BIPARTITE *B*-MATCHING

In traditional matching, any vertex can be matched at most once

***b*-matching: given G = (V,E), and a length-|V| vector *b* of nonnegative integers …**

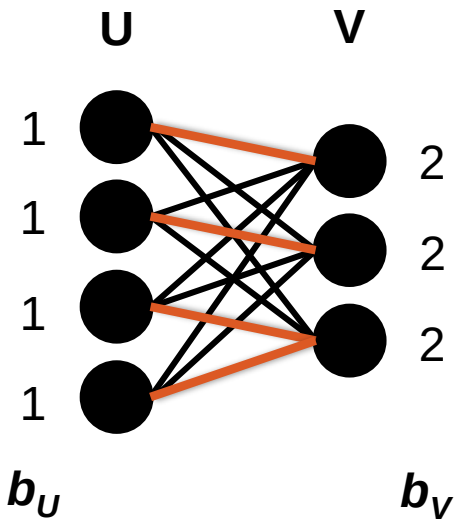
- Any vertex *i* can be matched at most *b*(*i*) times
- Generalizes traditional matching: ***b*** = **1**

Bipartite *b*-matching: given bipartite graph G = (U,V,E) ...

- PTIME for maximum cardinality/weight [Kleinschmidt et al. 1995, & earlier]

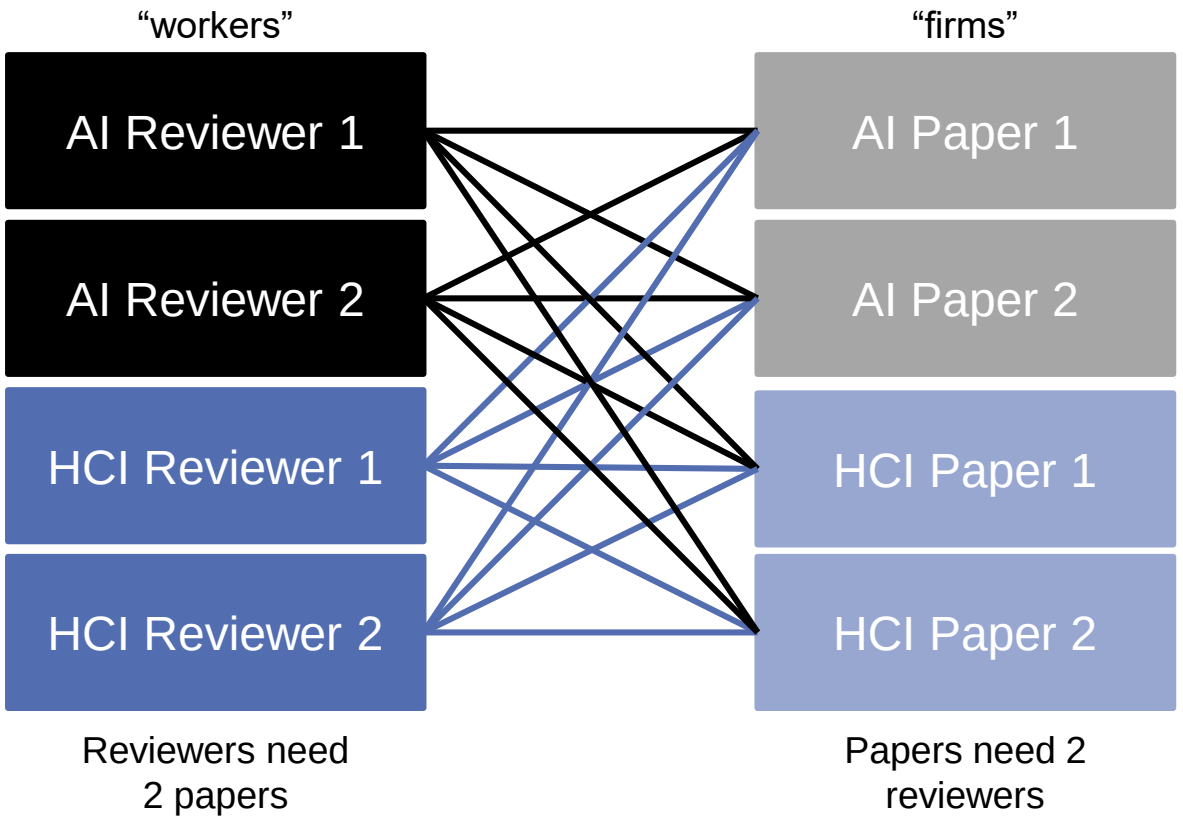
Further generalization: lower and upper bounds

- Vertex *i* must be matched at least $b_-(i)$, and at most $b_+(i)$, times
- NP-hard in many settings, even for existence



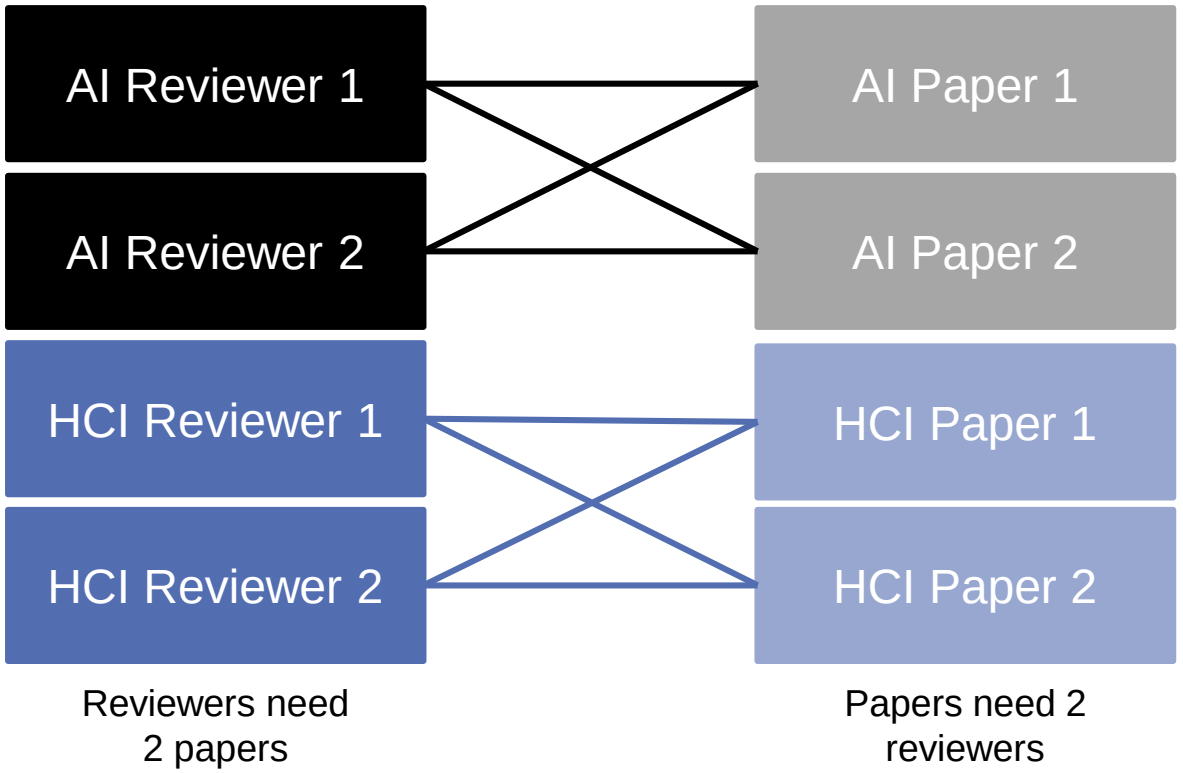
DIVERSITY IN MATCHING MARKETS

New goal: provide "good" coverage over different classes of items or agents



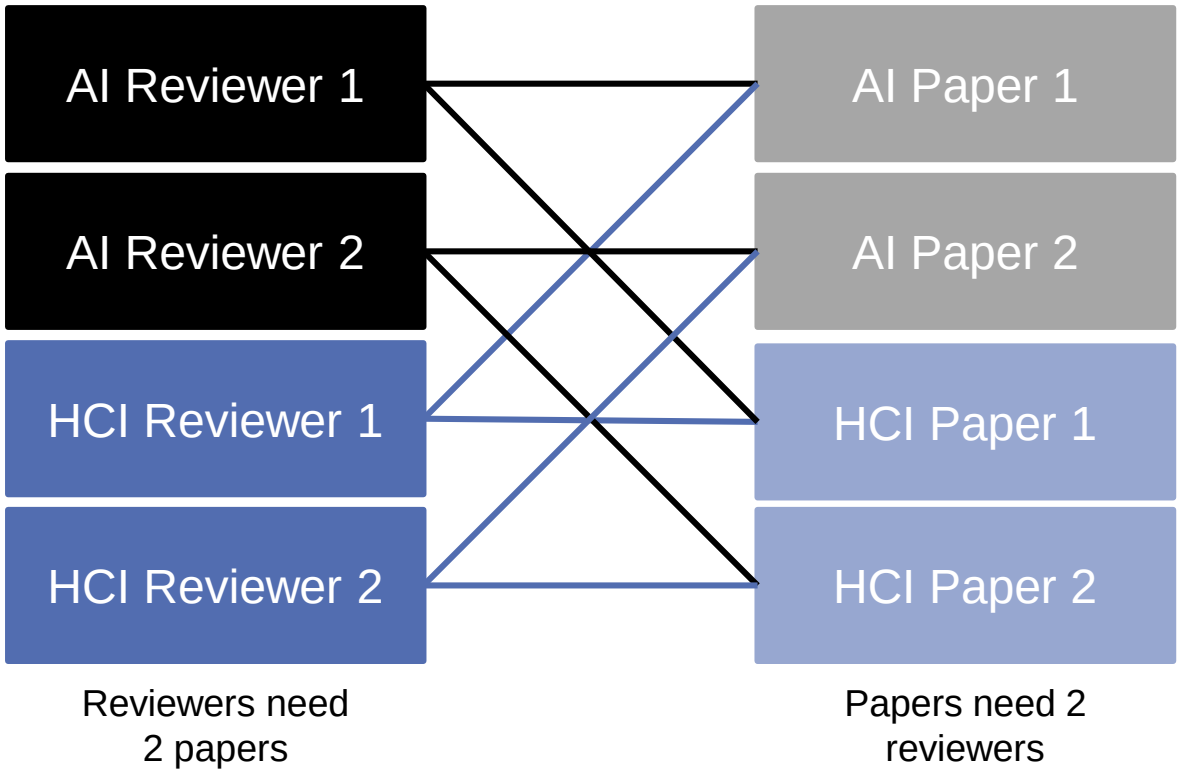
DIVERSITY IN MATCHING MARKETS

Maximum weighted matching will treat individual reviewer matchings as independent of the full review set for a paper



DIVERSITY IN MATCHING MARKETS

Maximum diverse weighted matching will balance individual quality with the diversity of opinion in the paper review set



HOW TO DEFINE DIVERSITY?

Given *K* classes on one side of the market …

- **{AI, HCI, Systems, Theory} paper classes → *K* = 4**
- **… want marginal gain of same-class matches to decrease.**

$$f_j(S_j) = \sum_{k=1}^{K} \sqrt{\sum_{\{i \mid i \in P_k \wedge (i,j) \in S_j\}} w_{i,j}}$$

Single node *j*

Full Objective

$$\max_{S \in \mathcal{S}} \sum_{i \in \text{Left}} f_i(S_i) + \sum_{j \in \text{Right}} f_j(S_j)$$

… each cluster *k* on the opposing side of the market …

… the more times node *j* is matched to nodes in cluster *k*, the lower the gain

SOLVING THIS PROBLEM

Basic maximum weight bipartite matching: PTIME

Max weight bipartite b-matching with conflict constraints: NP-hard [Chen et al. '16]

- **Integer linear program (so, ~solvable)**

Our problem: at least as hard ☹

- **Mixed integer quadratic program (so, harder)**
- **(Also, the program is enormous)**

We can show that an obvious PTIME greedy algorithm:

- **Guarantees 1 – 1/e ~ 0.63 of optimality (for some special cases)!**
- **Open question for the general case.**

ENTROPY GAIN & THE PRICE OF DIVERSITY

We use entropy to measure the gain in diversity:

- Entropy is zero if all matches come from the same cluster
- Entropy is maximized if matches are “spread evenly” across clusters
- (Edge weights, aka individual match quality, affects this.)

Entropy gain: relative gain in entropy compared to max weight

Price of diversity: relative loss in efficiency when compared to a maximum weight (aka, efficient) matching

- Want: no price of diversity with high gain in entropy!
- We show that the price of diversity can be very bad in theory ☹.

BUT WHAT ABOUT IN PRACTICE?

MovieLens 1M dataset [Harper&Konstan '16]

- One million ratings of movies (we use a standard collaborative recommender system to fill in blanks)

SIGIR and KDD reviewer bidding [Karimzadehgan&Zhai '09, Sugiyama&Kan '10]

	Solve to optimality		Solve approximately	
Dataset	**PoD**	**EG**	**PoD**	**EG**
MovieLens	0.01	1.45	0.01	1.45
SIGIR	0.08	1.63	0.17	1.60
KDD	0.06	4.28	0.07	4.28

INITIAL TAKEAWAY

We can greatly increase the diversity/coverage of a recommended matching at almost no cost to overall efficiency.

(Not in theory, but in practice, and in the **static case** …)

APPLICATION: HIRING WORKERS

INTRODUCTION

Imagine we're a company hiring a team of workers from a large pool of applicants

Wants:

- **High individual worker quality**
- **Good interplay between workers**

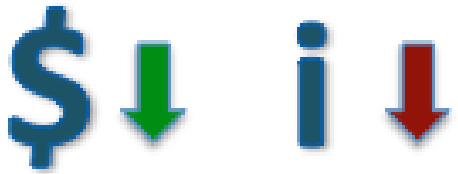
Constraints:

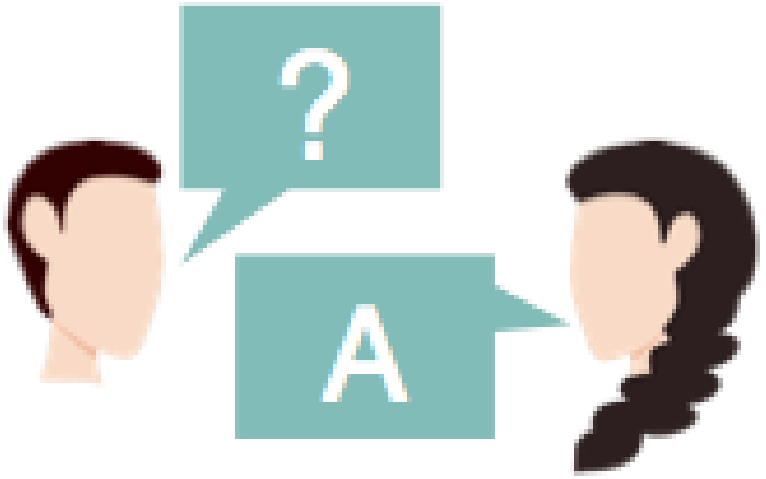
- **Interviewing budget / cost**
- **Uncertainty over individual quality**

INFORMATION GATHERING

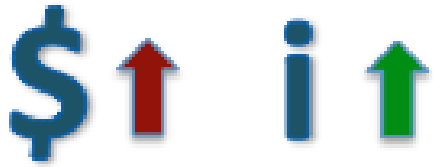
For the sake of this talk, assume we have two ways to gain information:

Resume screenings





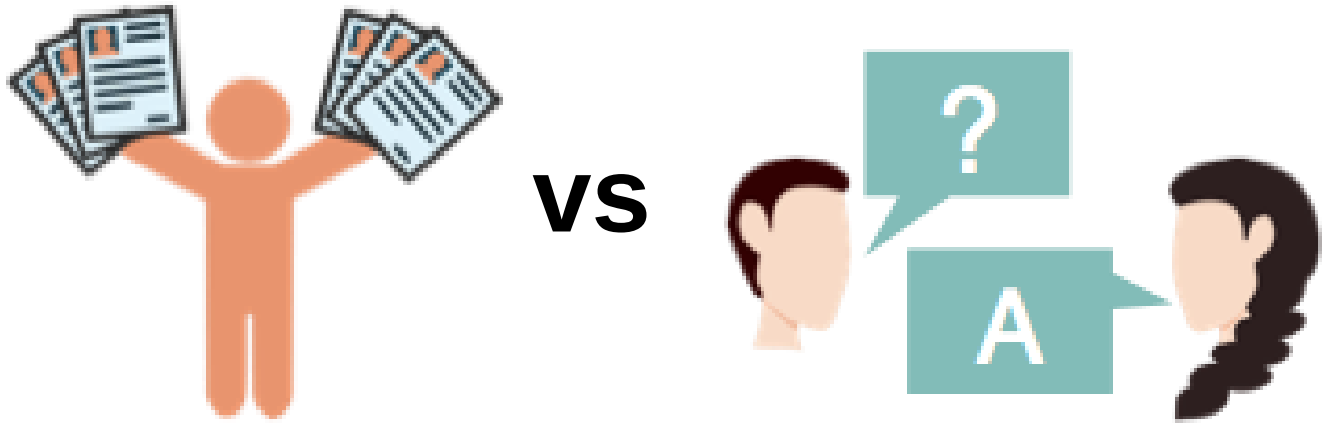
In-person interviews



KEY QUESTIONS

How should a company allocate its limited interviewing resources to select the best cohort of new employees from a large set of job applicants?

How should that company allocate cheap but noisy resume screenings and expensive but in-depth in-person interviews?



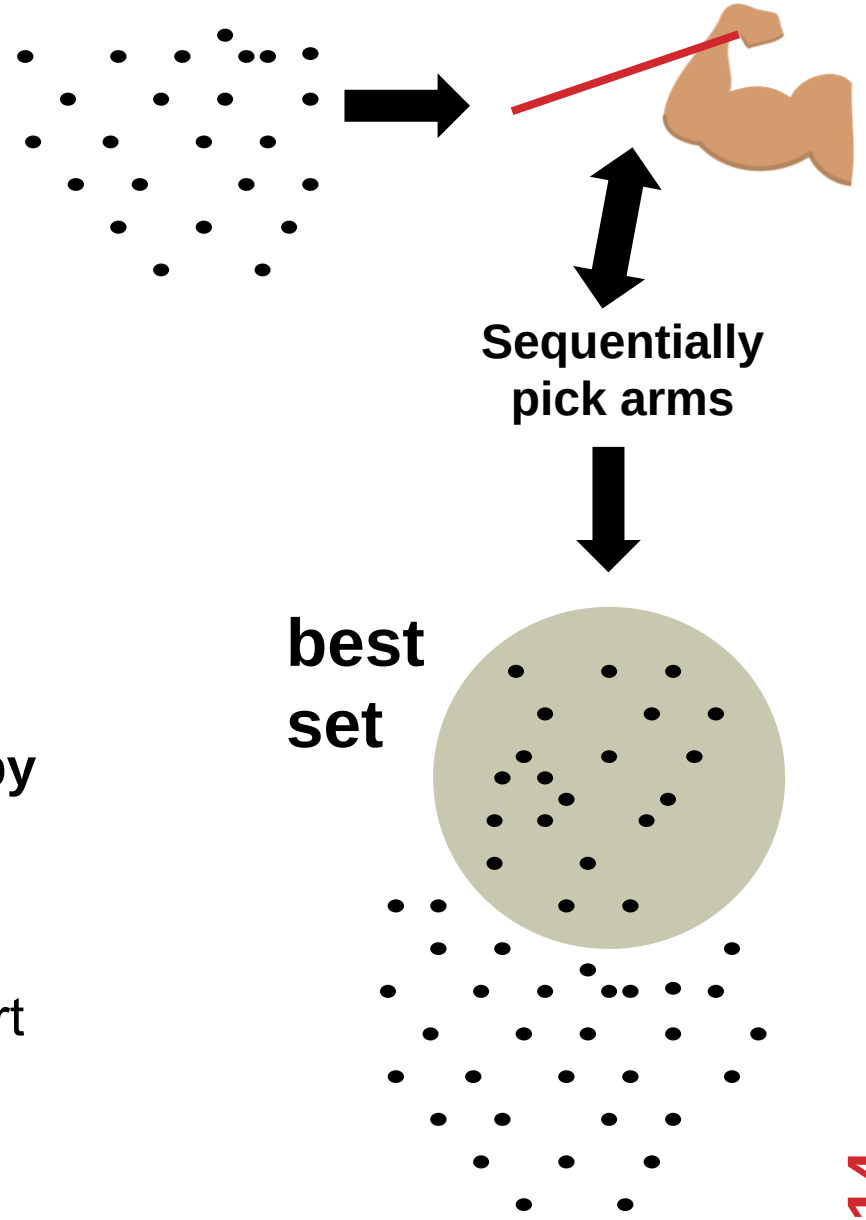
HIGH-LEVEL APPROACH

Model as a multi-armed bandit (MAB) problem in the Combinatorial Pure Exploration (CPE) setting

- Each applicant is represented by an arm
- Uncertainty over true value of that individual arm
- Can “pull” arms at some cost to gain information

Goal: find the optimal cohort maximizing some objective by selectively pulling arms:

- “Pure exploration” – pull arms first, and then select cohort
- Only care about how much effort it takes to find “best” cohort



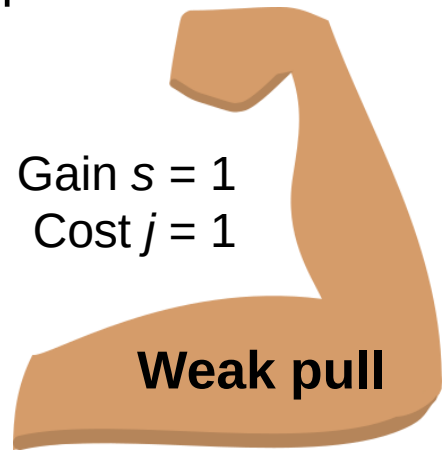
SETTING UP THE MODEL

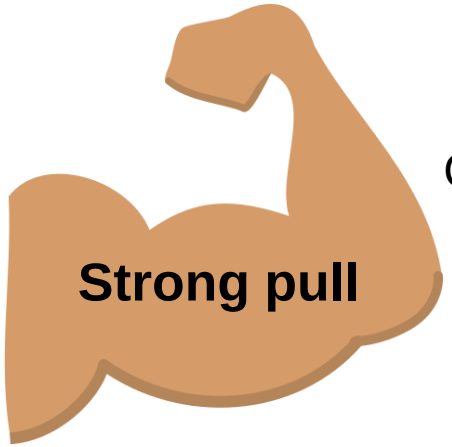
Might build on recent work due to Chen et al. [2014+]:

- Select a subset of arms with certain combinatorial structure (size-K, matching, etc)
- Looked at fixed confidence and fixed budget settings

A generalization: two ways to gather information about an arm's utility:

- Weak pull, equivalent to a resume screening of a candidate
- Strong pull, equivalent to an interview of a candidate

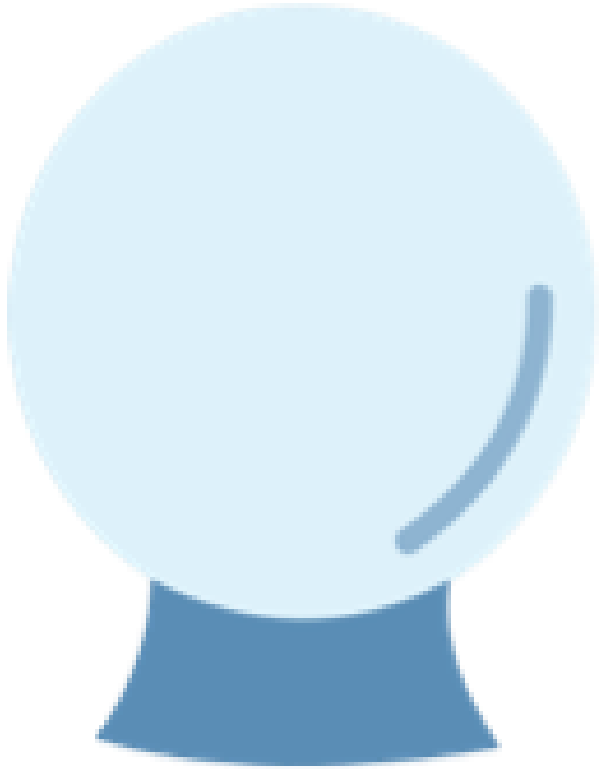




Gain s > 1
Cost j > 1

SWAP: STRONG-WEAK ARM PULL

We present the strong-weak arm pull (SWAP) algorithm, which chooses which arms to pull, and when, based on input from an oracle that maximizes a monotone submodular objective function.

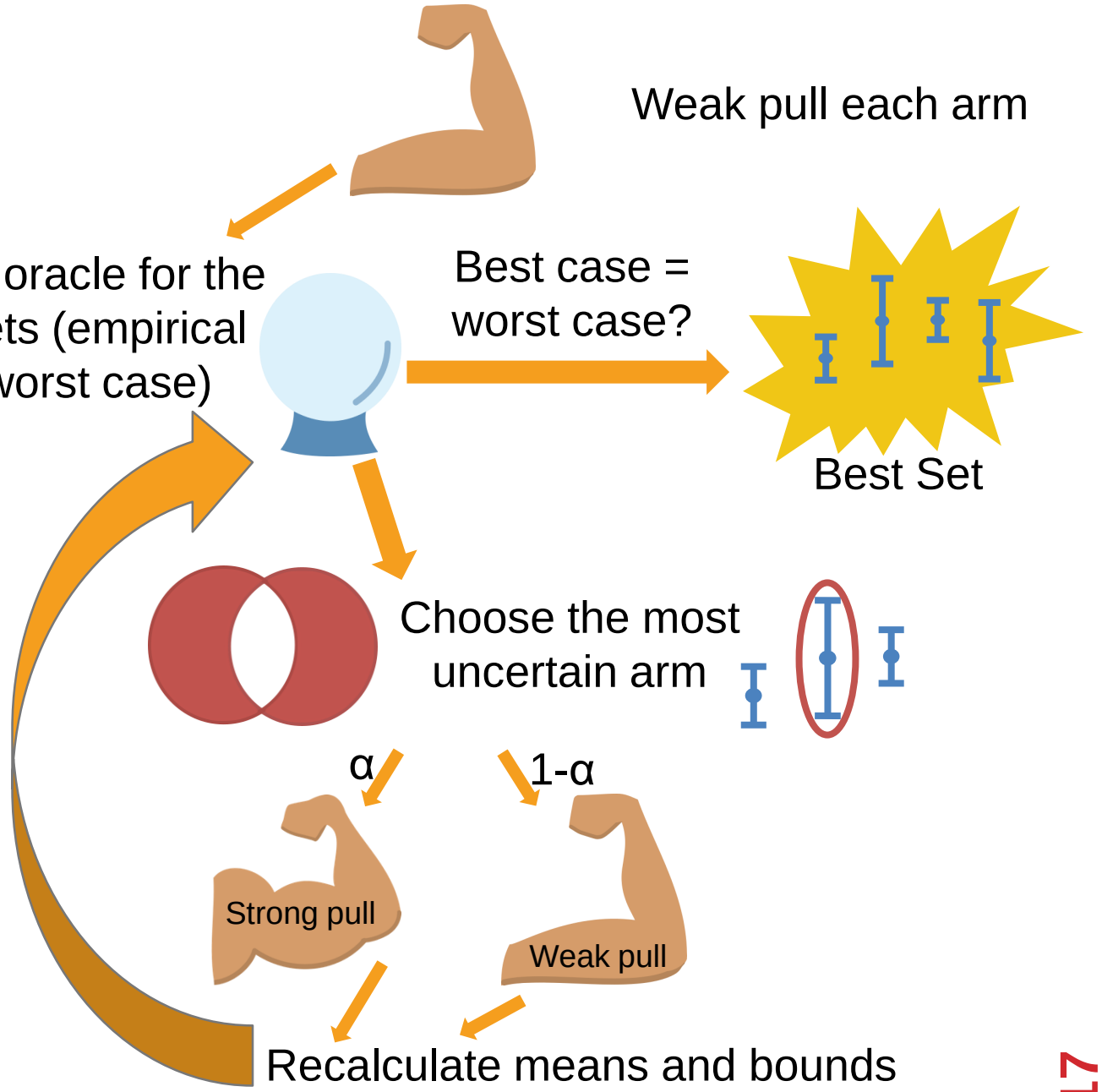


SWAP ALGORITHM

Initialize empirical means

Repeat until confident:

1. **Am I done?**
2. **Take the symmetric difference between optimistic and pessimistic cohorts**
3. **Choose the most uncertain arm from that subset of arms**
4. **Probabilistically weak/strong pull**
5. **Update estimates and repeat**



THEORETICAL RESULTS

Theorem 2. *Given any* $\delta_1, \delta_2, \delta_3 \in (0,1)$*, any decision class* $\mathcal{M} \subseteq 2^{[n]}$*, and any expected rewards* $\mathbf{u} \in \mathbb{R}^n$*, assume that the reward distribution* φ_a *for each arm* $a \in [n]$ *has mean* $u(a)$ *with an* σ*-sub-Gaussian tail. Let* $M_* = \arg\max_{M \in \mathcal{M}} w(M)$ *denote the optimal set. Set* $rad_t(a) = \sigma\sqrt{2\log\left(\frac{4nCost_t^3}{\delta}/T_t(a)\right)}$ *for all* $t > 0$ *and* $a \in [n]$*, set* $\epsilon_1 = \sigma\sqrt{2\log\left(\frac{1}{2}\delta_2/T\right)}$*, and set* $\epsilon_2 = \sigma\sqrt{2\log\left(\frac{1}{2}\delta_3/n\right)}$*. Then, with probability at least* $(1-\delta_1)(1-\delta_2)(1-\delta_3)$*, the SWAP algorithm (Algorithm 1) returns the optimal set* $\texttt{Out} = M_*$ *and*

$$T \leq O\left(\frac{\sigma^2 width(\mathcal{M})^2 \mathbf{H} \log\left(nR^2\left(\bar{X}_{Cost} - \epsilon_1\right)^3 \mathbf{H}/\delta\right)}{\bar{X}_{Gain} - \epsilon_2}\right), \quad (6)$$

where T *denotes the number of samples used by Algorithm 1,* $\mathbf{H}$ *is defined in Eq. 3*

Overview of theoretical results:

- We extend the results due to Chen et al. to the case of "arm pulls that cost *j* and give you *s*"
- We also give results for general probabilistic pulling policies like "strong pull with probability s/j and weak pull with prob. 1 – s/j"
- We give some initial results relating SWAP to other algorithms
- **These theoretical results are only for the linear case – monotone submodular is open!**

AN ONGOING (SIMULATION-BASED) EXPERIMENT WITH SWAP

Initial motivation: incorporating diversity into real-world hiring processes

- **Academic real world = graduate admissions**

Used actual admissions data from the University of Maryland's Department of Computer Science to run experiments using SWAP to simulate admissions of a diverse cohort of graduate students

- (IRB approval, & support of university and department.)

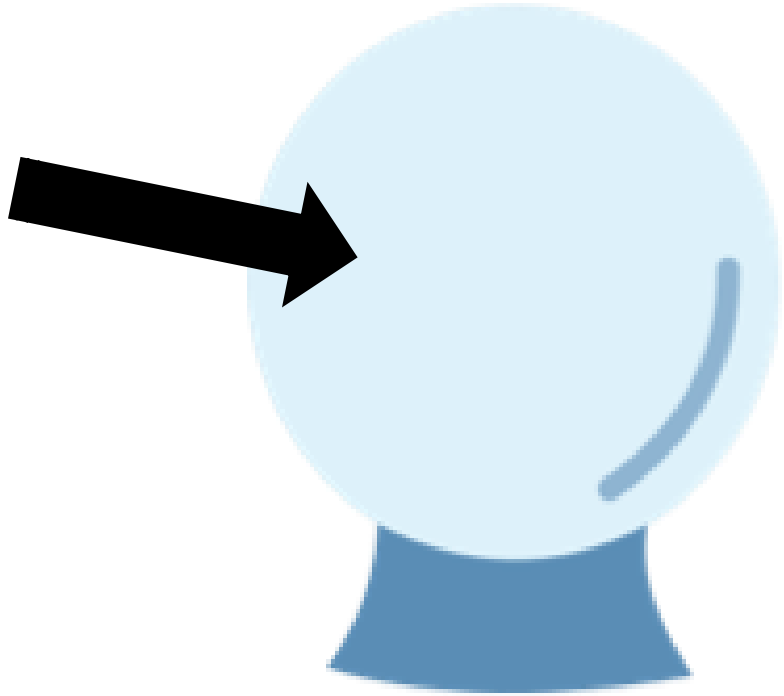
Would be interested in eventually setting up a live experiment, & happy to talk offline!

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

FORMALLY PROMOTING DIVERSITY WITH SWAP

In our experiments, we define diversity via an adaptation of the submodular function from earlier:

$$f(M) = \sum_{i=1}^{K} \sqrt{\sum_{j \in P_i \cap M} u(a)}$$



Recall: function takes K disjoint* classes as input

In our setting:

- **Applicant attributes like region or gender**
- **Applicant interests like AI, ML, HCI, and so on**

*extends to overlapping classes

GRADUATE ADMISSIONS EXPERIMENT SETUP

Trained a classifier on past admissions (2013 – 2015) data to model decisions of the graduate admissions committee

- Numerical score in [-2, +2] of quality, and actual admissions decision in {0,1}

Used different text-based and numerical features of applicants:

- Statements of purpose (OCR → word counts, LDA, and so on)
- Letters of recommendation (OCR → word counts, LDA, and so on)
- Academic data (GPA, GRE, …)
- Demographic data (Gender, region, …)
- “Standout”, “grindstone”, “ability”, and other word groups from related literature

GRADUATE ADMISSIONS EXPERIMENT SETUP, CONTINUED

Using our classifier, we ran simulations of SWAP diversifying for different sets of attributes

- Gender: {Male, Female}
- Region of Origin: {North America, China, India, Asia-Other, Middle East, Europe, Africa, Other}

Estimates of *s* and *j* from a small-scale survey of past committee members

Compared SWAP's results with the results of past admissions decisions

- Not a complete proxy for utilitarian matching!
- E.g., if the top K=100 individual applicants are in HCI, it is unlikely that the graduate chair would accept a cohort of only K=100 HCI students

Can compare against a simulated utilitarian matching based on committee scores

CAVEATS

These are not policy recommendations!

Feedback – negative and positive – is welcome and desired!

Issues:

- We are learning from past review scores, which are biased → biased classifier
- Real scores given without additional information gained from Skype interviews → our simulated Skype interviews only give a certain type of "additional information"
- When comparing against reality, may not include soft operational constraints (e.g., ideas of budget, hiring pushes for a particular year, and so on)
- Can we really even measure the "real utility" of a student …?

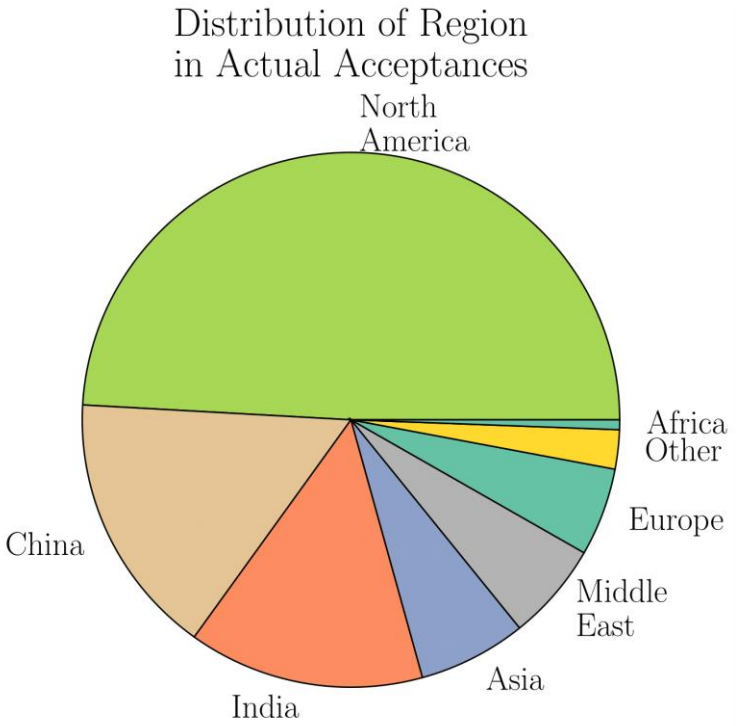
GRADUATE ADMISSIONS EXPERIMENT RESULTS

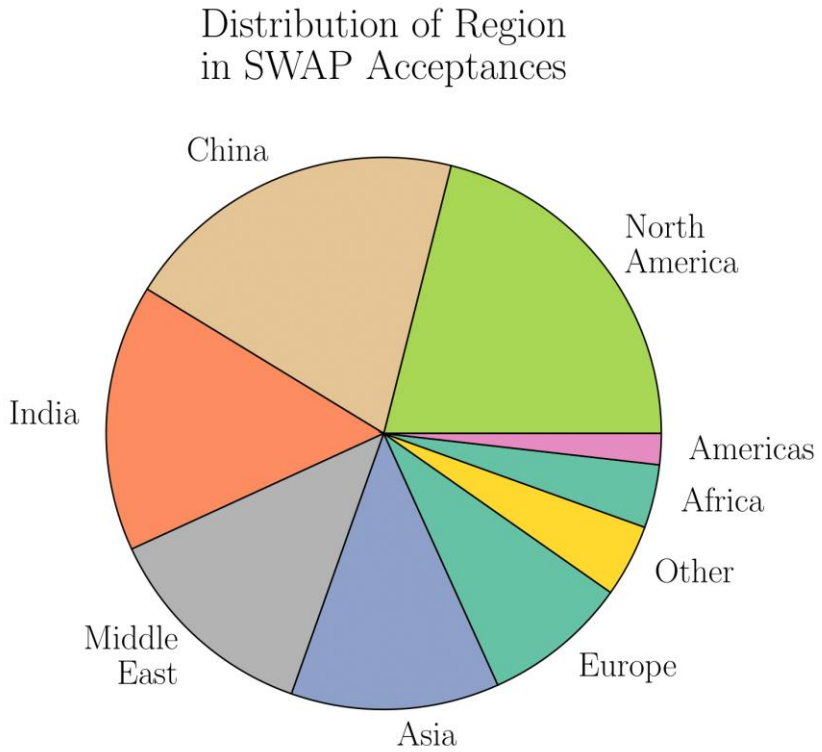
- **Large gains in diversity (by design!)**
- **SWAP spent roughly the same amount of total resources as the admissions committee**
- **Slight drop in general fit (versus Top-K Utility)**

	M/F Ratio
SWAP	1.3:1 (0.02)
Actual	2.9:1

	Gender		Region of Origin	
	$\sqrt{w_{\text{TOP}}}$	w_{DIV}	$\sqrt{w_{\text{TOP}}}$	w_{DIV}
SWAP	8.5 (0.03)	12.1 (0.06)	8.0 (0.03)	22.1 (0.03)
Actual	8.6	11.8	8.6	20.47

GRADUATE ADMISSIONS EXPERIMENT RESULTS





→ ***SWAP-style approaches could serve as a useful decision support tool to promote diversity in practice***

	Gender		Region of Origin	
	$\sqrt{w_{\text{TOP}}}$	w_{DIV}	$\sqrt{w_{\text{TOP}}}$	w_{DIV}
SWAP	8.5 (0.03)	12.1 (0.06)	8.0 (0.03)	22.1 (0.03)
Actual	8.6	11.8	8.6	20.47

ONGOING WORK

Tiered hiring via structured interviews:

- How do we sub-select through resume → phone → on-site → offers?

Group fairness: e.g., incorporation of fair treatment (*vis a vis* sensitive attribute) of arms

What does diversity even mean?

- Picked a fairly arbitrary submodular function – human judgment aggregation?

How should we partition? Can we learn a good partitioning?

Detecting bias in application materials → incorporate this into automated scoring