Lecture 2

Arrow's Impossibility Theorem

Aggregating individual preferences is hard.

1 What Environment Are We In?

- Finite set $\mathbf{A} = \{A, B, C, \ldots\}$ of at least three different policy options
- Finite number N of different individuals $i = 1, 2, \ldots, N$
- Each person i has preferences over the policy options $\succsim_i$, which are *complete* and *transitive*, and for simplicity, we'll assume they are *strict* as well (no indifferences)
  - so for any individual i and any two policies a and b, either $a \succ_i b$ or $b \succ_i a$
  - and for any three policies a, b, and c, if $a \succ_i b$ and $b \succ_i c$, then $a \succ_i c$
- And that's it.

2 What Are We Trying To Do?

- We're looking for a way to *aggregate preferences* – that is, we want a way to turn each possible set of individual preferences $\{\succ_1, \succ_2, \ldots, \succ_N\}$ into a preference relation $\succsim^*$ for "society"
  - We won't require $\succsim^*$ to be strict
- The mapping from individual preference relations into social preferences relations is alternatively called a *social welfare function*, or a *preference aggregation rule*, or a *Constitution*
- I'll use *social welfare function* – just remember that the SWF is not a set of preferences itself, but a rule for generating a set of preferences for society for each set of individual preferences
  - So if we let X be the set of all possible preference relations over the set of policies $\mathbf{A}$, a SWF is a mapping from X^N to X
- We want our social welfare function to satisfy certain properties

2.1 Social preferences need to be defined for any set of individual preferences.

- Our SWF has to specify some set of social preferences $\succsim^*$ for *any* given set of individual preferences $\{\succ_1, \succ_2, \ldots, \succ_N\}$
- This is also called "universal domain" – we're not ruling out any possible preferences.

2.2 Social preferences should be complete and transitive.

- Just like individual preferences $\succsim_i$ need to be complete and transitive to be "reasonable"...
- we want the social preferences $\succsim^*$ chosen by our SWF to be complete and transitive for each set of individual preferences
- (otherwise we won't have a coherent choice rule for society)
- This rules out something like pairwise majority voting

  Why? Suppose there are three people and three policies, with preferences

$$A \succ_1 B \succ_1 C$$
$$B \succ_2 C \succ_2 A$$
$$C \succ_3 A \succ_3 B$$

  If we go by majority rule, two out of three prefer A to B, and two out of three prefer B to C, and two out of three prefer C to A; so social preferences would have to be $A \succ^* B \succ^* C$ but $C \succ^* A$, which isn't transitive.

2.3 Social preferences should respect unanimity.

- If *everyone* in society agrees that policy a is strictly better than b, then the social preferences defined by our SWF should also strictly prefer a to b.
- If not, it's doing a pretty bad job of aggregating the individuals' preferences.
- This rules out stupid rules like "society ranks all options equally" or "regardless of individual preferences, society ranks policies alphabetically"

2.4 The SWF should satisfy *independence of irrelevant alternatives*

- Basically, this says that if we're trying to figure out whether society prefers a to b, what people think of c shouldn't matter.
- Formally, suppose we start with some set of individual preferences $\{\succ_1, \ldots, \succ_N\}$, and the SWF picks a social preference function under which $a \succsim^* b$

  Now modify one guy's preferences $\succ_i$ such that his preference between a and b stays the same, but his preferences for other things (including whether he prefers c to a or b) changes

  Then under the new preferences, the SWF should still pick a social preference function that prefers a to b.

- This rules out a rule like the Borda count.

  Suppose there are three policies, and we ask everyone for a rank-order list; and then we give a policy 3 points for each person who ranked it first, 2 points for each person who ranked it second, and 1 point for every person who ranked it third, and we say society prefers policies with more points.

  This rule violates IIA. For example, with two voters, if

$$\begin{array}{ccccc} A & \succ_1 & B & \succ_1 & C \\ B & \succ_2 & C & \succ_2 & A \end{array}$$

  then the Borda count would prefer B to A. But if preferences were

$$\begin{array}{ccccc} A & \succ_1 & C & \succ_1 & B \\ B & \succ_2 & A & \succ_2 & C \end{array}$$

  it would prefer A to B. But in both cases, $A \succ_1 B$ and $B \succ_2 A$ – all we did was change parts of preferences that "aren't supposed to matter" for choosing between A and B. So a Borda count rule violates IIA.

2.5 (Aside: note the ordinality of preferences)

- Note that preferences here are ordinal, not cardinal
- All we know is whether an individual prefers one policy to another – we have no language to even talk about *how much* he prefers one to another
- We don't have money, to allow people to barter (I prefer A to B, but I'll prefer B if you give me \$11), or anything like that
- All we have is ordinal pairwise preferences, and we need to aggregate them to preferences for society

2.6 What's Left?

- That's what Arrow's theorem tells us. Not much.
- One more thing to define: a SWF is a *dictatorship* if the social preference always just reflects the same one guy's preferences, that is, if there's some individual k such that regardless of anyone else's preferences, $a \succ^* b$ if and only if $a \succ_k b$.

3 The Result

Theorem (Arrow)**.** *Any SWF which respects transitivity, unanimity, and independence of irrelevant alternatives is a dictatorship.*

- So basically, you have four choices:

  1. you can violate transitivity (with something like majority voting)
  2. you can violate unanimity (with something like, social preferences are fixed regardless of individual preferences)
  3. you can violate independence of irrelevant alternatives (with something like the Borda count)
  4. or you can have a dictatorship

  But those are the only choices.

- The proof has several steps. We assume we have a SWF which satisfies transitivity, unanimity, and IIA, and then show that this means there must be some voter whose preferences always match the social preferences, who is therefore a dictator.

4 The Proof

- Throughout the proof, we will maintain the assumptions that social preferences $\succsim^*$ are always transitive, and that the SWF satisfies unanimity and IIA.

4.1 Part 1 – the Extremal Lemma

Lemma 1 (Extremal Lemma)**.** *For any policy b, if every individual i ranks b* either *strictly best or strictly worst, then $\succsim^*$ must rank b either strictly best or strictly worst as well.*

- We'll prove this by contradiction.

- Suppose the lemma were false. Then there would be some set of individual preferences and two other policies a and c such that $a \succsim^* b \succsim^* c$, even though each individual has b as either their favorite or least favorite policy.

- Now modify individual preferences as follows.

  For each individual i who has b at the top of their list: move c up to second on the list, so it is now strictly above a if it wasn't already.

  And for each individual i who has b at the *bottom* of their list, move c to the top of the list, so it's strictly above a if it wasn't already.

- This doesn't change any individual's preferences between b and a: if b was at the top of your list, it's still at the top; and if it was at the bottom, it's still at the bottom. So by IIA, society still prefers a to b

- And similarly, it doesn't change any individual's preferences between b and c, so by IIA, b is still preferred to c.

- And by transitivity, $a \succsim^* b$ and $b \succsim^* c$ imply $a \succsim^* c$

- But now, *everyone* has c ranked above a, so unanimity would require $c \succ^* a$, giving a contradiction

- So that proves that b must be either strictly best or strictly worst according to $\succsim^*$ – if it wasn't, we could generate this type of contradiction.

4.2 Part 2 – Find a pivotal guy and give him a name.

- Pick some random policy b
- We know (*unanimity*) that if *everyone in society* puts b last, then $\succsim^*$ puts b last too; and if everyone puts b first, $\succsim^*$ puts b first
- And we just showed that if some people put b first and the rest put b last, then $\succsim^*$ puts b *either* first or last
- So start out with a set of individual preferences where *everyone in society puts b last.* By unanimity, $\succsim^*$ must put b last as well.
- Now change voter 1's preferences by moving b from last to first. Since everyone in society either likes b the most or the least, b must be either first or last in $\succsim^*$.
- Now change voter 2's preferences by moving b from last to first. Again, since everyone in society either likes b the most or the least, b must be either first or last in $\succsim^*$.
- Keep going like this. By the time we've switched *everyone's* preferences to having b first, by unanimity, society must have b first as well.
- Now find the voter where the first switch happened. That is, the first time that the social preference $\succsim^*$ switched from having b at the bottom to b at the top. Call him Bob.

  That is, given the other preferences we started with, if everyone with a lower number than Bob has b first on their list, and Bob and everyone after has b last on their list, then society puts b last; but if Bob switches to having b first, then society puts b first.
- Call the first profile of preferences – where everyone up to Bob puts b first, and Bob and the rest put it last – profile I; and call the second set of preferences – where everyone up to Bob, and Bob himself, put b first, and the rest put it last – profile II.
- The rest of the proof involves showing that Bob is a dictator – that to satisfy IIA and transitivity, $\succsim^*$ has to *always* agree with Bob's preferences, regardless of what everyone else thinks

4.3 Part 3 – Proving Bob is a dictator

Part 3a: Bob is a dictator over policies that aren't b.

- That is, for any two policies a and c which aren't b, we'll show that if Bob prefers a to c, then no matter what everyone else's preferences are, $\succsim^*$ must put a strictly ahead of c as well.

- Start with arbitrary preferences where $a \succ_{Bob} c$, and call these (true) preferences Profile IV.

- Make the following changes to preferences:

  Move a to the top of Bob's preference list, and b to second on Bob's preference list

  For individuals 1 up to Bob, move b to the top of their list

  For individuals Bob $+1$ up to N, move b to the bottom of their list.

  Call this new set of preferences Profile III

- When we moved from IV to III, we didn't change anyone's ranking of a versus c – for Bob, a was above c, and we moved it to the top; for everyone else, all we did was move b

  So by IIA, the societal preference between a and c has to be the same at profile IV as at profile III

  So we'll show that at profile III, society has to prefer a to c

- Now, recall that at profile I, society put b last, which means $a \succ^* b$ at profile I.

  And everyone's ranking of a versus b is the same at profile I as at profile III.

  So by IIA, $a \succ^* b$ at profile III.

- At profile II, on the other hand, society put b first, which means $b \succ^* c$ at profile II

  And everyone's ranking of b versus c is the same at II as at III

  So by IIA, $b \succ^* c$ at profile III.

- Since preferences at profile III must be transitive, $a \succ^* c$.

- So at profile IV, $a \succ^* c$.

- So whenever $a \succ_{Bob} c$, $a \succ^* c$.

Part 3b: Bob is also dictator when it comes to b

- All that's left is to show is that Bob is also a dictator when one of the choices being considered is b

- This is the easy part. We need to show that if $b \succ_{Bob} a$, $b \succ^* a$; and if $a \succ_{Bob} b$, $a \succ^* b$.

- So now pick a policy that's different from those two – say, c – and repeat everything we already did

- Starting with the extremal lemma, move preferences from c worst to c best, find the pivotal guy, and prove that he has to be a dictator for *any two policies that aren't c*

  - So far, we don't know whether this new dictator is Bob or someone else
  - Just that *someone* is a dictator when c isn't involved, meaning, when choosing between a and b

- But we already know that Bob's preferences over b *sometimes* matter

  - When we moved from profile I to II, Bob's preferences for b were all that changed

    And that shifted society from putting b last to putting b first

- So if the new dictator isn't Bob, we have a contradiction; so the new dictator must also be Bob

- Which means Bob is dictator over any pair of policies that excludes b, *and* over any pair of policies that includes b

- So Bob's your dictator, and we're done. □

5 So What?

- So the result is, if you want to aggregate individual preferences and get a transitive social preference that respects unanimity and IIA, all that's left is a dictatorship.
- One way to interpret this is, IIA is a really strong restriction
  - Basically, by assuming IIA, we're ruling out inferring anything "cardinal" about preferences from where you rank other alternatives
  - If there are 100 policy choices, and you have a and b ranked #45 and #46, chances are, you're pretty close to indifferent between them

    OTOH, if you have a ranked #1 and b ranked #100, you probably like a a lot more

    IIA says that for choosing between a and b, we have to treat those two cases the same – all we're allowed to consider is that you like a more than b

    By assuming IIA, we're really ruling out having any way to elicit cardinal preferences

    Which is what makes this hard
- Contrast that with the first welfare theorem
  - In some sense, markets behave well exactly because prices elicit *cardinal* information
  - How much you're willing to buy of something at a given price reveals exactly how much you like it – not just whether you prefer it to something else
- In Arrow's world, we can't use prices – or anything else – to make this type of judgment
- So we can't aggregate preferences in a well-behaved, coherent way
- One other thing to notice: we've completely ignored the problem of *figuring out* peoples' preferences
  - Depending on the social choice function, people might have an incentive to lie about their preferences
  - The rest of the semester will, in some sense, be focused on that part of the problem

References

- Original result: Kenneth Arrow (1951), *Social Choice and Individual Values*, New York: Wiley
- Our proof: John Geanakoplos (2005), "Three Brief Proofs of Arrow's Impossibility Theorem," *Economic Theory* 26(1)