \frac{\bar{r}}{1-\beta} \right] : \phi_{X,b}(m) \geq m \right\}$. This “min \geq ” captures the intuition of “fair value”

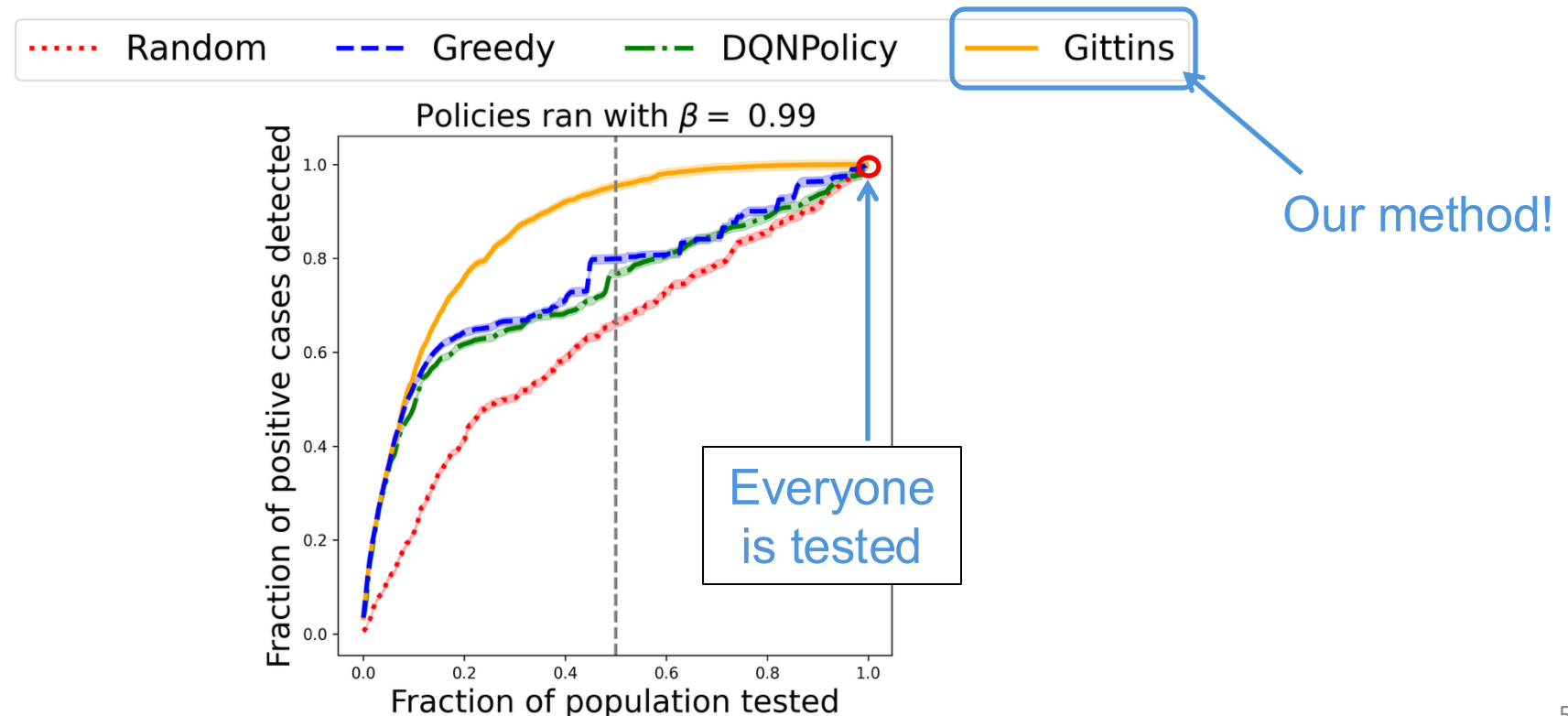
We show that our adaptive testing problem can be reduced to “branching bandits”

- Optimality of Gittins indices for rooted forests proven in [KO03]
- We provide the first efficient polynomial time dynamic programming (DP) method (with working Python code) in the 20 years since [KO03] for computing ϕ and Φ for *discrete labels*
 - Prove and exploit piecewise linearity of ϕ and Φ , enabling efficient representation of ϕ and Φ in the DP
 - Number of pieces scales well with the number of nodes and number of labels

Empirical evaluation* on HIV interaction graph

With only budget to test half the population, Gittins detects almost all positive cases in expectation while the other methods still miss about 20% of the positive cases

- Reduction to branching bandits, with provable optimality on tree-structured instances

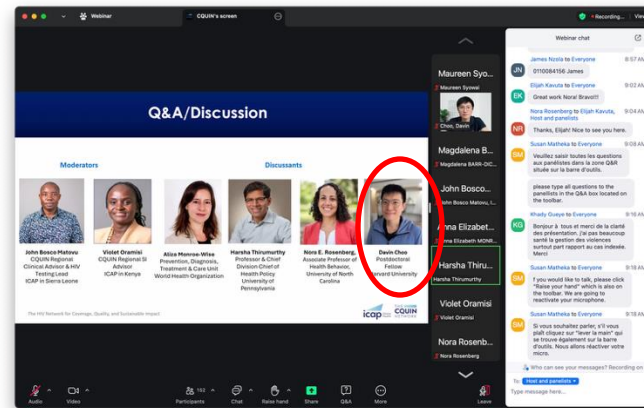


* We also have experiments on settings where policies only know a noisy version of the underlying distribution \mathcal{P}

Towards real-world impact

Invited speaker and panelist for CQUIN and WHO webinar* on 30 Oct 2025

- CQUIN is a vibrant network of 21 countries working together to exchange best practices, innovations, evidence, and tools to optimize HIV systems and services



**Alastair
van Heerden**
University of Witwatersrand
Wits Health Consortium

Strong support from domain expert collaborator

Our shared work convinced me that computer science and public health, together, can deliver healthcare that is both more inclusive and more efficient.



* <https://cquin.icap.columbia.edu/event/closing-the-gaps-with-network-based-testing-launch-of-a-new-toolkit-to-expand-self-testing-reach/>

Towards real-world impact

What's next for the HIV project? [CPW+25]

- Working towards trial-run deployment in KwaZulu-Natal, South Africa
 - In collaboration with WHO and the Gates foundation, under the HIV LIFT project
- Comprehensive retrospective study using de-identified data from Tanzania
 - In collaboration with FHI360, under the EpiC project

Towards real-world impact

“Often, the most important step is making the right notion, defining the right notion. Once you have the right notion, you know, the rest of the theory, theorems, proofs, [and] constructions follow.”

What's next for the HIV project? [CPW+25]

- Working towards trial-run deployment in KwaZulu-Natal, South Africa
 - In collaboration with WHO and the Gates foundation, under the HIV LIFT project
- Comprehensive retrospective study using de-identified data from Tanzania
 - In collaboration with FHI360, under the EpiC project
- Formulating and improving other processes in the HIV treatment cascade



Avi Wigderson,
2023 Turing award winner

[CPW+25] Davin Choo, Yuqi Pan, Tonghan Wang, Milind Tambe, Alastair van Heerden, Cheryl Johnson. *Adaptive Frontier Exploration on Graphs with Applications to Network-Based Disease Testing*. Conference on Neural Information Processing Systems (NeurIPS), 2025.

Avi Wigderson, Turing Award Lecture: “Alan Turing: A TCS Role Model”, 2024. Available at <https://www.youtube.com/live/f2NiGO8zC1c?t=3211s>

Towards real-world impact

“Often, the most important step is making the right notion, defining the right notion. Once you have the right notion, you know, the rest of the theory, theorems, proofs, [and] constructions follow.”

What's next for the HIV project? [CPW+25]

- Working towards trial-run deployment in KwaZulu-Natal, South Africa
 - In collaboration with WHO and the Gates foundation, under the HIV LIFT project
- Comprehensive retrospective study using de-identified data from Tanzania
 - In collaboration with FHI360, under the EpiC project
- Formulating and improving other processes in the HIV treatment cascade



Avi Wigderson,
2023 Turing award winner

One other project: Facility location in Ethiopia [CTG+25]

- In collaboration with Ethiopian public health institute and Ministry of Health, and the Harvard School of Public Health. Our proposed method has a learning-augmented component where advice is pre-selected set of facility locations from domain experts

[CPW+25] Davin Choo, Yuqi Pan, Tonghan Wang, Milind Tambe, Alastair van Heerden, Cheryl Johnson. *Adaptive Frontier Exploration on Graphs with Applications to Network-Based Disease Testing*. Conference on Neural Information Processing Systems (NeurIPS), 2025.

Avi Wigderson, Turing Award Lecture: “Alan Turing: A TCS Role Model”, 2024. Available at <https://www.youtube.com/live/f2NiGO8zC1c?t=3211s>

[CTG+25] Davin Choo, Yohai Trabelsi, Fentabil Getnet, Samson Warkaye Lamma, Wondesen Nigatu, Kasahun Sime, Lisa Matay, Milind Tambe, Stéphane Verguet. *Optimizing Health Coverage in Ethiopia: A Learning-augmented Approach and Persistent Proportionality Under an Online Budget*. Special Track on AI for Social Impact at AAAI Conference on Artificial Intelligence (AAAI), 2026.



Summary of efforts

RT1: Foundational AI / ML Research

- Statistical and causal problems
- Improving subtasks in modern ML pipelines

RT2: Algorithms with Imperfect Advice

- Test-and-Act framework

RT3: AI + X

- Network-based disease testing
- Facility location with domain expert advice

Towards theory-guided, impact-driven AI

RT1: Foundational AI / ML Research

RT2: Algorithms with Imperfect Advice

RT3: AI + X



Towards theory-guided, impact-driven AI



RT1: Foundational AI / ML Research

- Develop causality-aware AI / ML methods, i.e., going beyond statistical correlations
- Provide actionable insights and degrade gracefully even when assumptions are violated
- We can also improve modern ML pipelines through principled modeling and algorithmic rigor

RT2: Algorithms with Imperfect Advice

RT3: AI + X

Towards theory-guided, impact-driven AI



RT1: Foundational AI / ML Research

- Develop causality-aware AI / ML methods, i.e., going beyond statistical correlations
- Provide actionable insights and degrade gracefully even when assumptions are violated
- We can also improve modern ML pipelines through principled modeling and algorithmic rigor

RT2: Algorithms with Imperfect Advice

- Develop theory of advice operationalization
 - What advice is useful? How should it be elicited? When is it worth refining?
- Define advice complexity measures and building human-in-the-loop systems
 - How do we formalize and principally incorporate human experts as a form of advice?

RT3: AI + X

Towards theory-guided, impact-driven AI



RT1: Foundational AI / ML Research

- Develop causality-aware AI / ML methods, i.e., going beyond statistical correlations
- Provide actionable insights and degrade gracefully even when assumptions are violated
- We can also improve modern ML pipelines through principled modeling and algorithmic rigor

RT2: Algorithms with Imperfect Advice

- Develop theory of advice operationalization
 - What advice is useful? How should it be elicited? When is it worth refining?
- Define advice complexity measures and building human-in-the-loop systems
 - How do we formalize and principally incorporate human experts as a form of advice?

RT3: AI + X

- Engage organizations, both at national and regional levels
- Interdisciplinary research is one way for algorithmic rigor to effect real-world impact

Towards theory-guided, impact-driven AI



RT1: Foundational AI / ML Research

- Develop causality-aware AI / ML methods, i.e., going beyond statistical correlations
- Provide actionable insights and degrade gracefully even when assumptions are violated
- We can also improve modern ML pipelines through principled modeling and algorithmic rigor

RT2: Algorithms with Imperfect Advice

- Develop theory of advice operationalization
 - What advice is useful? How should it be elicited? When is it worth refining?
- Define advice complexity measures and building human-in-the-loop systems
 - How do we formalize and principally incorporate human experts as a form of advice?

RT3: AI + X

- Engage organizations, both at national and regional levels
- Interdisciplinary research is one way for algorithmic rigor to effect real-world impact

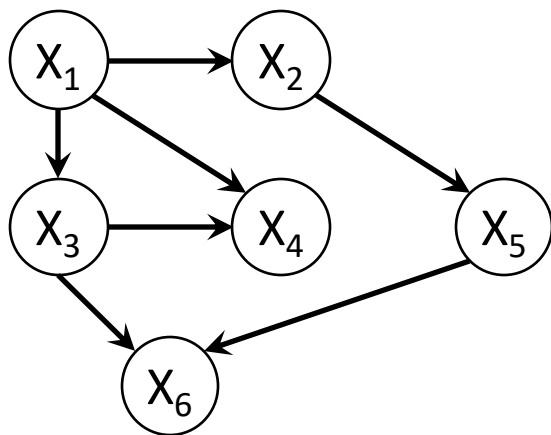
Thank you for your kind attention!

Back up slides

A glimpse of [BCG+22]

Insight: If network's in-degree is bounded, we can use less samples

- Suppose we get i.i.d. samples from a linear DAG with Gaussian noise



$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} + \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix}$$

$$X_i = \begin{cases} \eta_i + \sum_{X_j \in \text{Pa}(X_i)} a_{i,j} X_j & \text{if } \text{Pa}(X_i) \neq \emptyset \\ \eta_i & \text{if } \text{Pa}(X_i) = \emptyset \end{cases}$$

$$\mathbf{X} = \mathbf{A}\mathbf{X} + \boldsymbol{\eta}$$

$$\Rightarrow \mathbf{X} = (\mathbf{I}_n - \mathbf{A})^{-1} \boldsymbol{\eta}$$

$\Rightarrow \mathbf{X}$ is a multivariate Gaussian, in general

\Rightarrow Need $\tilde{\Omega}\left(\frac{n^2}{\varepsilon^2}\right)$ i.i.d. samples to learn \mathbf{X} “ ε -well”

- Here, max in-degree $d = 2$



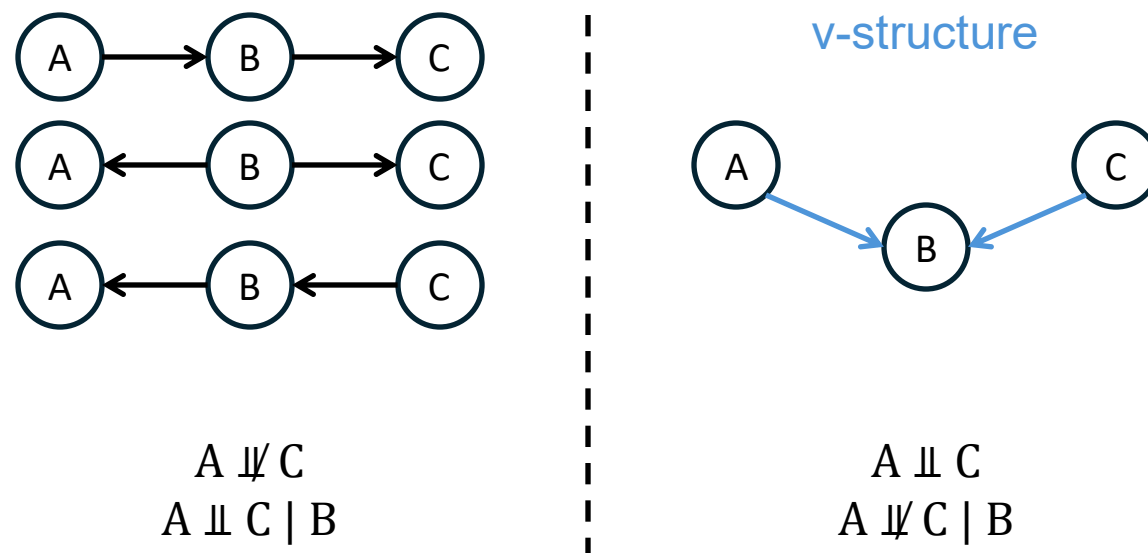
A rough intuition: All nd covariance matrix entries “matter”, in general

Turns out $\tilde{O}\left(\frac{nd}{\varepsilon^2}\right)$ samples suffices: Least squares at each node + careful analysis

A glimpse of [CSB22]

With just observational data, there are statistically indistinguishable models

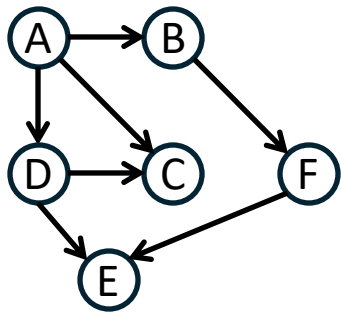
- Even with infinite observational data, can only determine causal graph up to some equivalence class where all conditional independence relations agree. E.g., $X_1 \rightarrow X_2$ vs. $X_1 \leftarrow X_2$ earlier



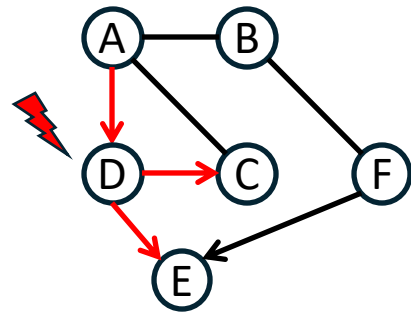
Interventions can help recover causal graphs

Meek rules [Mee95] are sound and complete for maximizing information about arc orientations in DAGs

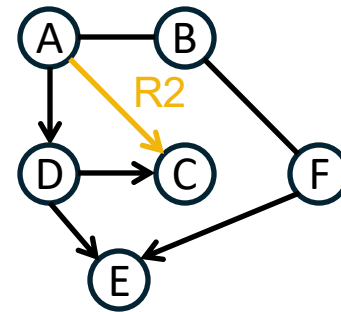
- Converge in polynomial time [WBL21]



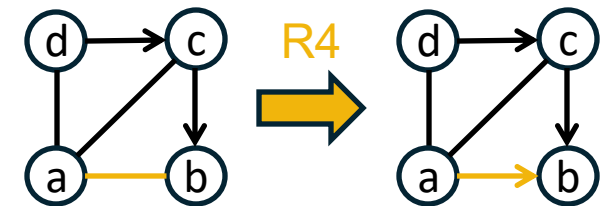
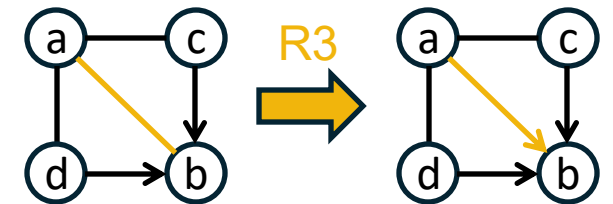
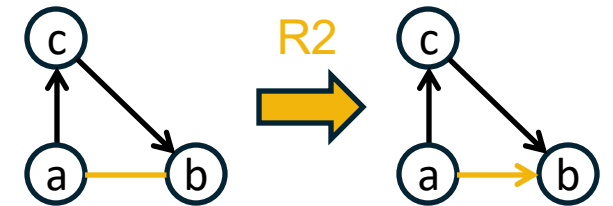
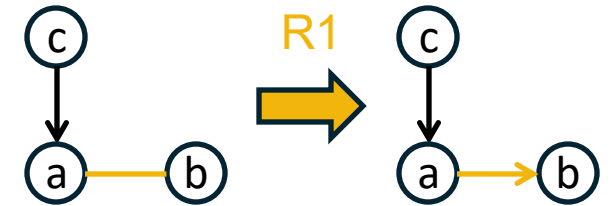
True underlying DAG



After intervening on D



Apply Meek rules



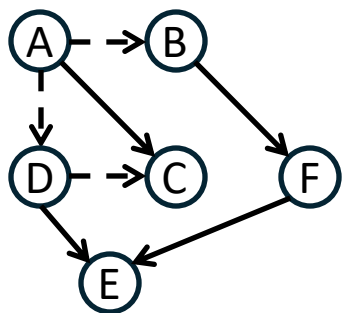
Characterization of the verification number

Trivial bound on searching and verifying

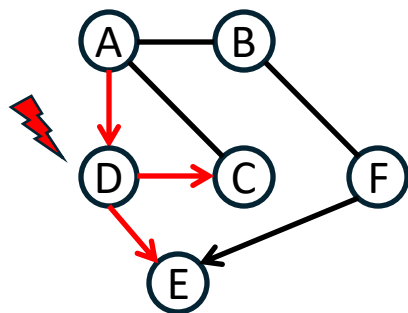
- Let $MVC(.)$ represent “minimum vertex cover”
- Always works: Intervene on $MVC(\text{unoriented edges in the essential graph of underlying DAG})$

Verification number can be efficiently characterized by covered edges

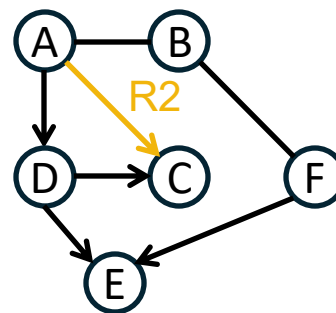
- Surprisingly, enough to compute MVC on a *subset* of edges called covered edges [Chi95]
- $u - v$ is covered edge if and only if $Pa(u) \setminus \{v\} = Pa(v) \setminus \{u\}$ in the underlying DAG (dashed)



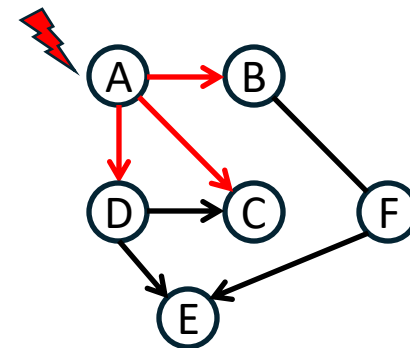
True underlying DAG



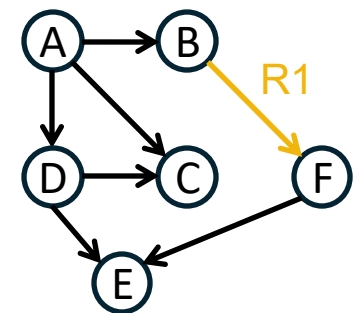
After intervening on D



Apply Meek rules



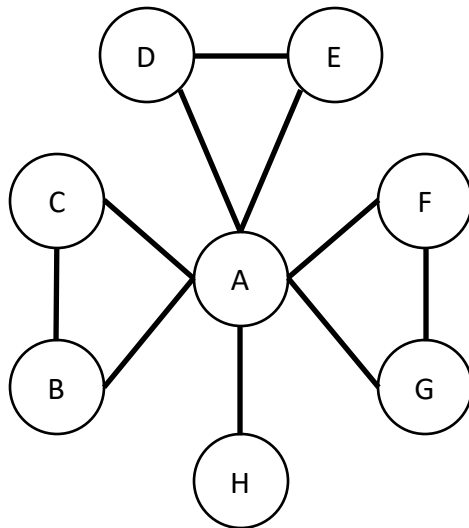
After intervening on A



Apply Meek rules

Comparison

- ← Maximal clique size
 1. $\nu(G) \geq \left\lfloor \frac{\omega(G)}{2} \right\rfloor$ Number of maximal cliques [SMG+20]
↓
 2. $\left\lfloor \frac{n-r}{2} \right\rfloor \leq \nu(G) \leq n - r$ [PSS22]



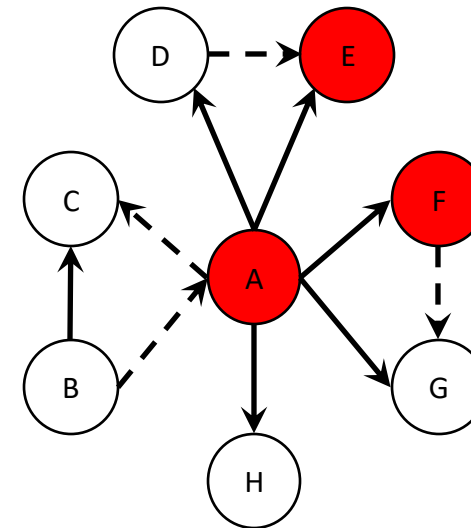
MEC $[G^*]$

$$n = 8, \omega(G) = 3, r = 4$$

1. $1 \leq \nu(G)$
2. $2 \leq \nu(G) \leq 4$

We can compute
exact $\nu(G)$ for any
 given $G \in [G^*]$

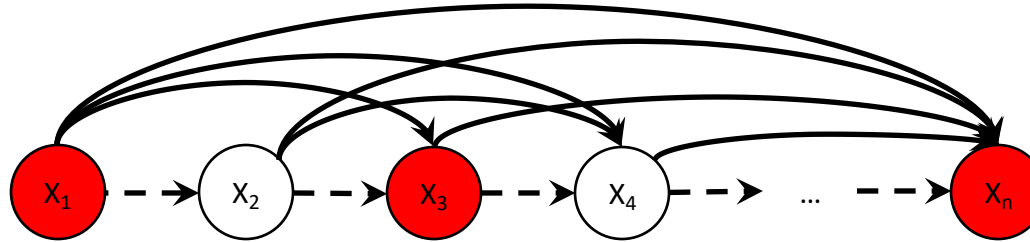
In fact, $\nu(G) \in \{3, 4\}$
 for any $G \in [G^*]$



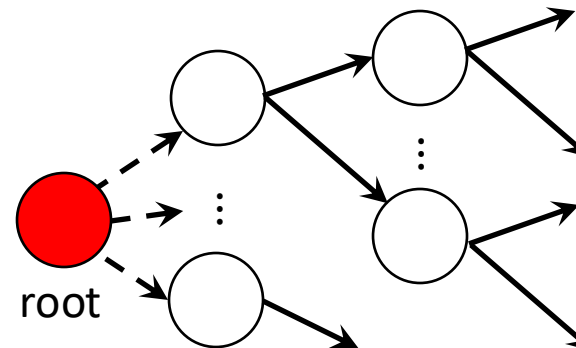
One possible DAG from $[G^*]$

Through the lens of covered edges

- Covered edges cannot have both endpoints as sink of any maximal clique, so $\nu(G) \leq n - r$
- G is a clique \Rightarrow Prior work: $\nu(G) = \left\lfloor \frac{n}{2} \right\rfloor$



- G is a tree \Rightarrow
Prior work: $\nu(G) = 1$



A glimpse of [CSB22]

With just observational data, there are statistically indistinguishable models

Interventions can help distinguish between causal models

Verifying and searching causal DAGs using *adaptive* interventions

- Let MVC represent “minimum vertex cover” and $\nu(G^*)$ be “verification number”
- Trivial search solution: Intervene on MVC of unoriented edges
- [CSB22]: **Exact characterization** of verification number $\nu(G^*)$ via MVC of covered edges [Chi95]
 - Proof: Simple (but subtle) induction using the notion of covered edges and Meek rules
 - Necessary: If not MVC of covered edges, there are two causal DAGs that are consistent with data
 - Sufficient: If MVC of covered edges, always guaranteed to recover the true underlying DAG
 - **Implication**: Easy re-interpretation of known facts and improves on prior bounds on $\nu(G^*)$
 - Claim: Covered edges form a forest, so MVC is linear time computable via dynamic programming

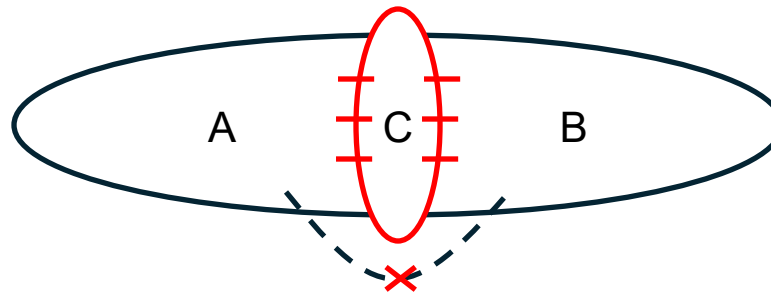
A glimpse of [CSB22]

With just observational data, there are statistically indistinguishable models

Interventions can help distinguish between causal models

Verifying and searching causal DAGs using *adaptive* interventions

- Let MVC represent “minimum vertex cover” and $\nu(G^*)$ be “verification number”
- Trivial search solution: Intervene on MVC of unoriented edges
- [CSB22]: **Exact characterization** of verification number $\nu(G^*)$ via MVC of covered edges [Chi95]
- [CSB22]: $\mathcal{O}(\nu(G^*) \cdot \log n)$ *adaptive* interventions suffice for searching. Worst case necessary.
 - Use chordal graph separators [GRE84] on unoriented part. $|A|, |B| \leq \frac{|G|}{2}$. Recurse $\mathcal{O}(\log n)$ times.
 - Proof involves proving stronger lower bound on $\nu(G^*)$ than what was known previously in the literature



[CSB22] Davin Choo, Kirankumar Shiragur, Arnab Bhattacharyya. *Verification and search algorithms for causal DAGs*. Conference on Neural Information Processing Systems (NeurIPS), 2022.

[Chi95] David Maxwell Chickering. *A Transformational Characterization of Equivalent Bayesian Network Structures*. Uncertainty in Artificial Intelligence (UAI), 1995

[GRE84] John R. Gilbert, Donald J. Rose, Anders Edenbrandt. *A Separator Theorem for Chordal Graphs*. SIAM Journal on Algebraic Discrete Methods, 1984.

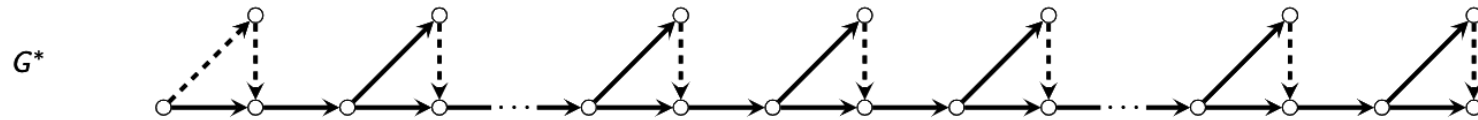
A

lower bound

Intuition [HB12,14]: In any interventional essential graph, interventions across different “connected components” *do not* help.

Claim: Fix an essential graph and some DAG G in it. Then,

$$\nu(G) \geq \sum_{\substack{\text{connected components} \\ H \in \text{after removing oriented arcs}}} \left\lfloor \frac{\omega(H)}{2} \right\rfloor$$



$$\text{Lower bound from claim: } \nu(G^*) \geq \left\lfloor \frac{3}{2} \right\rfloor = 1$$

But, from our covered edge characterization, we know that $\nu(G^*) \approx \frac{n}{2}$

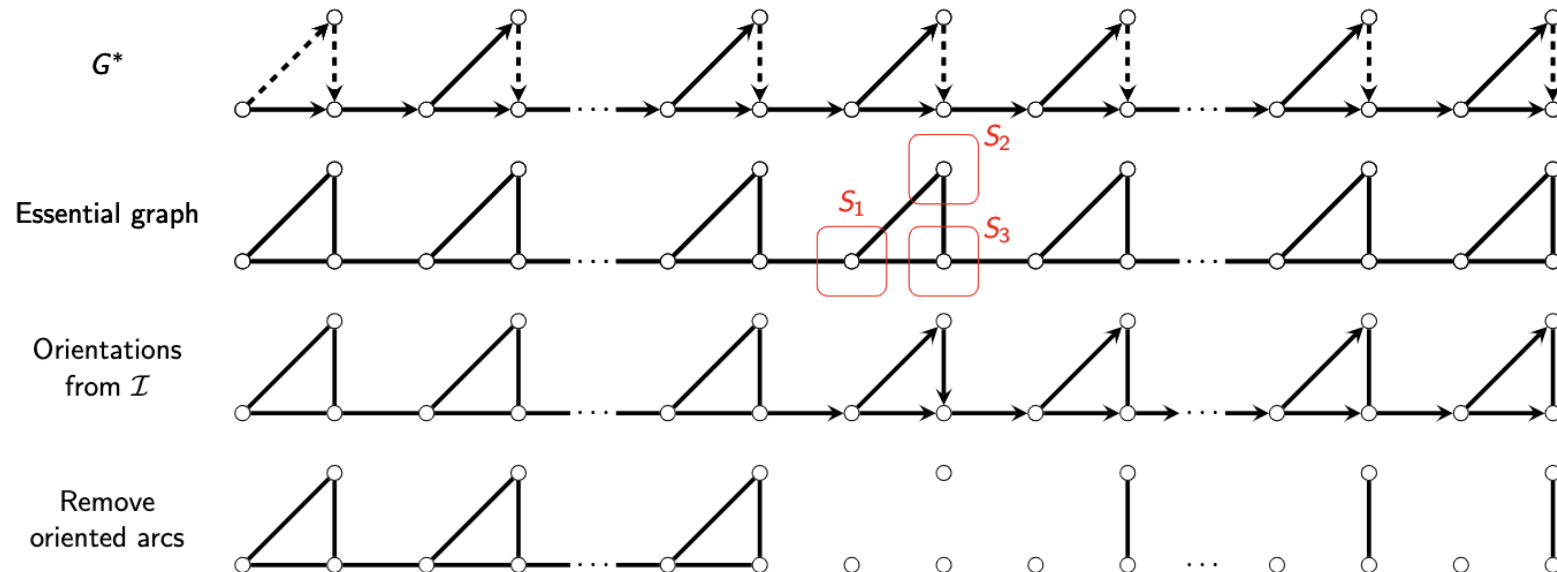
A stronger (but not computable) lower bound

Intuition [HB12,14]: In any interventional essential graph, interventions across different “connected components” *do not* help.

Claim: Fix an essential graph and some DAG G in it. Then,

$$\nu(G) \geq \sum_{\substack{\text{connected components} \\ H \in \text{after removing oriented arcs} \\ \text{after interventions } S_1, \dots, S_t}} \left\lfloor \frac{\omega(H)}{2} \right\rfloor$$

[CSB22]
max
atomic
interventions
 S_1, \dots, S_t



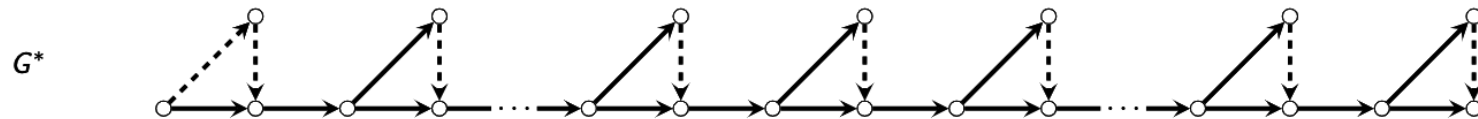
A stronger (but not computable) lower bound

Intuition [HB12,14]: In any interventional essential graph, interventions across different “connected components” *do not* help.

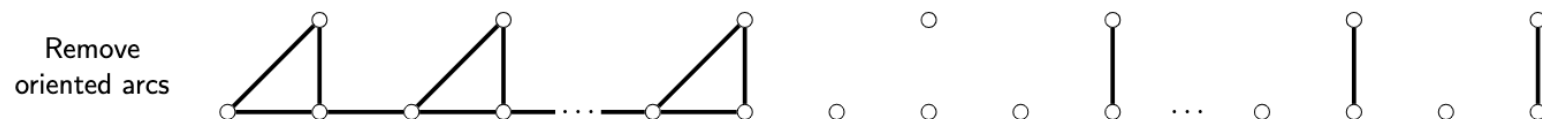
Claim: Fix an essential graph and some DAG G in it. Then,

$$\nu(G) \geq \sum_{\substack{\text{connected components} \\ H \in \text{after removing oriented arcs} \\ \text{after interventions } S_1, \dots, S_t}} \left\lfloor \frac{\omega(H)}{2} \right\rfloor$$

[CSB22]
 max atomic interventions S_1, \dots, S_t



$$\nu(G^*) \geq \left\lfloor \frac{3}{2} \right\rfloor + 1 + \dots + 1 \in \Omega(n)$$



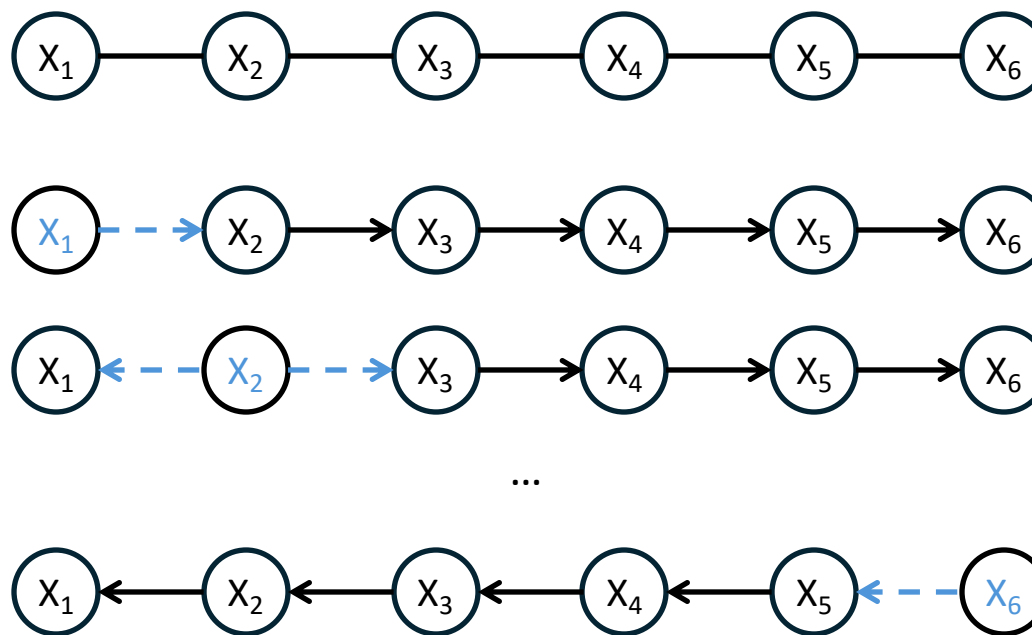
A glimpse of [CSB22]: Simple line example

Suppose there is a chain of variables with 1 single **root cause**

- No v-structure in the ground truth DAG

Fully unoriented from
observational data

$n = 6$
possible
ground truth
DAGs



In this example, **only the edges incident to the root cause are covered edges (dashed)**

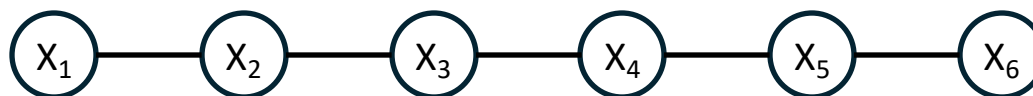
So, $\nu(G^*) = 1$ for all possible ground truth DAGs

A glimpse of [CSB22]: Simple line example

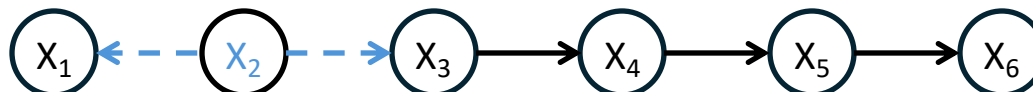
Suppose there is a chain of variables with 1 single **root cause**

- No v-structure in the ground truth DAG

Fully unoriented from
observational data



Hidden ground
truth DAG G^*



$$v(G^*) = 1$$

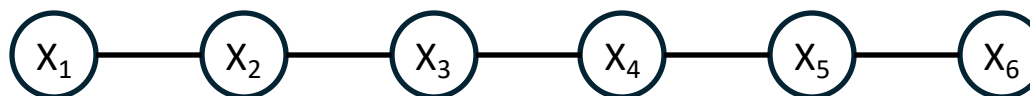
We do not see G^* or know $v(G^*)$

A glimpse of [CSB22]: Simple line example

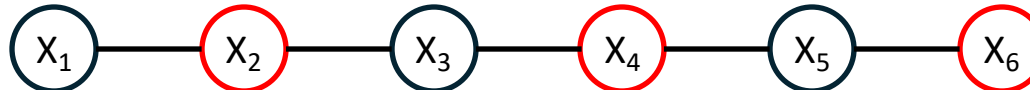
Suppose there is a chain of variables with 1 single **root cause**

- No v-structure in the ground truth DAG

Fully unoriented from
observational data

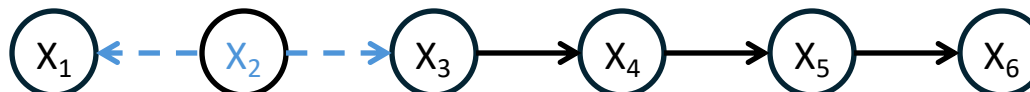


Non-adaptive
interventions



Make all intervention decisions upfront
before seeing outcomes, so we must
intervene on **MVC of unoriented edges**
since we don't know the true root cause

Hidden ground
truth DAG G^*



$$v(G^*) = 1$$

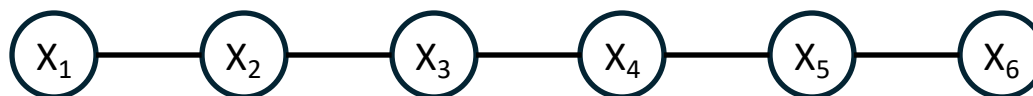
We do not see G^* or know $v(G^*)$

A glimpse of [CSB22]: Simple line example

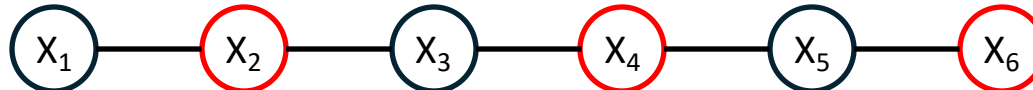
Suppose there is a chain of variables with 1 single **root cause**

- No v-structure in the ground truth DAG

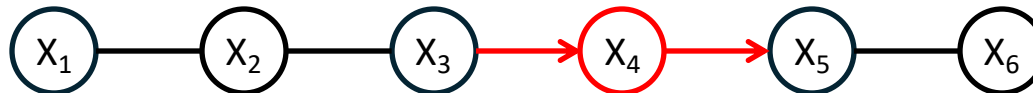
Fully unoriented from
observational data



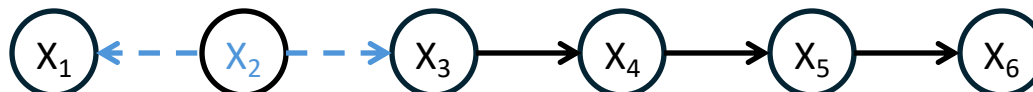
Non-adaptive
interventions



Adaptive
interventions



Hidden ground
truth DAG G^*



$$v(G^*) = 1$$

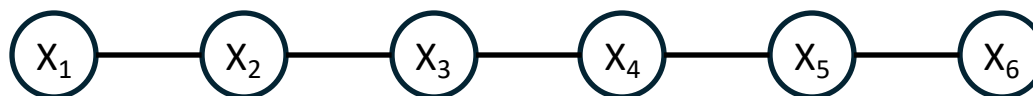
We do not see G^* or know $v(G^*)$

A glimpse of [CSB22]: Simple line example

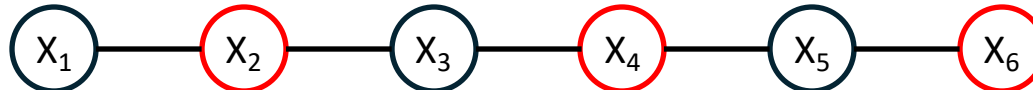
Suppose there is a chain of variables with 1 single **root cause**

- No v-structure in the ground truth DAG

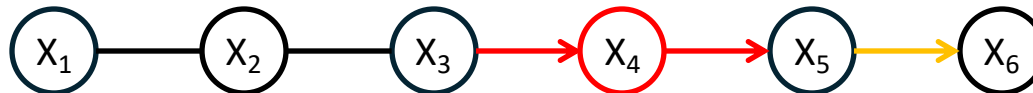
Fully unoriented from
observational data



Non-adaptive
interventions

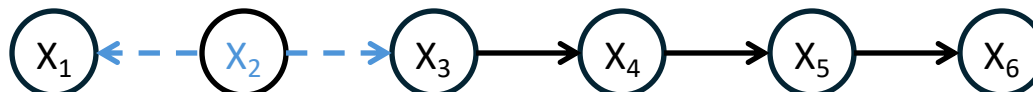


Adaptive
interventions



Because we don't
have v-structures

Hidden ground
truth DAG G^*



$$v(G^*) = 1$$

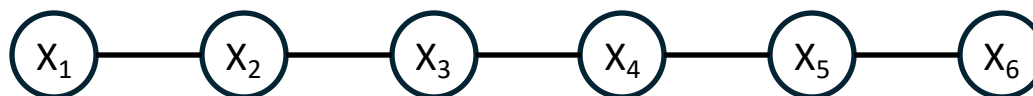
We do not see G^* or know $v(G^*)$

A glimpse of [CSB22]: Simple line example

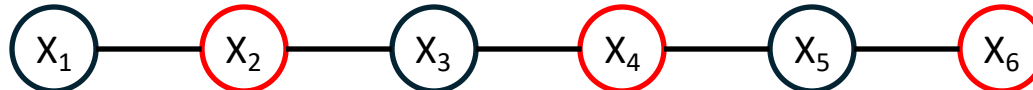
Suppose there is a chain of variables with 1 single **root cause**

- No v-structure in the ground truth DAG

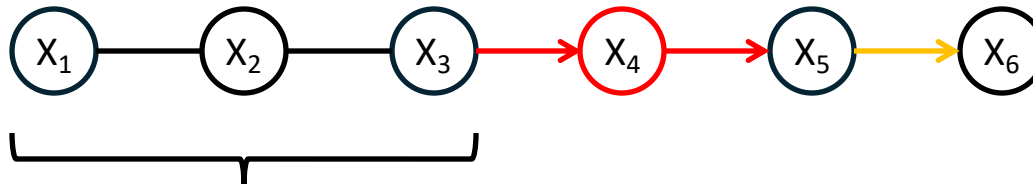
Fully unoriented from
observational data



Non-adaptive
interventions



Adaptive
interventions

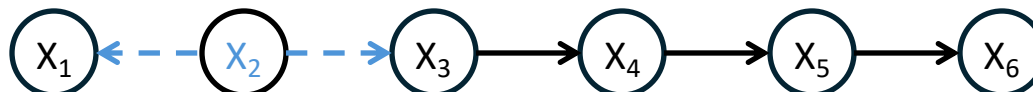


Non-adaptive:
 $\approx \frac{n}{2}$ interventions

Exponential gap!

Recurse on left half, just like in binary search: $\approx \log n$ adaptive interventions

Hidden ground
truth DAG G^*



$$v(G^*) = 1$$

We do not see G^* or know $v(G^*)$

A glimpse of [BCGG24]

Learning Gaussians: $\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$

Tolerant testing Gaussians: $\tilde{\Theta}\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$ i.i.d. samples from $N(\mu, I)$

[BCGG24]: Efficient algorithm for improving sample complexity given advice $\tilde{\mu} \in \mathbb{R}^d$

- Why ℓ_1 -norm?
 - Short answer: It just pops out of analysis
 - Small ℓ_1 -norm \Rightarrow Difference is approximately sparse
 - Sample complexity is linear in sparsity, not in dimension d
 - If advice close in ℓ_2 -norm, linear $\Omega(d)$ sample complexity lower bounds apply unfortunately
- How to estimate ℓ_1 efficiently using ℓ_2 tolerant tester?
 - Naively using $\|x\|_2 \leq \|x\|_1 \leq \sqrt{d} \cdot \|x\|_2$ does not escape $\Omega(d)$ sample complexity
 - Idea: Exploit independence in the coordinates in the mean vector
 - Estimate ℓ_1 of length k chunks, then optimize the parameters in analysis
 - Similar idea works for covariance matrix
 - Disjoint principal submatrices define independent Gaussians of smaller dimensions

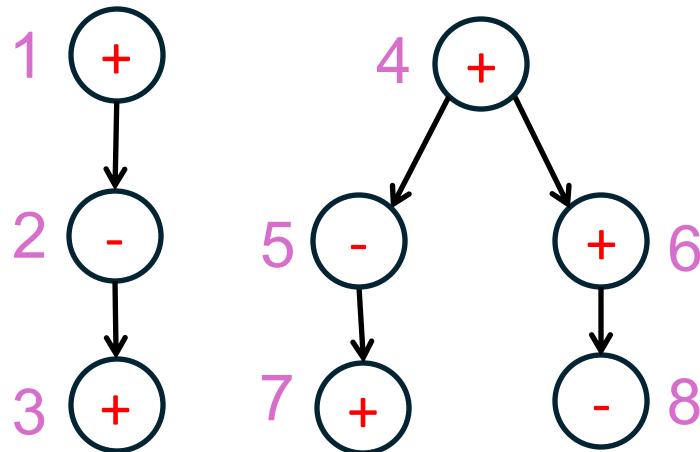
$$\tilde{\Theta}\left(\frac{d}{\varepsilon^2}\left(\min\left\{1, \frac{\|\mu - \tilde{\mu}\|_1^2}{\varepsilon^2 d}\right\}\right)\right)$$

Example on how to compare testing sequences

We want to detect positive cases faster, for any fixed amount of testing budget

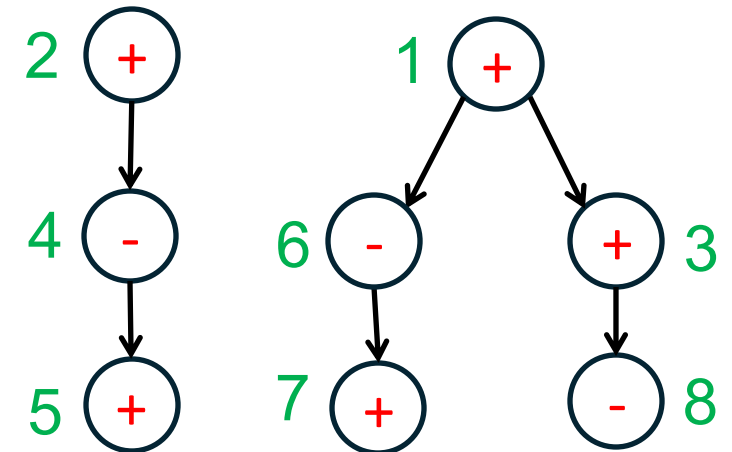
- Why? Early detection → early intervention. There may be sudden budget cuts.
- Suppose we know underlying disease statuses. Which testing sequence is better?

Testing budget	1	2	3	4	5	6	7	8
# positive detected	1	1	2	3	3	4	5	5
	1	2	3	3	4	4	5	5



Green sequence on the right is “better”.

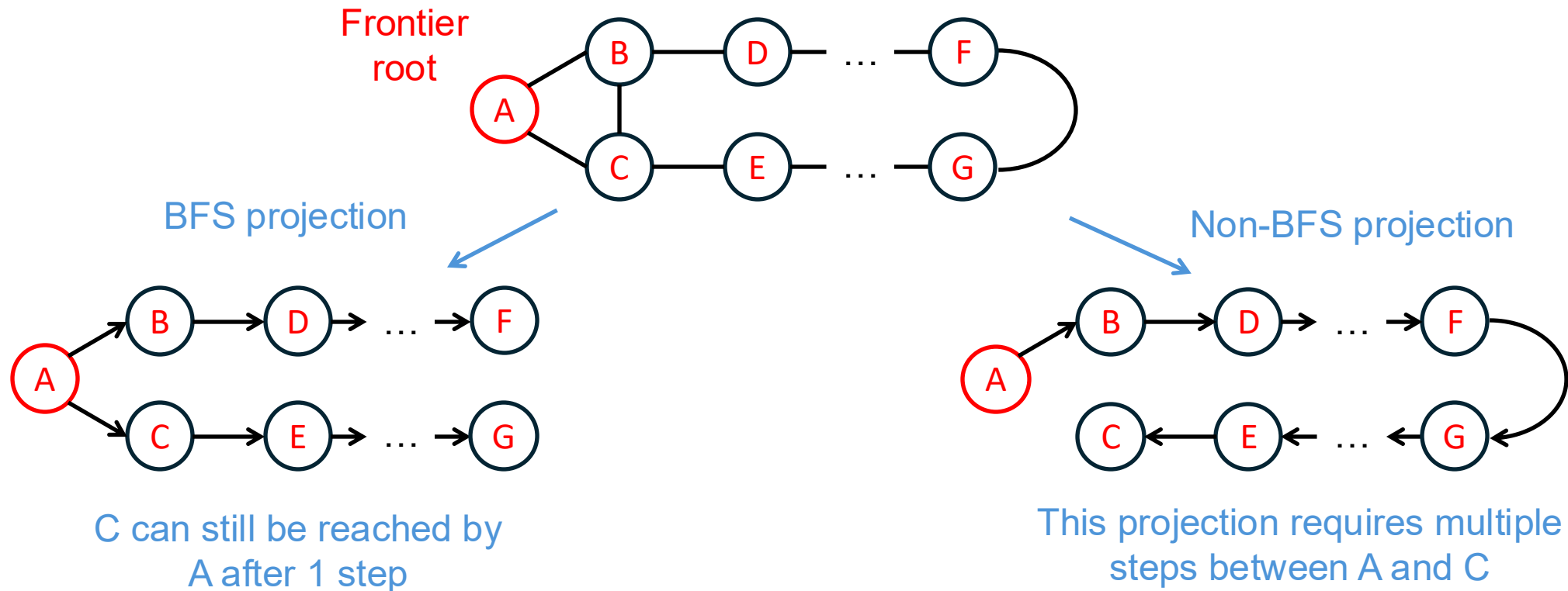
For any testing budget, it discovers same or more positive cases



In practice, use BFS to project non-trees into trees before applying Gittins computation

Transmission graphs \mathcal{G} may not be forests in general

- Run breadth-first search (BFS) on \mathcal{G} from the frontier roots in each component
- This minimizes height to root, reducing artificial frontier constraint due to projection



Leveraging on domain expertise for health facility location in Ethiopia

Can consider polynomially many subsets

High-level idea of our learning-augmented algorithm

- Run the greedy algorithm with different subsets \mathbf{A}_i of advice selection \mathbf{A} as an initial selection
- That is, maximize marginal gain when selecting next element, starting from \mathbf{A}_i instead of \emptyset
- When $\mathbf{A}_i = \emptyset$, we recover $\mathbf{U} = \mathbf{G}$
- When $\mathbf{A}_i = \mathbf{A}$, we recover $\mathbf{U} = \mathbf{A}$

$$f(\mathbf{U}) \geq \max \{ f(\mathbf{A}), f(\mathbf{G}) \}$$

Original	Greedy	Advice	Learning-augmented
3 3 5 1 5	3 3 5 1 5	3 3 5 1 5	3 3 5 1 5
3 3 2 5 1	3 3 2 5 1	3 3 2 5 1	3 3 2 5 1
1 5 10 5 1	1 5 10 5 1	1 5 10 5 1	1 5 10 5 1
1 7 5 5 5	1 7 5 5 5	1 7 5 5 5	1 7 5 5 5
3 1 3 5 3	3 1 3 5 3	3 1 3 5 3	3 1 3 5 3
Coverage = 0	Coverage = 62	Coverage = 60	Coverage = 67