MECHANISM DESIGN

JOHN P DICKERSON & MARINA KNITTEL

Lecture #10 – 02/23/2022

CMSC498T
Mondays & Wednesdays
2:00pm – 3:15pm

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

SOCIAL CHOICE & MECHANISM DESIGN PRIMER

Thanks to: AGT book, Conitzer (VC), Parkes (DP), Procaccia (AP), Sandholm (TS)

SOCIAL CHOICE

A mathematical theory that focuses on aggregation of individuals' preferences over alternatives, usually in an attempt to collectively choose amongst all alternatives.

- A single alternative (e.g., a president)
- A vector of alternatives or outcomes (e.g., allocation of money, goods, tasks, jobs, resources, etc)

Agents reveal their preferences to a center

A social choice function then:

- aggregates those preferences and picks outcome

Voting in elections, bidding on items on eBay, requesting a specific paper/lecture presentation in CMSC498T, …

FORMAL MODEL OF VOTING

Set of voters *N* and a set of alternatives *A*

Each voter ranks the alternatives

- Full ranking
- Partial ranking (e.g., US presidential election)

A preference profile is the set of all voters' rankings

1	2	3	4
a	*b*	*a*	*c*
b	*a*	*b*	*a*
c	*c*	*c*	*b*

VOTING RULES

A voting rule is a function that maps preference profiles to alternatives

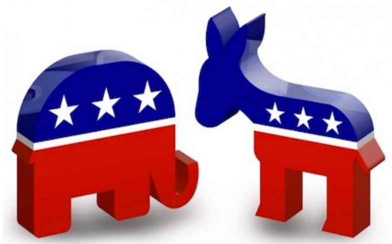
Many different voting rules – we'll discuss more later

Plurality: each voter's top-ranked alternative gets one point, the alternative with the most points wins

1	2	3	4
a	*b*	*a*	*c*
b	*a*	*b*	*a*
c	*c*	*c*	*b*

??????????

a: 2 points; *b*: 1 point; *c*: 1 point → *a* wins



SINGLE TRANSFERABLE VOTE

Wasted votes: any vote not cast for a winning alternative

- Plurality wastes many votes (US two-party system …)
- Reducing wasted votes is pragmatic (increases voter participation if they feel like votes matter) and more fair

Single transferable vote (STV):

- Given *m* alternatives, runs *m*-1 rounds
- Each round, alternative with fewest plurality votes is eliminated
- Winner is the last remaining alternative
- (General: If there is more than one seat, stop when #seats remain)

Ireland, Australia, New Zealand, a few other countries use STV (and coincidentally have more effective "third" parties…)

- You might hear this called "instant run-off voting" – this is equivalent to the single-winner version of STV

STV EXAMPLE

Starting preference profile:

1	2	3	4	5
a	*a*	*b*	*b*	*c*
b	*b*	*a*	*a*	*d*
c	*c*	*d*	*d*	*b*
d	*d*	*c*	*c*	*a*

1	2	3	4	5
a	*a*	*b*	*b*	*c*
b	*b*	*a*	*a*	*b*
c	*c*	*c*	*c*	*a*

Round 1, *d* has no plurality votes

Round 2, *c* has 1 plurality vote

1	2	3	4	5
a	*a*	*b*	*b*	*b*
b	*b*	*a*	*a*	*a*

1	2	3	4	5
b	*b*	*b*	*b*	*b*

Round 3, *a* has 2 plurality votes

AP

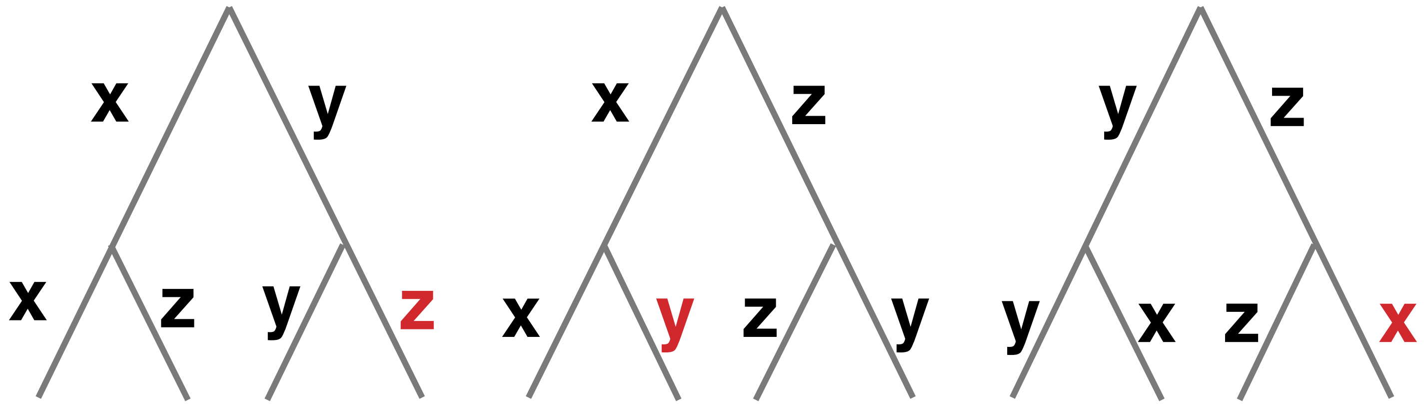
MANIPULATION: AGENDA PARADOX

Preference profile:

1. x > z > y (35%)
2. y > x > z (33%)
3. z > y > x (32%)

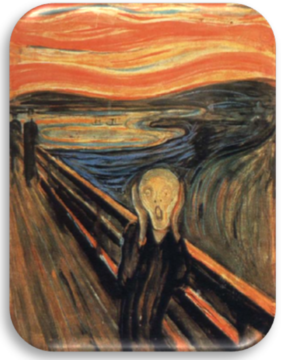
Binary protocol (majority rule), aka "cup"

Three types of agents:



Power of agenda setter (e.g., chairman)

Under plurality rule, x wins
Under STV rule, y wins



HOW SHOULD WE DESIGN VOTING RULES?

Take an axiomatic approach!

Majority consistency:

- If a majority of people vote for *x* as their top alternative, then *x* should win the election

Is plurality majority consistent?

- Yes

Is STV majority consistent?

- Yes

Is cup majority consistent?

- Yes

HOW SHOULD WE DESIGN VOTING RULES?

Given a preference profile, an alternative is a Condorcet winner if it beats all other alternatives in pairwise elections

- Wins plurality vote against any candidate in two-party election

Doesn't always exist! Condorcet Paradox:

1	2	3
x	z	y
y	x	z
z	y	x

$x > y$ (2-1); $y > z$ (2-1); $z > x$ (2-1) $\rightarrow$ $x > y > z > x$

Condorcet consistency: chooses Condorcet winner if it exists

- Stronger or weaker than majority consistency …?

HOW SHOULD WE DESIGN VOTING RULES?

1. **Strategyproof**: voters cannot benefit from lying.
2. **Computational tractability** of determining a winner?
3. **Unanimous**: if all voters have the same preference profile, then the aggregate ranking equals that.
4. **(Non-)dictatorial**: is there a voter who always gets her preferred alternative?
5. **Independence of irrelevant alternatives** (IIA): social preference between any alternatives *a* and *b* only depends on the voters' preferences between *a* and *b*.
6. **Onto**: any alternative can win

Gibbard-Satterthwaite (1970s): if $|A| \geq 3$, then any voting rule that is strategyproof and onto is a dictatorship.

COMPUTATIONAL SOCIAL CHOICE

There are many strong impossibility results like G-S

- We will discuss more of them (e.g., G-S, Arrow's Theorem) during the voting theory lectures in a month and a half

Computational social choice creates "well-designed" implementations of social choice functions, with an eye toward:

- Computational tractability of the winner determination problem
- Communication complexity of preference elicitation
- Designing the mechanism to elicit preferences truthfully

Interactions between these can lead to positive theoretical results and practical circumventions of impossibility results.

MECHANISM DESIGN: MODEL

Before: we were given preference profiles

Reality: agents reveal their (private) preferences

- Won't be truthful unless it's in their individual interest; but
- We want some globally good outcome

Formally:

- Center's job is to pick from a set of outcomes O
- Agent i draws a private type θ_i from Θ_i, a set of possible types
- Agent i has a public valuation function $v_i : \Theta_i \times O \rightarrow \Re$
- Center has public objective function $g : \Theta \times O \rightarrow \Re$
  - Social welfare max aka efficiency, maximize $g = \Sigma_i v_i(\theta_i, o)$
  - Possibly plus/minus monetary payments

MECHANISM DESIGN WITHOUT MONEY

A (direct) deterministic mechanism without payments z maps $\Theta \rightarrow O$

A (direct) randomized mechanism without payments z maps $\Theta \rightarrow \Delta(O)$, the set of all probability distributions over O

Any mechanism z induces a Bayesian game, Game(z)

A mechanism is said to implement a social choice function f if, for every input (e.g., preference profile), there is a Nash equilibrium for Game(z) where the outcome is the same as f

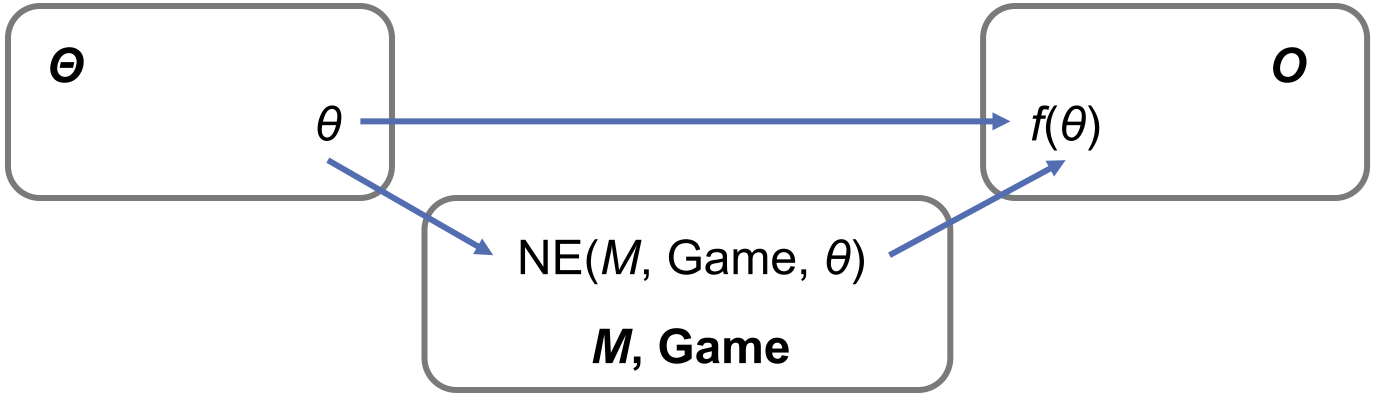
PICTORIALLY …

Agents draw private types θ from Θ

If those types were known, an outcome $f(\theta)$ would be chosen

Instead, agents send *messages M* (e.g., report their type as θ', or bid if we have money) to the mechanism

Goal: design a mechanism whose Game induces a Nash equilibrium where the outcome equals f(θ)



A (SILLY) MECHANISM THAT DOES NOT IMPLEMENT WELFARE MAX

2 agents, 1 item

Each agent draws a private valuation for that item

Social welfare maximizing outcome: agent with greatest private valuation receives the item.

Mechanism:

- Agents send a message of {1, 2, …, 10}
- Item is given to the agent who sends the lowest message; if both send the same message, agent $i = 1$ gets the item

Equilibrium behavior: ??????????

- Always send the lowest message (1)
- Outcome: agent $i = 1$ gets item, even if $i = 2$ values it more

MECHANISM DESIGN WITH MONEY

We will assume that an agent's utility for

- her type being θ_i,
- outcome *o* being chosen,
- and having to pay π_i,

can be written as $v_i(\theta_i, o) - \pi_i$

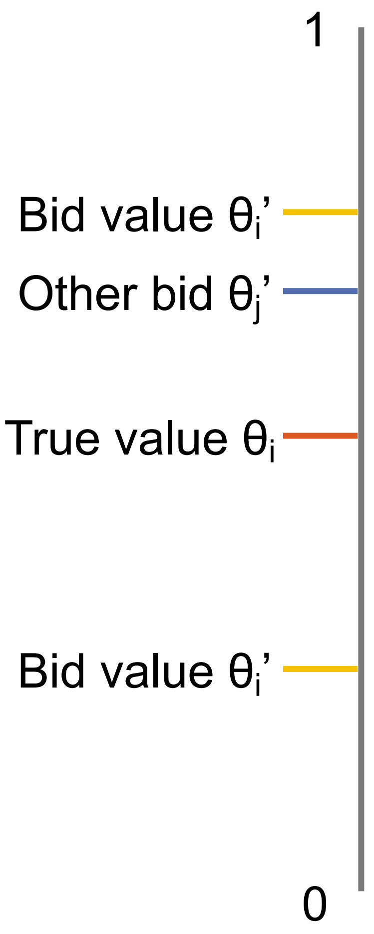
Such utility functions are called quasilinear

- "quasi" – linear with respect to one of the raw inputs, in this case payment π_i, as well as a function of the rest (i.e., $v_i(\theta_i, o)$)

Then, (direct) deterministic and randomized mechanisms with payments additionally specify, for each agent *i*, a payment function $\pi_i : \Theta \rightarrow \Re$

VICKREY'S SECOND PRICE AUCTION ISN'T MANIPULABLE

(Sealed) bid on single item, highest bidder wins & pays second-highest bid price



Bid $\theta_i' > \theta_i$ and win:

- Second-highest bid $\theta_j' > \theta_i$?
  - Payment is θ_j', pay more than valuation!
- Second-highest bid $\theta_j' < \theta_i$?
  - Payment from bidding truthfully is the same

Bid $\theta_i' > \theta_i$ and lose: same outcome as truthful bidding

Bid $\theta_i' < \theta_i$ and win: same outcome as truthful bidding

Bid $\theta_i' < \theta_i$ and lose:

- Winning bid $\theta_j' > \theta_i$?
  - Wouldn't have won by bidding truthfully, either
- Winning bid $\theta_j' < \theta_i$?
  - Bidding truthfully would've given positive utility

THE CLARKE (AKA VCG) MECHANISM

The Clarke mechanism chooses some outcome *o* that maximizes $\Sigma_i v_i(\theta_i', o)$

To determine the payment that agent *j* must make:

- Pretend *j* does not exist, and choose o_{-j} that maximizes $\Sigma_{i\neq j} v_i(\theta_i', o_{-j})$
- *j* pays $\Sigma_{i\neq j} v_i(\theta_i', o_{-j}) - \Sigma_{i\neq j} v_i(\theta_i', o) =$
  $= \Sigma_{i\neq j} (v_i(\theta_i', o_{-j}) - v_i(\theta_i', o))$

We say that each agent pays the externality that she imposes on the other agents

- Agent i's externality: (social welfare of others if *i* were absent) - (social welfare of others when *i* is present)

(VCG = Vickrey, Clarke, Groves)

INCENTIVE COMPATIBILITY

Incentive compatibility: there is never an incentive to lie about one's type

A mechanism is dominant-strategies incentive compatible (aka strategyproof) if for any *i*, for any type vector $\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_n$, and for any alternative type θ_i', we have

$$v_i(\theta_i, o(\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_n)) - \pi_i(\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_n) \geq$$
$$v_i(\theta_i, o(\theta_1, \theta_2, \ldots, \theta_i', \ldots, \theta_n)) - \pi_i(\theta_1, \theta_2, \ldots, \theta_i', \ldots, \theta_n)$$

A mechanism is Bayes-Nash equilibrium (BNE) incentive compatible if telling the truth is a BNE, that is, for any *i*, for any types θ_i, θ_i',

$$\Sigma_{\theta_{-i}} P(\theta_{-i}) [v_i(\theta_i, o(\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_n)) - \pi_i(\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_n)] \geq$$
$$\Sigma_{\theta_{-i}} P(\theta_{-i}) [v_i(\theta_i, o(\theta_1, \theta_2, \ldots, \theta_i', \ldots, \theta_n)) - \pi_i(\theta_1, \theta_2, \ldots, \theta_i', \ldots, \theta_n)]$$

VCG IS STRATEGYPROOF

Total utility for agent *j* is

$v_j(\theta_j, o) - \Sigma_{i \neq j} (v_i(\theta_i', o_{-j}) - v_i(\theta_i', o))$

$= v_j(\theta_j, o) + \Sigma_{i \neq j} v_i(\theta_i', o) - \Sigma_{i \neq j} v_i(\theta_i', o_{-j})$

But agent *j* cannot affect the choice of o_{-j}

→ *j* can focus on maximizing $v_j(\theta_j, o) + \Sigma_{i \neq j} v_i(\theta_i', o)$

But mechanism chooses *o* to maximize $\Sigma_i v_i(\theta_i', o)$

Hence, if $\theta_j' = \theta_j$, *j*'s utility will be maximized!

Extension of idea: add any term to agent *j*'s payment that does not depend on *j*'s reported type

- This is the family of Groves mechanisms

INDIVIDUAL RATIONALITY

A selfish center: "All agents must give me all their money." – but the agents would simply not participate

- This mechanism is not individually rational

A mechanism is ex-post individually rational if for any *i*, for any known type vector $\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_n$, we have

$$v_i(\theta_i, o(\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_n)) - \pi_i(\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_n) \geq 0$$

A mechanism is ex-interim individually rational if for any *i*, for any type θ_i,

$$\Sigma_{\theta_{-i}} P(\theta_{-i}) [v_i(\theta_i, o(\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_n)) - \pi_i(\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_n)] \geq 0$$

Is the Clarke mechanism individually rational?

WHY ONLY TRUTHFUL DIRECT-REVELATION MECHANISMS?

Bob has an incredibly complicated mechanism in which agents do not report types, but do all sorts of other strange things

- Bob: “In my mechanism, first agents 1 and 2 play a round of rock-paper-scissors. If agent 1 wins, she gets to choose the outcome. Otherwise, agents 2, 3 and 4 vote over the other outcomes using the STV voting rule. If there is a tie, everyone pays $100, and …”

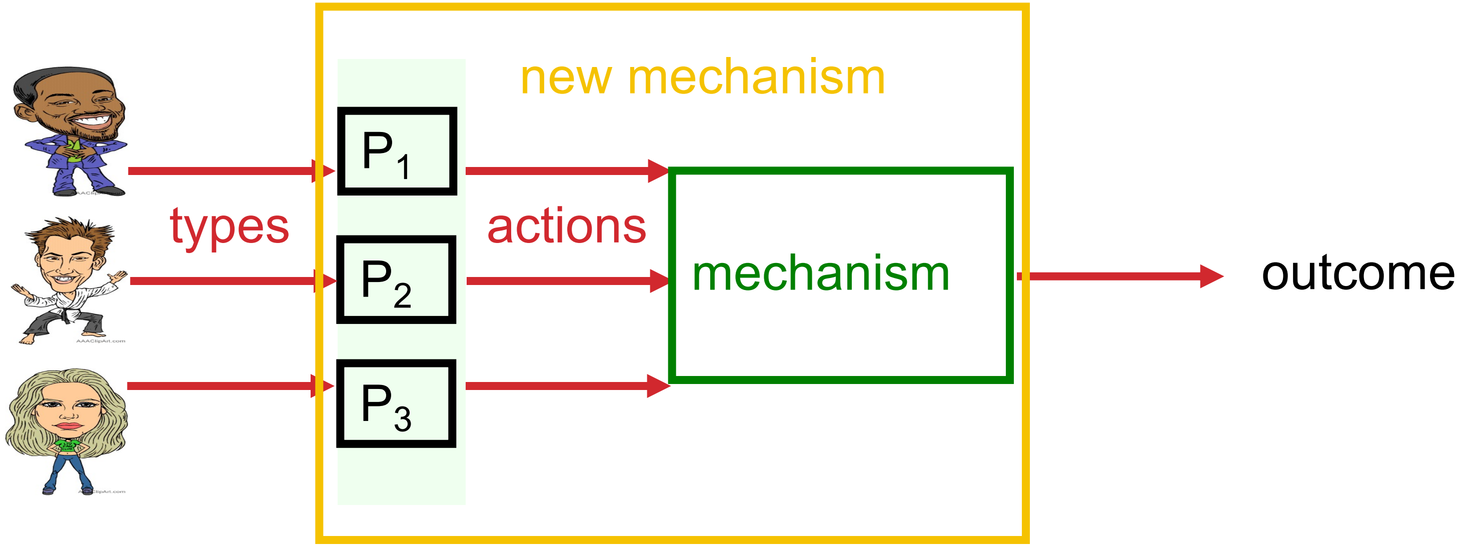
Bob: “The equilibria of my mechanism produce better results than any truthful direct revelation mechanism.”

- Could Bob be right?

THE REVELATION PRINCIPLE

For any (complex, strange) mechanism that produces certain outcomes under strategic behavior (dominant strategies, BNE)…

… there exists a {dominant-strategies, BNE} incentive compatible direct-revelation mechanism that produces the same outcomes!



REVELATION PRINCIPLE IN PRACTICE

"Only direct mechanisms needed"

- But: strategy formulator might be complex
  - Complex to determine and/or execute best-response strategy
  - Computational burden is pushed on the center (i.e., assumed away)
  - Thus the revelation principle might not hold in practice if these computational problems are hard
  - This problem traditionally ignored in game theory
- But: even if the indirect mechanism has a unique equilibrium, the direct mechanism can have additional bad equilibria

REVELATION PRINCIPLE AS AN ANALYSIS TOOL

Best direct mechanism gives tight upper bound on how well any indirect mechanism can do

- Space of direct mechanisms is smaller than that of indirect ones
- One can analyze all direct mechanisms & pick best one
- Thus one can know when one has designed an optimal indirect mechanism (when it is as good as the best direct one)

COMPUTATIONAL ISSUES IN MECHANISM DESIGN

Algorithmic mechanism design

- Sometimes standard mechanisms are too hard to execute computationally (e.g., Clarke requires computing optimal outcome)
- Try to find mechanisms that are easy to execute computationally (and nice in other ways), together with algorithms for executing them

Automated mechanism design

- Given the specific setting (agents, outcomes, types, priors over types, …) and the objective, have a computer solve for the best mechanism for this particular setting

When agents have computational limitations, they will not necessarily play in a game-theoretically optimal way

- Revelation principle can collapse; need to look at nontruthful mechanisms

Many other things (computing the outcomes in a distributed manner; what if the agents come in over time (online setting); …) – many good project ideas here ☺.

RUNNING EXAMPLE: MECHANISM DESIGN FOR KIDNEY EXCHANGE

THE PLAYERS AND THEIR INCENTIVES

Clearinghouse cares about global welfare:

- How many patients received kidneys (over time)?

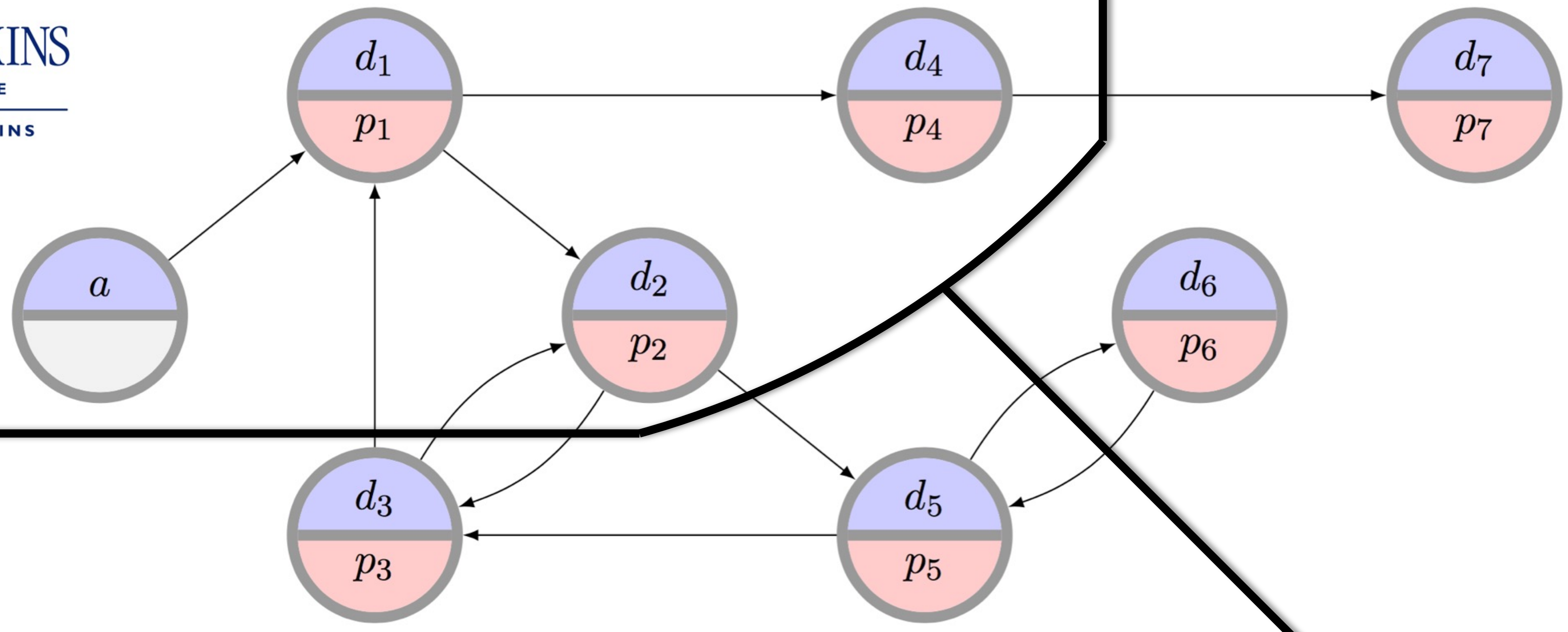
Transplant centers care about their individual welfare:

- How many of my own patients received kidneys?

Patient-donor pairs care about their individual welfare:

- Did I receive a kidney?
- (Most work considers just clearinghouse and centers)

PRIVATE VS GLOBAL MATCHING



MODELING THE PROBLEM

What is the type of an agent?

What is the utility function for an agent?

What would it mean for a mechanism to be:

- **Strategyproof**
- **Individually rational**
- **Efficient**

KNOWN RESULTS

Theory **[Roth&Ashlagi 14, Ashlagi et al. 15, Toulis&Parkes 15]:**

- **Can't have a strategy-proof and efficient mechanism**
- **Can get "close" by relaxing some efficiency requirements**
- **Even for the undirected (2-cycle) case:**
  - No deterministic SP mechanism can give 2-eps approximation to social welfare maximization
  - No randomized SP mechanism can give 6/5-eps approx
- **But! Ongoing work by a few groups hints at dynamic models being both more realistic and less "impossible"!**

Reality: transplant centers strategize like crazy! **[Stewert et al. 13]**