Computational Learning Theory **Spring Semester, 2010**

Lecture 9: Mechanism Design and Social Choice

Lecturer: Yishay Mansour *Scribe: Eyal Altshuler, Nir Gazit*[1]

9.1 Mechanism Design

9.1.1 Overview

Mechanism Design is the closest game theory topic to practical engineering. We want to build an algorithm (mechanism) that, given player preferences as an input, will suggest an output that will be fair, optimal, etc. for all players. Our main problem here will be the strategic behavior of the players. Namely, we would prefer if players would tell their real preferences.

In the following lectures we will discuss these topics concerning Mechanism Design:

1. *Social Choice* (This lecture). We'll mainly talk on elections without money involved. We want the players to combine their preferences without payments.

2. *VCG* (Next lecture). The mechanism will have payment to and from the players in order to make sure they report their real preferences.

3. *Sponsered Search.*

4. *Combinatorial Auctions.*

9.2 Social Choice

Social choice is the general problem of mapping a set of multiple individual preferences to a single social preference, that will best reflect the aggregate individual preferences. Common examples for such scenarios are public elections, voting, etc. We shall show that given very simple and reasonable requirements, no mechanism that exhibits the above behavior exists.

We shall show two theorems for two slightly different social choice scenarios: Arrow's Impossibility Theorem and the Gibbard-Satterthwaite Theorem.

[1] based in part on the scribe notes of Dean Agron , Avi Sedaka , Shai Embon from 2005/6.

9.2.1 Motivation

Let us start by defining our main goal. Given the preferences of players, we need to define the Social Choice.

Example: Public Elections with 2 candidates.

A simple mechanism will be to ask each player about his preferred candidate, and then choose the candidate with the higher number of voters. Notice that each player has to tell the truth about his preference. This is the only way he can contribute to the election of his candidate.

Another Example: Public Elections with 3 (or more) candidates.

- A simple solution will be to ask each player for a list of preferences. The first candidate in the list will be his most preferred, and the last candidate will be his least preferred. Then,we can choose between any 2 candidates A and B simply by going through all the votes and checking which candidate is graded higher than his opponent most of the times.

  In 1785, the *Marquis de Condorcet* discovered a problem! Suppose there are 3 candidates: A, B, C, and three players that prefer:

  Player 1: $A >_1 B >_1 C$

  Player 2: $B >_2 C >_2 A$

  Player 3: $C >_3 A >_3 B$

  Now, according to this mechanism:

  $A > B$ (by players 1 & 3)

  $B > C$ (by players 1 & 2)

  $C > A$ (by players 2 & 3)

  We get a cyclic social choice: $A > B > C > A$. Also, for each candidate we choose, 2 of the players will prefer a different candidate.

- Plurality - Each player votes for one candidate, and the winner is the candidate that got the most votes.

- Borda Count - Each player gives a list of candidate preferences. The i-th candidate in each list gets $n - i$ points. The winner is the candidate with most points.

  The last 2 mechanisms are problematic due to the fact that they encourage startegic voting. For example, let us assume that a player has a preference $A > B > C$.

However, if he 'knows' that A will not win, then in the Plurality mechanism he'll vote for B, and in Borda Count mechanism he'll place B to be the first in his vote list - which is different from his real preferences.

Strategic voting is a problematic issue because it depends on the other players' preferences. Moreover, A player that plays strategically should expect that the other players will also play strategically. Such a behavior will demage the election mechanism, in such way the the results will not reflect the real preferences of the voters.

So, our goal here will be to find a mechanism in which each player will always prefer to report the truth. However, ironically, in the next section, we'll see that such a mechanism does not exist.

9.2.2 Arrow's Impossibility Theorem

Arrow's Impossibility Theorem deals with social ordering. Given a set of alternatives $\mathcal{A} = \{A, B, C, \ldots\}$, a *transitive preference* is a ranking of the alternatives from top to bottom, with ties allowed. Given a set of individuals (a society, so to speak), a *social preference function* is a function associating any tuple of personal transitive preferences, one per individual, with a single transitive preference called the *social preference.*

Definition A **Transitive Preference** is an ordering of $\mathcal{A}$ with no ties allowed. (Actually the proof will allow also for ties.)

For example: $A > B > C > D > E$.

Definition Given alternatives $\mathcal{A}$ and a set of individuals $\mathcal{N}$, a **Social Profile** is an association of a transitive preference per individual.

We will typically denote individual by numbers $1 \ldots N$, in which case a social profile will be represented by an N-tuple of transitive preferences. We will also represent a profile by a matrix, where each column is the transitive preference of single individual, ranked from top to bottom. For example:

1	2	3	4
A	B	C	A
B	D	D	B
C	A	B	C
D	C	A	D

Definition A **Social Preference Function** is a function associating each profile with a transitive preference, called the social preference.

1	2	3	4		Social
A	B	C	A		A
B	D	D	B	→	C
C	A	B	C		D
D	C	A	D		B

Definition A social preference function respects **Unanimity** if the social preference strictly prefers α over β whenever all of the individuals strictly prefer α over β.

Definition A social preference function respects **I.I.A. (Independence of Irrelevant Attributes)** if the relative social ranking (higher and lower) of α and β depends only on their relative ranking in the profile (and not on the ranking of other 'attributes', or alternatives).

Definition Given a social preference function, an individual n is a **Dictator** if the social preference strictly prefers α over β whenever n strictly prefers α over β. If a dictator exists, the social preference function is a **Dictatorship**.

Unless specifically stated, relationships (such as prefer, above, first, last, etc) are non-strict. We are now ready to present Arrow's Impossibility Theorem.

Theorem 9.1 ***Arrow's Impossibility Theorem*** *For 3 or more alternatives, any social preference function that respects unanimity and I.I.A., is a dictatorship.*

Proof: Assume a social preference function meeting the conditions set in the theorem. Let $B \in \mathcal{A}$ be chosen arbitrarily.

Claim 9.2 *For any profile where B is always either strictly first or strictly last in the ranking (for all individuals), the social preference must place B either strictly first or strictly last as well.*

Proof: Assume to the contrary that the social preference does not place B strictly first or strictly last. Then there exist two alternatives A and C (different from each other and B) such that in the social preference, $A \geq B$ and $B \geq C$. Since all individuals place B strictly first or strictly last, moving C strictly above A for an individual is possible without changing the relative preference between A and B or B and C for this individual. This is depicted in Figure 9.1. Due to I.I.A., we conclude that moving C strictly above A for all

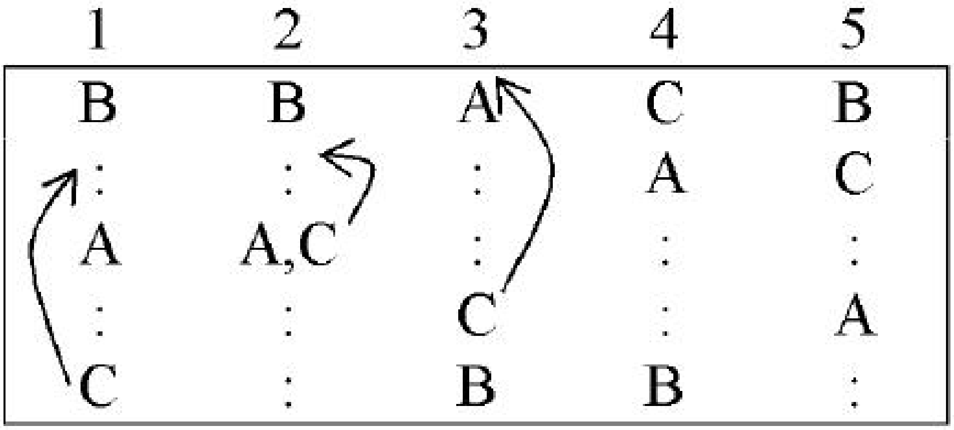

Figure 9.1: Moving C strictly above A without changing relative preference to B

individuals should not change the relative social preference between A and B, and between B and C, so we still have $A \geq B$ and $B \geq C$, which implies $A \geq C$ due to transitivity. But this contradicts unanimity, because now all individuals strictly prefer C over A.

Therefore, the social preference must place B strictly first or strictly last. □

Claim 9.3 *There exists an individual n^* and a specific profile such that n^* can swing B from the strictly last position in the social preference to the strictly first position by changing his preference.*

Proof: We observe an arbitrary profile where B is strictly last for all individuals. Due to unanimity, B must be strictly last in the social preference. Now let the individuals, from 1 to N, move B from the strictly last position to the strictly first position successively. Due to the previous claim, in any stage B must be strictly first or strictly last in the social preference. Because it starts strictly last, and must end strictly first, there must be an individual whose change causes B to move from the former position to the latter. We denote this individual by n^*. Denote by profile I the profile just before n^* changes his preference, and by profile II the profile just after the change. Profile I is the profile mentioned in the claim, and n^* is the individual. This is depicted in Figure 9.2. □

Note that n^* will have this behavior for any profile where all individuals $i < n^*$ place B strictly first and all individuals $i \geq n^*$ place B strictly last. The reason is that the (strict) relative preferences between B and any other alternative in such a profile and in profile I are identical, so this must hold in the social preference, and thus B must still be strictly last in any such profile. The same is true for the changed profile and profile II, where B must be strictly first. Therefore the choice of n^* is only dependent on B,and the order of

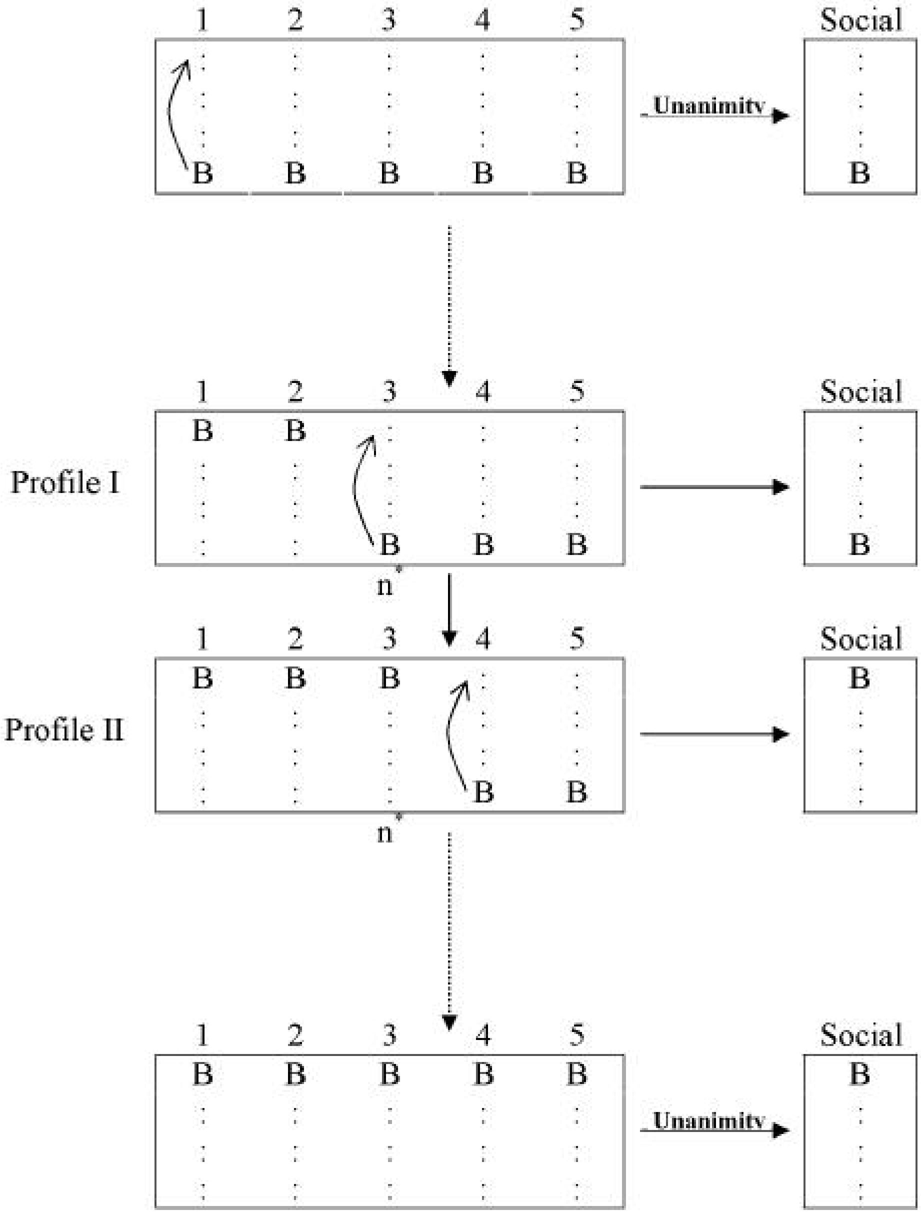

Figure 9.2: Existence of n^*

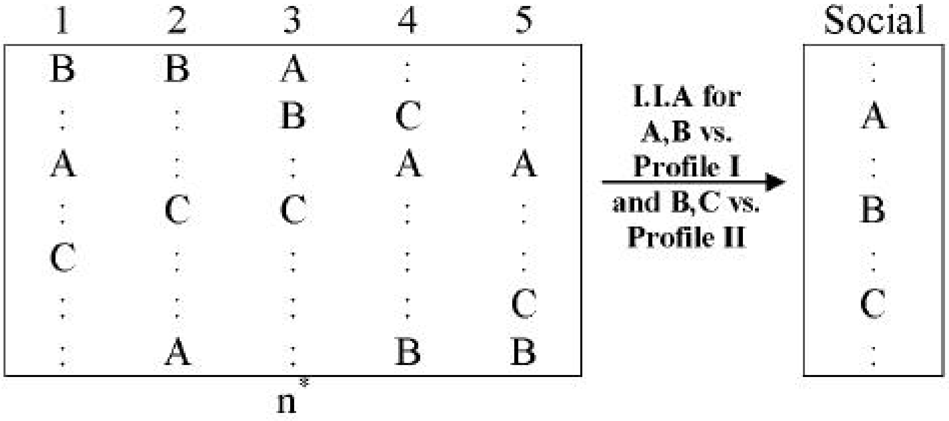

Figure 9.3: Profile IV

the players, not the profile. Later we will show that n^* is also independent of the order of the players.Therefore we can denote $n^* = n(B)$.

Claim 9.4 *n^* is a dictator for any pair of alternatives A and C that does not include B.*

Proof: Recall that by I.I.A., the social preference between A and C depend only on the individuals preference between A and C. Given any profile III where for n^*, $A > C$, create profile IV from profiles II and III by:

1. Start with profile II
2. Have n^* move A strictly above B, without changing the relative preferences among all alternatives other than A.
3. Have all other individuals change the preference between A and C to be identical to profile III, while B remains in it's profile II position.

In profile IV, the relative preference between A and B is identical to their relative preference in profile I for all individuals, and due to I.I.A. we must have $A > B$ in the social preference for profile IV (because in profile I , B is strictly last). The relative preference between C and B is in profile IV is identical to that of profile II for all individuals, thus (I.I.A) we must have $B > C$ in the social preference for profile IV (because in profile II B is strictly first). This is depicted in figure 9.3.

Therefore we must have $A > C$ in the social preference for profile IV. But the relative preferences between A and C in profiles III and IV are identical, so we must also have $A > C$ in the social preference for profile III. This is true for any profile III with n^* strictly preferring A over C, thus n^* is a dictator for A and C. □

Claim 9.5 *n^* is a dictator for any pair of alternatives A and B.*

Choose a third alternative C. By the same construction above, there exists $n(C)$ that is a dictator for any pair of alternatives exclusive of C, such as A and B. But $n(B)$ definitely effects the relative social preference of A and B, so he is the only possible dictator for A and B, thus $n(C) = n(B)$. Therefore $n(B)$ is also a dictator for A and B.

We have shown that there is a single individual that is dictator for any pair of alternatives, thus the social preference function is a dictatorship. □

Claim 9.6 *The identity of n^* is independent of the order of the individuals.*

Proof: For contradiction assume that using a different order we get $k^* \neq n^*$. Consider a profile where k^* has $A > C$ and n^* has $A < C$. By Claim 11.6 n^*(or k^*) is a dictator for A and C. But only one can hold. Therefore n^*=k^*. □

9.2.3 Gibbard-Satterthwaite Theorem

We shall now deal with an even simpler problem. The general scenario is similar to the one described previously. The difference is that we will only be interested in a single "most-desired" alternative, instead of an entire social preference. Instead of looking for a social preference function, we are looking for an election mechanism.

Definition An **Election Mechanism** is a function $f : L^n \to S$ mapping each social profile to a single alternative (the **elected alternative**) .

Definition An election mechanism M is said to be onto if for every alternative S, there exists a social profile L^n such that $f(L^n) = S$.

Definition A mechanism M that decides an election is defined to be a **dictatorial** one, when: There exist a dictator, namely a voter v such that if v votes for candidate A, then A will win the election regardless of the other voters' votes.

Definition A mechanism M that decides an election is defined to be a **strategy proof** one, when: The dominant strategy of each voter is voting in the order of his real preferences ("telling the truth"). Namely, if the voter prefers candidate A over B, his dominant strategy will be to rank A above B.

Formally, player i can do a strategic manipulation if there exist $L_1, ..., L_n$ (votes) and L'_i (another vote for the i-th player) such that $f(L_1, ..., L_n) = A$ and $f(L_1, ..., L_{i-1}, L'_i, L_{i+1}, ..., L_n) = A'$ where A and A' are alternatives and $A <_i A'$. The problem is that player i can tell L'_i as his preference and according to the definition of f, his status will improve.

Mechanism f is called strategy-proof (or incentive-compatible) if no player can do a strategic manipulation.

We are now ready to present the central theorem of this section.

Theorem 9.7 ***(Gibbard-Satterthwaite Theorem)*** *An election mechanism for 3 or more alternatives which is:*

- *Onto*
- *Strategy proof*

is a dictatorship.

This theorem will also be referred to as the **GS Theorem**. We precede the proof with a few lemmas.

Definition An election mechanism f is monotonic if: $f(L_1, ..., L_n) = A \neq A' = f(L_{-i}, L'_i)$ implies $A' <_{L_i} A$ and $A <_{L_{i'}} A'$.

Intuitively, if the mechanism chose A' instead of A, when only player i changed his input, it implies that player i actually prefers A' over A.

Claim 9.8 *An election mechanism f is startegy-proof if and only if f is monotonic.*

Proof: Suppose f is strategy proof. Let us take player i and look at his preference L_i. If player i changes his preference to L'_i, and the result of f is changed, then it must be worse for player i in L'_i (because of the strategy-proof property of f). Thus, f is monotonic.

On the other hand, suppose f is monotonic. For a player i, suppose he has 2 preferences L_i and L'_i, where L_i is better for him. Then, because of the montonicity of f, the result A of the mechaism will be better for player i then the result A' for L'_i. In other words, he can't do a strategic manipulation. □

Proof of the GS Theorem:

The idea: we will build a mechanism F from f, and show that F satisfies the attributes of *Arrow's theorem*. From that we conclude that F doesn't exist, and so f.

Definition Given an order L we shall define L^S by moving the elements of S to the beginning of the order L. Formally:

- if $a, b \in S$ then $a <^S b \Leftrightarrow a < b$
- if $a, b \notin S$ then $a <^S b \Leftrightarrow a < b$
- if $a \notin S, b \in S$ then $a <^S b$

Claim 9.9 *For every social profile $L_1, ..., L_n$ and a set S: $f(L_1^S, ..., L_n^S) \in S$.*

Proof: We choose an alternative $A \in S$. There exists $(L'_1, ..., L'_n)$ such that $f(L'_1, ..., L'_n) = A$. Sequentially, we change the inputs from L'_i to L_i^S. We claim that the output we always be in S. By contradiction, we assume that the output has changed to $B \notin S$, because $B <_i^S A$ it's a contradiction to the monotonicity. □

Now we show how to build F from f - For every two alternatives A, B, we define an order $>_F$ as follows: $f(L_1^{\{A,B\}}, ..., L_n^{\{A,B\}}) = A \Leftrightarrow A >_F B$

Intuitively, we move A, B to the beginning and preserve the order between them in every preference, and then we take f.

Example:

1	2	3
C	B	D
B	D	C
A	A	A
D	C	B

→

1	2	3
B	B	A
A	A	B
C	D	D
D	C	C

Claim 9.10 *This claim is divided into 2 parts:*

1. *if f is onto and strategy proof then F is a social choice function (defines an order).*
2. *if f is not dictatorship then F is unanimous, I.I.A and not dictatorship.*

Proof:

1. - *Antisymmetric:* $S = \{A, B\} \Rightarrow f(L_1^S, ..., L_n^S) \in \{A, B\}$ and thus we defined the order between A and B.
   - *Transitive:* In contradiction, we assume that there exist a circle $A > B > C > A$. We choose S to be the set of the alternatives in the circle.
     $f(L_1^S, ..., L_n^S) = A \in S$
     Assuming $B \in S : F(L_1^{\{A,B\}}, ..., L_n^{\{A,B\}}) = A$ and thus $A > B$ for all $A \neq B \in S$ - in contradiction with the circle.

2. - *Unanimity:* If for all i $A >_i B$ then $\left(L_i^{\{A,B\}}\right)^{\{A\}} = L_i^{\{A,B\}}$, and-
     $f(L_i^{\{A,B\}^{\{A\}}}) = A = f(L_i^{\{A,B\}})$.
   - *I.I.A:* Let us look at social profiles L_i and L'_i. We assume that for every i $A >_i B \Leftrightarrow A >_{i'} B$ Thus - $f((L'_1)^{\{A,B\}}, ..., (L'_n)^{\{A,B\}}) = f((L_1)^{\{A,B\}}, ..., (L_n)^{\{A,B\}})$
   - *Not dictatorship:* implied directly from the fact that f is not a dictatorship.

□

To summarize, we have assumed by a way of contradiction that such a mechanism f as in GS theorem exists, and shown that if it exists we can build a mechanism F for social choice that satisfies unanimity, I.I.A and not dictatorship, which contradicts Arrow's theorem. This proves GS Theorem.

□

9.3 Examples of Strategy-Proof Algorithms

9.3.1 Median Algorithm

The scope of alternatives is $[0, 1]$. for each player i we have a point z_i s.t: $|z_i - y| > |z_i - x| \Leftrightarrow x >_i y$. A strategy-proof algorithm can be: The preferences of each player is well-defined by telling his z_i. The algorithm takes the median between all the z_i's (assuming there is an odd number of players).

Claim 9.11 *The median algorithm is SP.*

Proof: We sort the z_i's:

$$\begin{array}{ccccc} z_1 & z_2 & z_3 & z_4 & z_5 \\ & & \uparrow & & \end{array}$$

The only way for a player whose z_i is not the median to change the median will make the median further away from him. Doing this will just make his status worse. Thus the algorithm is SP.

□

Notice that each choice of the k'th element (by sorted order) is also SP.

9.3.2 Allocating problems

We are given n players where each player has been assigned a task. The players can switch tasks between themselves. The algorithm should match between players and tasks, given a preference order for each player.

Definition Blocking Coalition - A subset C of players that can switch tasks between themselves such that there exists at least one player that can definitely improve his status after the switches (while the others' imrove themselves or remain the same).

Definition The core of the game is defined as the set of allocations that don't have a blocking coalition.

We now suggest an algorithm to solve the problem:

TTCA - Top Trading Cyclic Algorithm:
In phase K:

- We build a graph with the remaining players.
- There is a node for each player.
- There is an edge from player i node to player j node if player i prefers player j's task in the k'th place or before (self edges are allowed).
- We choose all the circles in the graph.
- In each circle we switch all the tasks along the circle.
- We drop the players in each such circle from the graph.

Theorem 9.12 *1. The algorithms returns an allocation in the core of the game (and it is the only possible legal allocation).*

2. The algorithm is SP.

Proof:

1. Let's mark N_k to be the set of players that were assigned tasks in the k'th phase. The players of N_1 won't join any coalition because they got their best preference. In addition, in every allocation in the core the players from N_1 will get the same allocation as in the algorithm (otherwise they will build a coalition). Inductively, let us assume for each $i < k$ that the assumption is correct. Now we look at the players in N_k. Players in N_i, where $i < k$ won't join to a coalition with players of N_k because they can just lose from such an action. Players of phase K won't join a coalition with players of N_i (where $K < i \leq n$) because the players in N_k will lose from such an action. Therefore the assignment is in the core.

2. Let us look at player $i \in N_k$. Suppose i doesn't report his true k'th preference and changes the algorithm allocation from π to π' and then gets task a' instead of a. The TTCA algorithm will work the same for the first $k-1$ phases so the task a' that player i got in π' must in the k'th priority (or less) of player i. However, the task a is also the best for player i from priority k or less. Thus, player i doesn't gain anything. In conclusion, it is better for i to report his true preference order.

□