# Mechanism design
## (strategic voting)

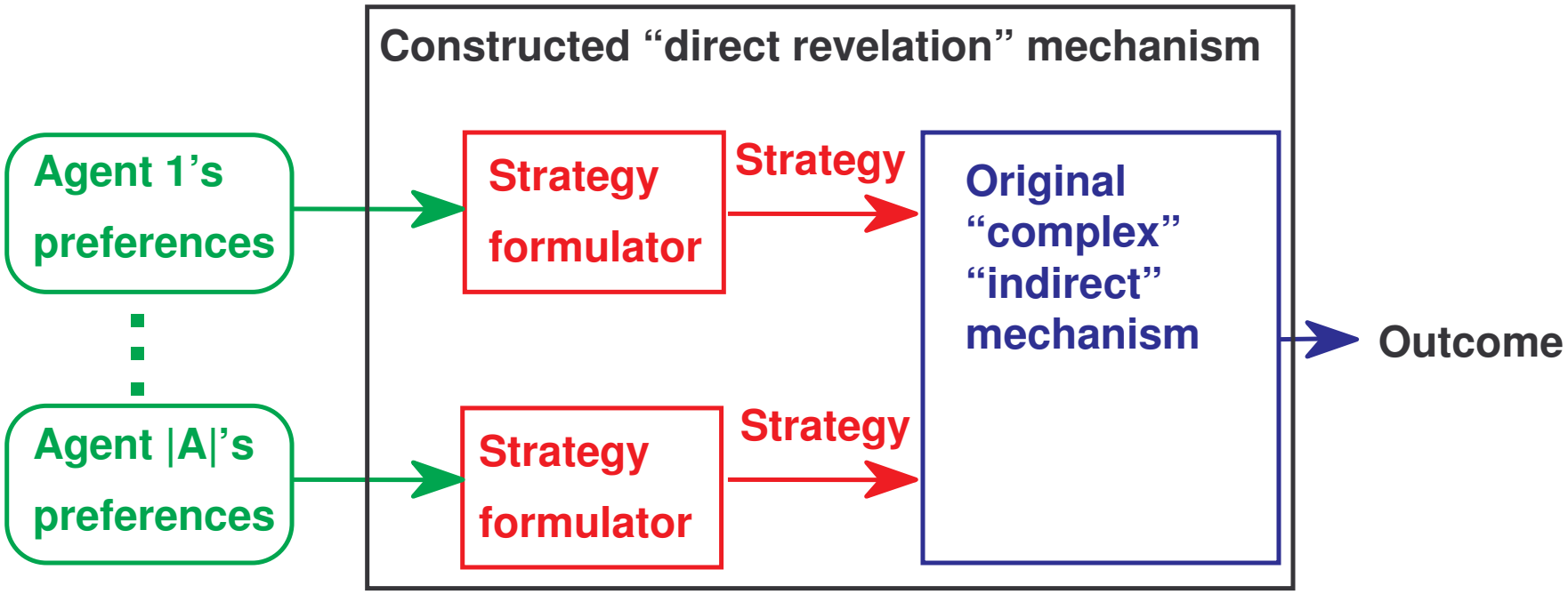
**Tuomas Sandholm**
Professor
Computer Science Department
Carnegie Mellon University

# Goal of mechanism design

- ***Implementing*** **a social choice function f(R) using a game**
  - **Actually, say we want to implement $f(u_1, \ldots, u_{|A|})$**
- **Center = "auctioneer" does not know the agents' preferences**
- **Agents may lie**
  - **unlike in the theory of social choice which we discussed in class before**
- **Goal is to design the rules of the game (aka mechanism) so that in equilibrium $(s_1, \ldots, s_{|A|})$, the outcome of the game is $f(u_1, \ldots, u_{|A|})$**
- **Mechanism designer specifies the strategy sets $S_i$ and how outcome is determined as a function of $(s_1, \ldots, s_{|A|}) \in (S_1, \ldots, S_{|A|})$**
- **Variants**
  - **Strongest: There exists exactly one equilibrium. Its outcome is $f(u_1, \ldots, u_{|A|})$**
  - **Medium: In every equilibrium the outcome is $f(u_1, \ldots, u_{|A|})$**
  - **Weakest: In at least one equilibrium the outcome is $f(u_1, \ldots, u_{|A|})$**

# Revelation principle

- **Any outcome that can be supported in Nash (dominant strategy) equilibrium via a complex "indirect" mechanism can be supported in Nash (dominant strategy) equilibrium via a "direct" mechanism where agents reveal their types truthfully in a single step**



# Uses of the revelation principle

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

# Implementation in dominant strategies

Strongest form of mechanism design

**Tuomas Sandholm**
Computer Science Department
Carnegie Mellon University

# Implementation in dominant strategies

- **Goal is to design the rules of the game (aka mechanism) so that in dominant strategy equilibrium ($s_1$, …, $s_{|A|}$), the outcome of the game is f($u_1$, …, $u_{|A|}$)**
- **Nice in that agents cannot benefit from counterspeculating each other**
  - **Others' preferences**
  - **Others' rationality**
  - **Other's endowments**
  - **Other's capabilities …**

# Gibbard-Satterthwaite impossibility

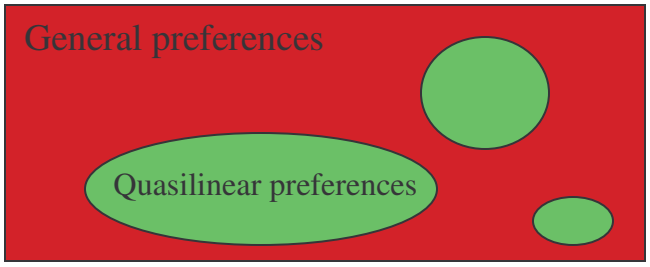
- Thrm. If $|O| \geq 3$ (and each outcome would be the social choice under f for some input profile $(u_1, \ldots, u_{|A|})$ ) and f is implementable in dominant strategies, then f is dictatorial
- Proof. (Assume for simplicity that utility relations are strict)
  - By the revelation principle, if f is implementable in dominant strategies, it is truthfully implementable in dominant strategies with a direct revelation mechanism
    - (maybe not in unique equilibrium)
  - Since f is truthfully implementable in dominant strategies, the following holds for each agent i: $u_i(f(u_i,u_{-i})) \geq u_i(f(u_i',u_{-i}))$ for all $u_{-i}$
  - Claim: f is monotonic. Suppose not. Then there exists u and u' s.t. f(u) = x, x maintains position going from u to u', and f(u') ≠ x
    - Consider converting u to u' one agent at a time. The social choices in this sequence are e.g. x, x, y, z, x, z, y, …, z. Consider the first step in this sequence where the social choice changes. Call the agent that changed his preferences agent i, and call the new social choice y. For the mechanism to be truth-dominant, i's dominant strategy should be to tell the truth no matter what others reveal. So, truth telling should be dominant even if the rest of the sequence did not occur.
    - Case 1. $u'_i(x) > u'_i(y)$. Say that $u'_i$ is the agent's truthful preference. Agent i would do better by revealing $u_i$ instead (x would get chosen instead of y). This contradicts truth-dominance.
    - Case 2. $u'_i(x) < u'_i(y)$. Because x maintains position from $u_i$ to $u'_i$, we have $u_i(x) < u_i(y)$. Say that $u_i$ is the agent's truthful preference. Agent i would do better by revealing $u'_i$ instead (y would get chosen instead of x). This contradicts truth-dominance.
  - Claim: f is Paretian. Suppose not. Then for some preference profile u we have an outcome x such that for each agent i, $u_i(x) > u_i(f(u))$.
    - We also know that there exists a u' s.t. f(u') = x
    - Now, choose a u'' s.t. for all i, $u_i''(x) > u_i''(f(u)) > u_i''(z)$, $\forall z \neq f(u), x$
    - Since f(u') = x, monotonicity implies f(u'') = x (because going from u' to u'', x maintains its position)
    - Monotonicity also implies f(u'') = f(u) (because going from u to u'', f(u) maintains its position)
    - But f(u'') = x and f(u'') = f(u) yields a contradiction because x ≠ f(u)
  - Since f is monotonic & Paretian, by strong form of Arrow's theorem, f is dictatorial. ■

# Ways around the Gibbard-Satterthwaite impossibility

- **Use a weaker equilibrium notion**
  - **E.g., Bayes-Nash equilibrium**
  - **In practice, agent might not know others' revelations**
- **Design mechanisms where computing a beneficial manipulation (insincere ranking of outcomes) is hard**
  - NP-complete in *second order Copeland* voting mechanism [Bartholdi, Tovey, Trick 1989]
    - *Copeland score*: Number of competitors an outcome beats in pairwise competitions
    - *2nd order Copeland*: Copeland, and break ties based on the sum of the Copeland scores of the competitors that the outcome beat
  - NP-complete in *Single Transferable Vote* mechanism [Bartholdi & Orlin 1991]
  - NP-hard, #P-hard, or PSPACE-hard in many voting protocols if one round of pairwise elimination is used before running the protocol [Conitzer & Sandholm IJCAI-03]
  - Weighted coalitional manipulation (and thus unweighted individual manipulation when the manipulator has correlated uncertainty about others) is NP-complete in many voting protocols, even for a constant #candidates [Conitzer, Sandholm & Lang JACM 2007]
  - "Typical case" complexity tends to be easy [Conitzer&Sandholm AAAI-06, Procaccia&Rosenschein JAIR-07]
- **Randomization**

IC => convex combination of
(some randomization to pick a dictator)
and
(some randomization to pick 2 alternatives)
[Gibbard Econometrica-77]

- **Agents' preferences have special structure**



# Quasilinear preferences: *Groves mechanism*

- **Outcome $(x_1, x_2, ..., x_k, m_1, m_2, ..., m_{|A|})$**
- ***Quasilinear* preferences: $u_i(x, m) = m_i + v_i(x_1, x_2, ..., x_k)$**
- ***Utilitarian* setting: Social welfare maximizing choice**
  - **Outcome $s(v_1, v_2, ..., v_{|A|}) = \max_x \sum_i v_i(x_1, x_2, ..., x_k)$**
- **Thrm. Assume every agent's utility function is quasilinear. A utilitarian social choice function f: v -> (s(v), m(v)) can be implemented in dominant strategies if $m_i(v) = \sum_{j \neq i} v_j(s(v)) + h_i(v_{-i})$ for arbitrary function h**
- **Proof. We show that every agent's (weakly) dominant strategy is to reveal the truth in this direct revelation (*Groves*) mechanism**
  - **Let v be agents' revealed preferences where agent i tells the truth**
  - **Let v' have the same revealed preferences for other agents, but i lies**
  - **Suppose agent i benefits from the lie: $v_i(s(v')) + m_i(v') > v_i(s(v)) + m_i(v)$**
  - **That is, $v_i(s(v')) + \sum_{j \neq i} v_j(s(v')) + h_i(v_{-i}') > v_i(s(v)) + \sum_{j \neq i} v_j(s(v)) + h_i(v_{-i})$**
  - **Because $v_{-i}' = v_{-i}$ we have $h_i(v_{-i}') = h_i(v_{-i})$**
  - **Thus we must have $v_i(s(v')) + \sum_{j \neq i} v_j(s(v')) > v_i(s(v)) + \sum_{j \neq i} v_j(s(v))$**
  - **We can rewrite this as $\sum_j v_j(s(v')) > \sum_j v_j(s(v))$**
  - **But this contradicts the definition of s()** ■

# Uniqueness of Groves mechanism

- **Thrm. Assume every agent's utility function is quasilinear. A utilitarian social choice function f: v -> (s(v), m(v)) can be implemented in dominant strategies for all v: A x O -> R only if $m_i(v) = \sum_{j \neq i} v_j(s(v)) + h_i(v_{-i})$ for some function h**
- **Proof.**
- **Can write $m_i(v) = \sum_{j \neq i} v_j(s(v)) + h_i(v_i , v_{-i})$**
- **We prove $h_i(v_i , v_{-i}) = h_i(v_{-i})$**
- **Suppose not, i.e., $h_i(v_i , v_{-i}) \neq h_i(v'_i , v_{-i})$**
- **Case 1. $s(v_i , v_{-i}) = s(v'_i , v_{-i})$. If f is truthfully implementable in dominant strategies, we have**
  - **that $v_i(s(v_i , v_{-i})) + m_i(v_i , v_{-i}) \geq v_i(s(v'_i , v_{-i})) + m_i(v'_i , v_{-i})$ and**
  - **that $v'_i(s(v'_i , v_{-i})) + m_i(v'_i , v_{-i}) \geq v'_i(s(v_i , v_{-i})) + m_i(v_i , v_{-i})$**
  - **Since $s(v_i , v_{-i}) = s(v'_i , v_{-i})$, these inequalities imply $h_i(v_i , v_{-i}) = h_i(v'_i , v_{-i})$. Contradiction**

# Uniqueness of Groves mechanism…

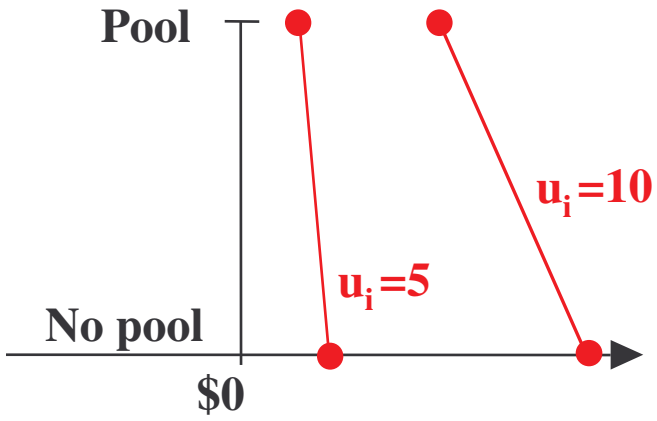
- **PROOF CONTINUES…**
- **Case 2. $s(v_i , v_{-i}) \neq s(v'_i , v_{-i})$. Suppose wlog that $h_i(v_i , v_{-i}) > h_i(v'_i , v_{-i})$**
- **Consider an agent with the following preference**
  - **Let $v''_i(x) = -\sum_{j\neq i} v_j(s(v_i , v_{-i}))$ if $x = s(v_i , v_{-i})$**
  - **Let $v''_i(x) = -\sum_{j\neq i} v_j(s(v'_i , v_{-i})) + \varepsilon$ if $x = s(v'_i , v_{-i})$**
  - **Let $v''_i(x) = -\infty$ otherwise**
- **We will show that $v''_i$ will prefer to report $v_i$ for small $\varepsilon$**
- **Truth-telling being dominant requires**
- **$v''_i(s(v''_i , v_{-i})) + m_i(v''_i , v_{-i}) \geq v''_i(s(v_i , v_{-i})) + m_i(v_i , v_{-i})$**
- **$s(v''_i , v_{-i}) = s(v'_i , v_{-i})$ since setting $x = s(v'_i , v_{-i})$ maximizes $v''_i(x) + \sum_{j\neq i} v_j(x)$**
  - **(This choice gives welfare $\varepsilon$, $s(v_i , v_{-i})$ gives 0, and other choices give $-\infty$ )**
- **So, $v''_i(s(v'_i , v_{-i})) + m_i(v''_i , v_{-i}) \geq v''_i(s(v_i , v_{-i})) + m_i(v_i , v_{-i})$**
- **From which we get by substitution:**
- **$-\sum_{j\neq i} v_j(s(v'_i , v_{-i})) + \varepsilon + m_i(v''_i , v_{-i}) \geq -\sum_{j\neq i} v_j(s(v_i , v_{-i})) + m_i(v_i , v_{-i}) \Leftrightarrow$**
- **$-\sum_{j\neq i} v_j(s(v'_i , v_{-i})) + \varepsilon + \sum_{j\neq i} v_j(s(v''_i , v_{-i})) + h_i(v''_i, v_{-i}) \geq -\sum_{j\neq i} v_j(s(v_i , v_{-i})) + \sum_{j\neq i} v_j(s(v_i , v_{-i})) + h_i(v_i, v_{-i})$**
- **$\Leftrightarrow \varepsilon + h_i(v''_i , v_{-i}) \geq h_i(v_i , v_{-i})$**
- **Because $s(v''_i , v_{-i}) = s(v'_i , v_{-i})$, by the logic of case 1, $h_i(v''_i , v_{-i}) = h_i(v'_i , v_{-i})$**
- **This gives $\varepsilon + h_i(v'_i , v_{-i}) \geq h_i(v_i , v_{-i})$**
- **But by hypothesis we have $h_i(v_i , v_{-i}) > h_i(v'_i , v_{-i})$, so there is a contradiction for small $\varepsilon$ ■**

- **Caveat to the theorem: Other mechanisms might work too *if v has special structure***

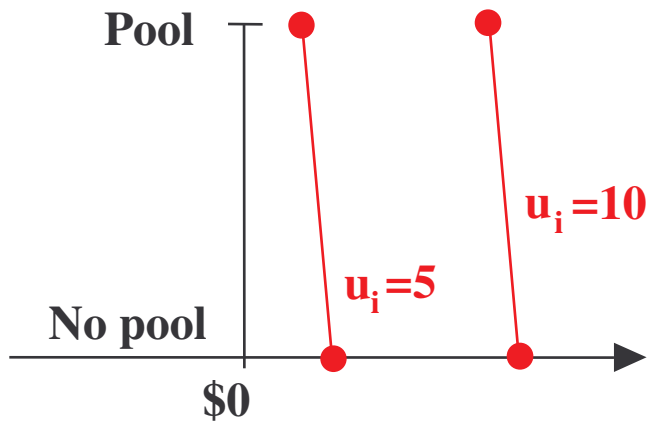
# Clarke tax "pivotal" mechanism

- Special case of Groves mechanism: $h_i(v_{-i}) = - \sum_{j \neq i} v_j(s(v_{-i}))$
- So, agent's payment $m_i = \sum_{j \neq i} v_j(s(v)) - \sum_{j \neq i} v_j(s(v_{-i})) \leq 0$ is a tax
- Intuition: Agent internalizes the negative externality he imposes on others by affecting the outcome
  - Agent pays nothing if he does not change ("pivot") the outcome
- Example: k=1, x1="joint pool built" or "not", mi = $
  - E.g. equal sharing of construction cost: $-c / |A|$, so $v_i(x_1) = w_i(x_1) - c / |A|$
  - So, $u_i = v_i(x_1) + m_i$

## General preferences




## Quasilinear preferences



# Clarke tax mechanism…

- **Pros**
  - **Social welfare maximizing outcome**
  - **Truth-telling is a dominant strategy**
  - **Ex post individually rational (i.e., even in hindsight each agent is no worse off by having participated)**
    - **Not all Groves mechanisms have this property, but Clarke tax does**
  - **Feasible in that it does not need a benefactor ($\sum_i m_i \leq 0$)**
- **Cons**
  - **Budget balance not maintained (in pool example, generally $\sum_i m_i < 0$)**
    - **Have to burn the excess money that is collected**
    - **Thrm. [Green & Laffont 1979]. Let the agents have quasilinear preferences $u_i(x, m) = m_i + v_i(x)$ where $v_i(x)$ are arbitrary functions. No social choice function that is (ex post) welfare maximizing (taking into account money burning as a loss) is implementable in dominant strategies**
    - **If there is some party that has no private information to reveal and no preferences over x, welfare maximization and budget balance can be obtained by having that party's payment be $m_0 = -\sum_{i=1..} m_i$**
      - **E.g. auctioneer could be agent 0**
  - **Vulnerable to collusion**
    - **Even by coalitions of just 2 agents**