Some Mathematical Aspects of Fairness

Daniel Kraft

March 20th, 2012

1 Introduction

For stability and success of human societies and to sustain and encourage cooperation between individuals, *fairness* and *justice* have been very important moral principles dating back to ancient times. So it is no surprise that questions of fairness as well as rules and laws to enforce just and fair behaviour already appeared thousands of years ago in human history. For instance, the Code of Hammurabi dates back nearly 4,000 years and is one of the oldest known texts with significant length — and it's purpose is precisely to define a set of laws governing how the people in Babylonia should interact with each other in a "fair way". The notion of justice described in this text also includes the famous principle of "eye for an eye, tooth for a tooth" as a very rough measure of "fair behaviour" [10] or in this case, fair punishment for an offender. Another example for early rules can be found in the Old Testament, which underlies the great monotheistic religions and can thus be seen as influential to a substantial amount of today's population of the world. The Ten Commandments [2, Ex 20:1–17] are the most famous example, and include important rules for a fair society (like "Thou shalt not steal." or "Thou shalt not bear false witness against thy neighbour."). One can also mention the Book of Leviticus [2, Lev] here, which consists nearly entirely of laws and rules; some of them concern correct behaviour with respect to God, but some are also of importance for human interaction. Of course, there are a lot more examples of early rules and codes of conduct aiming to stabilise human societies.

In legal science, there's a principle called *Equity* whose goal is to ensure fair judgement; in [11, p. 5] at the very beginning, it is defined as: "Equity is the means by which a system of law balances out the need for certainty in rule-making with the need to achieve fair results in individual circumstances." This is for instance reflected in Austrian law in article 879 of the ABGB [1], which states that "a transaction, that ... is against good morals, is void". So even in addition to the general usage of laws to ensure fair behaviour and cooperation, the aspect of fairness to each member of society is further emphasised by that principle.

Of course, laws in general are only concerned with fairness in a broad sense. In this work, I will consider three more specific aspects of this topic in the following sections. Section 2 will deal with fairness in game-theory and economy: There are situations in which the rational and theoretically optimal behaviour in the context of games is selfish and unfair to other players, and empiric evidence shows that humans tend to choose a fairer strategy despite it being sub-optimal when just considering the pay-offs under certain circumstances.

In Section 3, the main problem considered is that of fairly dividing some resources among applicants. Such problems are very important; a similar and very famous setting appears already in the Judgement of Solomon [2, 1Kings 3:16–28], where two women claim motherhood to one child. Another example is the classical problem of dividing a piece of cake between two people — with the well-known "Divide and Choose" solution (one person divides, the other then chooses the preferred piece). This famous procedure is also called the "Pie Rule" and used to balance advantages of the first move in several games, for instance for the game Hex [6]. Related is also the "riddle of the vanishing camel" which can be found in a lot of different versions, see for instance [17] for an elaborate treatment or [4] for a short problem statement:

> An Arabic king was dying and had his three sons come to his bed. He told them that except for his 17 camels, there was nothing they could inherit from him; his eldest son should get half of them, the second a third and the youngest son should be given a ninth. After his death, his sons were in the desert for one week with the camels and did not know how their

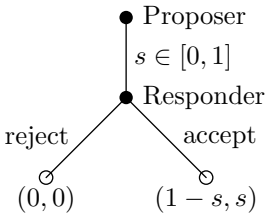

Figure 1: Tree of the Ultimatum Game.

> father's last wish could possibly be fulfilled, since 17 is neither divisible by two nor three nor nine. (And in fact a prime number.) They considered making parity via money, but that would not do their father justice and so they disregarded the idea. One day, a wise woman rode through the country on her camel. The sons asked her for advice; and she could finally give every son his due share, before continuing her journey on her camel. How did the woman accomplish that?

The crucial point is that $\frac{1}{2} + \frac{1}{3} + \frac{1}{9} < 1$, and so dividing the camels by the father's instructions would not have been possible even if the number was evenly divisible. However, when the woman added her own camel to the 17 of the king, she could give the first son $\frac{18}{2} = 9$ camels, the second $\frac{18}{3} = 6$ and the third $\frac{18}{9} = 2$. Then the sons had $9 + 6 + 2 = 17$ camels together, and because the fractions of the king do not add up to a whole unity, there would still be the camel of the woman left for her to take back.

As a final topic, I want to discuss *voting systems* in Section 4. There I will introduce the famous "Arrow Impossibility Theorem" (see [3]), which states that under certain definitions of fairness, voting can never be fully "fair". Clearly, that is a very important result for modern societies, where political and social decisions are made mainly by democratic voting! So naturally the question arises what alternatives there are to systems based on majority voting, where I also want to mention two possible other voting systems — one based on resistance votes in Subsection 4.2 and the so-called "Fractional Voting System" (see [15]) in Subsection 4.3, which is not subject to the restrictions of Arrow's theorem.

2 Inequity Aversion in Game-Theory

In this section, I will consider Ernst Fehr's theory of inequity aversion in humans. Basically, there are some examples of games that can be analysed theoretically and then tested on real humans experimentally — with the result that the strategy chosen by real players in the experiment does not agree with the usual assumption in economy that every ("rational") actor tries to maximise his or her own material pay-off. I will present two games from the seminal paper [5] and use them to explain Fehr's ideas about how this can be understood and how to model the behaviour of humans more reliably. Note that apart from [5], there's also current research going on about this topic; furthermore, not all authors agree that the actual game *outcome* (as discussed below) is important to induce, for instance, punishing behaviour, but rather the *intent* of other players. See [7] for a current research paper on this topic.

2.1 The Ultimatum Game

In the Ultimatum Game, two players — the Proposer and Responder — bargain about their share in some resource. Without loss of generality, we can normalise the available total pay-off to 1. Then the Proposer picks a share $s \in [0,1]$ and proposes that share to the Responder. In the next move, the Responder can either accept or reject the proposal. If she accepts, the she gets the share s while the Proposer gets the remaining $1 - s$. Otherwise nobody gets anything. The tree of this game is depicted in Figure 1.

If both players are purely selfish and try to get maximal material pay-off, then the Responder will accept any $s > 0$ since that is better than the pay-off of 0 in case of rejection. For $s = 0$, the Responder is indifferent between acceptance and rejection. The Proposer tries to maximise his own pay-off $1 - s$ and thus will choose s as small as possible. It follows that one Nash equilibrium of this game is a proposal of

$s = 0$, which is accepted by the Responder. In this case, neither of the players can profit by changing the strategy: If the Proposer picks $s' > s = 0$, he loses. If the Responder chooses to reject the offer, she is not worse off but also not better. Another equilibrium is $s = 0$, which is rejected. Any other combination of strategies can't be an equilibrium, since for $s > 0$, either the Proposer can improve himself by lowering s if the Responder accepts, or the Responder can get herself a better result by accepting instead of rejecting the offer.

So in practice, one may expect that a rational Proposer chooses a very small but non-zero s, while a rational Responder accepts any offer $s > 0$, at least. However, empiric results show that human players don't follow that strategy. Table 1 in [5, p. 827] shows that in different experiments across various countries, only very few offers with $s < 0.2$ were made and the majority of offers fell into the range $s \in [0.4, 0.5]$. Also, the experiments show that low but still strictly positive offers are frequently rejected by the Responder. This is a much "fairer" result than the assumption of rational players can explain.

So, how can these empiric facts be explained in a model? The idea of Fehr is that human players show what he calls *inequity aversion*: The driving force behind our decisions is not purely material pay-off, but rather our feeling of "total happiness" the outcome of a game incurs. And in addition to the real game pay-off, there are also penalties for non-equal pay-offs among the players included in that result. Namely, if one player gets lower pay-off than another, she feels treated unfairly and has an incentive to avoid that advantage of another over her. Also, if some player gets more than somebody else, he feels a penalty, too, because of his bad conscience about that. How much each player cares about those two cases can vary among them, and is set via parameters in the model's "extended" pay-off function Equation 1. We assume though that the aversion towards being unfairly treated by others is stronger than that towards treating others unfairly. Note also that in the model, each player only feels a penalty for unequal results between herself and others, but does not care about inequity between pairs of other players. All that is made precise in the following definition, which corresponds to [5, p. 822, eq. 1]:

Definition 1. Given $n \in \mathbb{N}$ players, let $x \in \mathbb{R}^n$ denote the vector of material pay-offs of those players. Assume that for $i = 1, \ldots, n$, $\beta_i \leq \alpha_i$ and $0 \leq \beta_i < 1$ are given. Then the utility function for player i is given by:

$$U_i(x) = x_i - \frac{\alpha_i}{n-1} \sum_{j \neq i} \max(x_j - x_i, 0) - \frac{\beta_i}{n-1} \sum_{j \neq i} \max(x_i - x_j, 0) \tag{1}$$

Let's examine Equation 1 closely: The first term in $U_i(x)$ is just the "real" game pay-off x_i and can be interpreted as the "base value" of utility, which then is modified by the other two terms which measure the two types of inequity aversion mentioned above. The second term (with the coefficient α_i in front) measures disadvantageous inequity (where player i feels unfairly treated by the others) — for each player $j \neq i$ who has a higher monetary pay-off ($x_j - x_i > 0$, thus the maximum function returns something positive instead of 0), he gets a penalty of the difference weighted by α_i. The division by $n - 1$ is there to normalise the penalty so that α_i can be fixed "independently of n". Similarly, the third term with β_i in front penalises advantageous inequity (where $x_i > x_j$). So α_i measures how much player i cares about being treated unfairly, while β_i does the same for treating others unfairly. Then it makes sense (by our assumptions above) to require $\beta_i \leq \alpha_i$. Furthermore, it is clear that those coefficients should be non-negative, as we want the additional terms in Equation 1 to be really negative penalties. ($\alpha_i < 0$ would mean that someone prefers to be worse off than others, which seems implausible. $\beta_i < 0$ indicates that player i likes to be better off than others, which may well be the case for some people in reality — this is the basis of social status seeking. Nevertheless, we exclude this case.) Finally, Definition 1 also assumes $\beta_i < 1$; to justify that, think about what $\beta_i = 1$ would mean: Then player i is indifferent between getting a higher monetary pay-off x_i, which increases inequity to the other players, and not doing so, which reduces the inequity aversion penalty. (Or as [5, p. 824] puts it: "... player i is prepared to throw away one dollar in order to reduce his advantage relative to player j which seems very implausible.") We want to exclude that case, and of course also the even more extreme cases with $\beta_i > 1$. On the other hand, there's no such natural bound on α_i — it may well be possible that some player prefers to sacrifice her monetary pay-off in order to punish another player for being unfair to her. ("If I can't get that much money, at least he also doesn't get it.") Then, having defined the utility functions U_i of the players, each one strives to maximise not her own pay-off, but rather this utility. Based on that function, we assume rational behaviour — and as we'll see below, this model can explain the observed outcomes of the Ultimatum Game (where it leads to fair behaviour), as well as the outcomes in a simple

market game (see Subsection 2.2) where also in experiments and despite the inequity aversion terms in Equation 1 the outcome can be purely selfish.

Having Fehr's model available, we can apply it to the Ultimatum Game. There, we only have two players (the Proposer and Responder). Let the Proposer be player 1 and the Responder player 2, i. e., (α_1, β_1) are the preferences of the Proposer and (α_2, β_2) of the Responder. For simplicity, assume that the Proposer knows the preferences of the Responder (for the more general case where only a probability distribution of the preferences is known, see [5]). Then the optimal strategy, assuming the utility functions for the players from Definition 1, is (see [5, p. 826, Proposition 1]):

Theorem 1. *For the Responder, reject s if and only if*

$$s < s'(\alpha_2) = \frac{\alpha_2}{1 + 2\alpha_2} < \frac{1}{2}. \tag{2}$$

For the Proposer, offer

$$s^* = \begin{cases} \frac{1}{2} & \beta_1 > \frac{1}{2} \\ s'(\alpha_2) & \beta_1 < \frac{1}{2} \end{cases}. \tag{3}$$

In the case $\beta_1 = \frac{1}{2}$, each offer $s \in \left[s'(\alpha_2), \frac{1}{2}\right]$ results in the same utility.

Proof. First of all, note that it is a trivial estimation to see that $s'(\alpha_2) < \frac{1}{2}$ holds for every $\alpha_2 \geq 0$ with the function s' from Equation 2. Now, in case of rejection, both players get a pay-off of 0 and thus $U_1(s) = U_2(s) = 0$ in that case. If $s \geq \frac{1}{2}$, the offer is advantageous for the Responder. Thus in this case, her utility function is:

$$U_2^a(s) = s - \beta_2(s - (1 - s)) = s - \beta_2(2s - 1) \geq s - (2s - 1) = 1 - s \geq 0$$

(The estimation is valid since $\beta_2 < 1$ by Definition 1.) Hence it is the best strategy for the Responder to accept any $s \geq \frac{1}{2}$. Now, let $s < \frac{1}{2}$. Then the offer is to the Responder's disadvantage and her utility function will be

$$U_2^d(s) = s - \alpha_2((1 - s) - s) = s - \alpha_2(1 - 2s) = (1 + 2\alpha_2)s - \alpha_2.$$

It is then best to accept the offer as long as $U_2^d(s) > 0$, and better to reject if $U_2^d(s) < 0$. (For $U_2^d(s) = 0$, it does not matter.) Solving $U_2^d(s) = 0$ for s, we get precisely the condition asserted in Equation 2.

Consider now the Proposer. We assume that he knows the Responder's preferences, and thus also knows how she will react on the offer (which is what was discussed above). Since $s = \frac{1}{2}$ will always be accepted and yields perfect equality, the offer $s = \frac{1}{2}$ is always better for the Proposer than any larger offer. It remains to consider possible offers $s < \frac{1}{2}$. In that case, assuming the offer will be accepted, the utility of the Proposer is given by (since the result is to his advantage):

$$U_1(s) = (1 - s) - \beta_1((1 - s) - s) = (1 - s) - \beta_1(1 - 2s) = (2\beta_1 - 1)s + (1 - \beta_1)$$

Note first that $U_1\left(\frac{1}{2}\right) = \frac{1}{2} > 0$, which means that all the "best cases" mentioned below are really positive and better than an offer that will be rejected. For $\beta_1 > \frac{1}{2}$, this is strictly increasing in s and hence $s = \frac{1}{2}$ is the best offer. For $\beta_1 = \frac{1}{2}$, it does not depend on s at all, and thus anything between $s'(\alpha_2)$ (so that the offer is accepted) and $\frac{1}{2}$ is possible. For $\beta_1 < \frac{1}{2}$, $U_1(s)$ is strictly decreasing in s, and hence the optimal offer is $s = s'(\alpha_2)$ which is the smallest one that gets still accepted. Taking all together, we find that also the Proposer's best strategy is as claimed in Equation 3. □

Concluding this subsection, we have found that inequity aversion in the players can lead to fair behaviour in the Ultimatum Game — which matches experimental results. If $\beta_1 > \frac{1}{2}$, the Proposer prefers to make a 50–50 offer over one more to his advantage. While this may not be the case for all participants in reality, more importantly, a large α_2 will lead to $s'(\alpha_2)$ in Equation 2 that is near $\frac{1}{2}$ (strictly speaking, $\lim_{\alpha_2 \to \infty} s'(\alpha_2) = \frac{1}{2}$ and convergence is monotone from below) — and hence Responders with sufficiently high α_2 have incentive to punish unfair Proposers (or rather, can convincingly threaten to do so), thus leading to offers which are nearly equal for both players.

2.2 Market Competition

Before concluding the discussion of Fehr's inequity aversion (Equation 1), I want to give another example of its application to an experimental game. This time, we consider a simple market game with Proposer competition — see also [5, p. 829ff]. In this game, we have n players. The first $n-1$ are Proposers, while the last player is a single Responder. Then in a first move, all Proposers choose and publish simultaneously an offer $s_i \in [0,1]$, $i = 1, \ldots, n-1$. (That is, no Proposer knows what the other is going to do and can not react on the strategies chosen by the others.) Afterwards, the Responder can accept the highest offer $\bar{s} = \max_{i=1,\ldots,n-1} s_i$ or reject it. In case of rejection, just as in the Ultimatum Game, nobody gets any payout. If she accepts the offer, then the Responder gets a monetary pay-off of $\bar{s}$, while *one* of the Proposers offering $s_i = \bar{s}$, chosen at random, gets paid $1 - \bar{s}$. All other players receive 0. This game is an aggressively simplified version of market competition between multiple producers, who want to sell some good, and a single buyer. The buyer demands only a single piece or unit of the good and will choose the best offer to buy, if it is not too high for him. All suppliers except the one which offers the best deal gain nothing in the trade.

First, consider what happens if all participants of this game are rational and only interested in the real, material pay-off: Then to have equilibrium, the Responder has to accept any offer $\bar{s} > 0$. If the highest proposed offer is less than 1, other Proposers have incentive to increase their own proposal in order to be chosen and achieve positive pay-off for themselves rather than 0. Thus, the equilibrium can only be at $\bar{s} = 1$. Furthermore, at least two Proposers have to offer $s_i = 1$ so that none has incentive to lower the offering because it would still be accepted. Thus the equilibria of this game are all situations in which at least two Proposers offer 1, and the Responder accepts. This outcome is of course by no means "fair", since all gain from trade is to the advantage of the single Responder! In the equilibrium, no Proposer (even the accepted one) is better off than not trading at all. Now, we can again consider what real players do in experiments. According to [5, p. 829], there exists experimental evidence from different countries that strongly show that the subjects behaved exactly in this way. Despite the unfair outcome, experimental results match the competitive market very well. Even though this outcome is not fair this time (as opposed to the Ultimatum Game in Subsection 2.1), one can show that Fehr's model also predicts these results:

Theorem 2. *The equilibrium outcome based on the utility functions in Equation 1 for the market game with Proposer competition is that at least two Proposers offer $s_i = 1$, and the Responder accepts.*

Proof. See [5, p. 830, Proposition 2] for an informal discussion of this result and [5, p. 856ff, Appendix] for the formal proof. □

So, we've seen that human subjects in experiments sometimes do not behave "rationally" in the sense that they maximise their monetary pay-off. Rather, they seem to prefer fair outcomes over unfair ones — this can be modelled with an utility function that takes result differences between players into account, which leads to Fehr's model of inequity aversion (Equation 1). This successfully explains experimental evidence of more equal outcomes in the Ultimatum Game (Subsection 2.1 and Theorem 1). On the other hand, there also exist experiments — for instance as discussed in Subsection 2.2 — where experimental results match the expectations of a fully competitive market and completely unfair distribution of trade gains. But also these situations can be explained by Fehr's model, see Theorem 2. The key point (which is stressed actually more than once in the discussion in [5, p. 834ff]) seems to be that unfair outcome is possible, if "no single player can enforce an equitable outcome". This is the case for the market game, while in the Ultimatum Game, the Responder *can* reject the offer to force a fair outcome (although to the disadvantage of both) if the offer is too uneven.

3 Matching Problems

In the Introduction, I promised to discuss an application of fair division of resources among applicants — in particular, places for internships among students. But before I can do so in Subsection 3.2, I need to discuss the so-called "Stable Marriage Problem" and the related Gale-Shapley-Algorithm for matching in Subsection 3.1 as prerequisite.

3.1 Stable Marriage Problem

The material discussed in this section was originally published in [9]; but since I didn't have access to that publication, I based my discussion on the review in [16]. The basic problem considered is that of matching participants in a two-sided market one-to-one. For instance, assume that we have finite sets of boys $B = \{b_1, \ldots, b_n\}$ and girls $G = \{g_1, \ldots, g_m\}$. When prom takes place at their school, each one has preferences about whom he or she wants to go dancing with. Lets say that for a boy b, $A(b) \subset G$ denotes all girls that he would accept as partners; instead of going with some $g \notin A(b)$, he'd even rather go alone. Similarly, for $g \in G$, $A(g) \subset B$ denotes the set of acceptable boys she can image to date. Furthermore, each student ranks all acceptable partners by a strict total ordering: For each person $p \in B \cup G$, let $<_p$ denote p's strict total ordering on the set $A(p)$, where $x <_p y$ for $x, y \in A(p)$ means that p prefers to date either x or y to being alone, but would choose y over x. Of course, it is questionable whether such a strict definition is a valid model for being in love or having relationships in real life; for one, the preferences of individuals in their choice of partners are surely not constant over time (and may not even be so over the weeks leading up to prom). Also, it may not even be the case that there exists a strict total ordering with all necessary properties (like transitivity) reflecting the answers someone would give to different people when asked out. But even when one does not believe in this framework for dating (or marriage as in the original work), it can still be applied to other situations (as in Subsection 3.2), so let's focus on the analysis of this model for now. For this, we need two additional notions (see [16, Section 1]):

Definition 2. Given the situation above, a *matching* is a pair of maps (γ, β) with $\gamma : B \to G \cup \{\emptyset\}$ and $\beta : G \to B \cup \{\emptyset\}$, such that $\gamma(b) \in A(b) \cup \{\emptyset\}$ for each $b \in B$ and $\beta(g) \in A(g) \cup \{\emptyset\}$ for each $g \in G$. For $b \in B$, $\gamma(b)$ is b's partner, as is $\beta(g)$ for $g \in G$. The special value $\emptyset$ means that he or she goes alone and is not matched to anyone. To get a meaningful matching, we also require that for all $(b, g) \in B \times G$, $\gamma(b) = g$ if and only if $\beta(g) = b$.

Note that the last property in Definition 2 readily implies that both functions are injective when restricted to arguments that do not map to $\emptyset$, which is of course also an important property of matching partners to each other and reflects the *one-to-one* property: Assume that $\gamma(b_1) = \gamma(b_2) = g \in G$. Then by this assumption, $b_1 = b_2 = \beta(g)$; to show injectivity of β, the same argument is employed the other way round.

Definition 3. A matching (γ, β) is *stable*, if there does not exist any pair of agents $(b, g) \in B \times G$ such that they both prefer being matched to each other over their actual partners, i. e., $\beta(g) <_g b$ and $\gamma(b) <_b g$. (Where we assume that $\emptyset <_p q$ for each person p and all $q \in A(p)$.) Such a pair is called *blocking* for the stability of a matching, if it exists.

In [16], the preferences are technically defined in a different (but on a higher level equivalent) way. Also in addition to pairs blocking stability of a matching, individuals may make a matching unstable, too, if they prefer to stay alone rather than being matched to their actual partner. However, with my framework that's not necessary since by Definition 2, *every valid matching* must respect all individuals' choices of acceptable partners *a-priori*. For the final results and also the basic ideas and understanding, it does not matter which formal set of definitions is chosen, though.

Gale and Shapley showed in [9] that for all preferences there exists a stable matching according to Definition 3. They proved it in a constructive way by presenting their "deferred matching algorithm" that can construct such a matching. To understand how that procedure works, consider first how the matching may be performed in a real situation of students going to prom: They would ask each other out and decide on whether to accept or reject a proposal based on their preferences. Let's assume the school decided that in the current year there should be ladies' choice for prom. Then each girl would ask her most preferred boy first. Some proposals would be accepted by the partners, and some rejected (because her favourite doesn't want to go out with her at all or has proposals by girls he likes better). The rejected girls would then go on and ask the still free boys in descending order of preference until they find partners or no boys acceptable to them remain free. The problem in this situation is that a boy, let's call him Mark, when asked out by some girl, say, Jessie, does not know whether Sarah, whom he likes better than Jessie, is going to ask him out later. If he goes with Jessie, he can't accept Sarah if she asks him later; but if he rejects and Sarah wants to go with someone else, he may find himself completely alone in the end. Thus, he may be compelled to accept Jessie even though he likes Sarah

better, just because Jessie asked first. But then, if Sarah really asks Mark and he has to reject, it may happen that Sarah goes alone and asks Mark for just one dance at the prom, but then they decide to stick together the whole evening. In other words, the matching would not be stable in this case because Mark may leave Jessie during the evening. (Formally, (Mark, Sarah) is the pair of agents that violates stability in Definition 3 in this example.)

The key to avoiding this problem is the word *deferred*: In the algorithm of Gale and Shapley, the final assignment is deferred until all proposals have been made. In other words, for the more conservative system of boys asking girls out, each girl may get proposals even if she already has a not-yet-rejected proposal. In that way, she can wait and see whether a boy whom she likes better than her current partner is going to ask her out, and reject the first partner only after that. More precisely, the algorithm goes like this (see [16, Section 1] again):

1. For each $i = 1, \dots, n$ with $A(b_i) \neq \emptyset$, b_i asks his first choice out.
2. Each girl g_j, with $j = 1, \dots, m$, rejects all proposals made to her by boys not in $A(g_j)$. If more than one proposal remains, she rejects all but the one by the most preferred boy. The not rejected proposals are kept "on hold".
3. Each boy who has been rejected in the last step asks the most preferred girl who hasn't yet rejected him out or goes alone if none remain.
4. Each girl rejects proposals as in step 2, also taking into account any proposal she may be "holding" from last steps.
5. Repeat steps 3 and 4 until no more proposals are made. Then match every girl to the boy whose proposal she is holding (if any).

Theorem 3. *The algorithm terminates after a finite number of steps (actually, the number of steps as well as the runtime time complexity can be bounded by $O(nm)$) and the produced matching is always stable. Consequently, a stable matching exists for all possible preferences.*

Proof. See also [16, Theorem 1]. At every iteration of the algorithm, at least one proposal is made. But since at most nm proposals can be made (each boy proposing to every girl, remember that no boy may every propose twice to a single girl!), it follows immediately that the runtime is bounded from above by $O(nm)$ and the algorithm must terminate.

Furthermore, it is clear that the algorithm produces a valid matching. Assume that the result is not stable. This means that there exists a pair (b, g) such that b prefers g over his assigned partner $\gamma(b)$ and vice-versa. But according to the algorithm, this implies that he proposed to g before he did to $\gamma(b)$. Thus, he had to be rejected by g, but that means that necessarily $b <_g \beta(g)$, which is a contradiction to the original assumption that g prefers b over her actual partner. □

Let's get back to the main topic of fairness: Theorem 3 shows that the matching produced by the Gale-Shapley-Algorithm is good and incorporates the preferences of the individuals to some degree, producing results that the participants like. However, it is also trivial to see that it is not symmetric with respect to the two parties (boys and girls). One of them proposes, while the other takes the role of responding to proposals. We'll see below that this is not only a technical detail but in fact introduces an unfair bias between the two sexes in the matching (see [16, Theorem 2]):

Definition 4. A matching is called *B-optimal* if every boy $b \in B$ likes his matched partner $\gamma(b)$ at least as good as all other partners $\gamma'(b)$ that can be matched to him in any stable matching. Similarly, a matching is called *G-optimal* if the reverse holds for every girl $g \in G$.

Theorem 4. *The matching produced by the Gale-Shapley-Algorithm with boys proposing is always B-optimal, and if the girls propose it is G-optimal.*

Proof. Assume without loss of generality that the boys propose and further that the produced matching is not B-optimal. Then there exists some $b \in B$ and a $g \in G$ such that $\gamma(b) \neq g$ but b and g could be matched together in a different stable matching, and $\gamma(b) <_b g$. But by the algorithm, b proposed to g before his actual partner $\gamma(b)$ (since g is preferred over $\gamma(b)$), and since $g \neq \gamma(b)$, g must have rejected him at some time. But that means that $(\beta(g), g)$ blocks any matching from being stable where b and g are partners, which is a contradiction. □

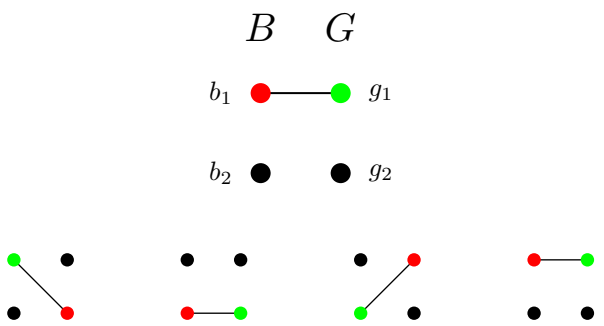

Figure 2: Example of a cycle in the naive matching idea. Lines connect partners matched together at each step, while single points denote people who are currently alone. Red marks unhappy partners who break up in the next step since they are part of a blocking pair with one of the singles, while green dots represent partners who are "optimally happy" with the current relationship.

Clearly, a B- or G-optimal matching is preferred by boys and girls, respectively, over non-optimal matchings. We've seen in Theorem 4 that it is always possible (and in fact, produced automatically by the presented algorithm) to achieve optimality for one side. However, it is not necessarily so that a matching is also optimal for the others — see Example 1 below for a simple counter-example. Thus the algorithm may be as good as it can be for producing a fair matching under certain definitions of fairness (in honouring the individual preferences), but it always discriminates against one of the two sides of the market. Note that it may be counter-intuitive at the first thought that girls should be worse off than boys in the matching, when boys propose but girls have the ultimate decision over accepting or rejecting proposals. However, the real advantage of the proposing side comes from the fact that its agents make the first move by proposing to the others in decreasing order of their own preference; and only if the top candidate is not available, they have to try with the less favourable ones. The responding agents on the other hand have the power to accept and reject proposals according to their will, but can not influence at all who proposes to them and thus can't act on their own to get the partner of their choosing.

Before concluding the section with the promised Example 1, I also want to briefly discuss a simpler idea for a matching algorithm one could have: Considering again the definition of stability, Definition 3, it is evident that the crucial point for stability are possible blocking pairs. So, a different approach for producing a stable matching could be to start with an arbitrary valid matching, and if a blocking pair exists, modify it such that the blocking pair is matched together (possibly breaking up other matchings and leaving the respective other partners alone for the moment). This could then be repeated until no more blocking pairs are there, which means that the produced matching is stable. In some sense, this is what could happen in reality with couples and relationships over time, and it seems intuitive that also this procedure should eventually result in a stable matching. However, the problem is that this may lead to infinite cycles and thus the algorithm does not always terminate; this is another thing shown by Example 1:

Example 1. As promised above, I will now conclude this section with a very simple example of the algorithm in operation; besides showing how to apply the description, this also clearly shows that the produced matching may only be optimal for one side, namely the proposing one. Assume that we have a total of four actors, two for each sex (i. e., $B = \{b_1, b_2\}$ and $G = \{g_1, g_2\}$). For each one, all persons of the other sex are acceptable ($A(b_1) = A(b_2) = G$ and $A(g_1) = A(g_2) = B$); the preferences are such that

$$b_2 <_{g_1} b_1, \quad b_1 <_{g_2} b_2, \quad g_1 <_{b_1} g_2, \quad g_2 <_{b_2} g_1.$$

Then, with boys proposing, b_1 proposes to g_2 and b_2 to g_1 in the first step. Neither is rejected, and so the matching produced is (b_1, g_2) and (b_2, g_1) — which is clearly B-optimal but not G-optimal. On the other hand, if girls propose, the same thing happens in reverse: g_1 proposes to b_1 and g_2 to b_2, also with no rejections; thus, the algorithm produces in this case the G-optimal (but not B-optimal) matching (b_1, g_1) and (b_2, g_2).

Finally, Figure 2 shows how the situation may "evolve" if we simply bring together a blocking pair at each step — which leads to a cycle and thus never finishes. The fundamental principle there is that

if only one couple is matched and the others are single, one of the partners in the couple prefers one of the singles over his or her actual partner — who also prefers him or her over being alone. So they come together in the next iteration, forming another couple with a happy and one not-so-happy party and another pair of singles, and the same thing happens again until we're back at the initial set-up. One can see in Figure 2, that only after four iterations the initial configuration (which is big on top) is reached again in the bottom row.

3.2 Two-Stage Matching for ELSA Internships

After I explained the basic principles of the Gale-Shapley-Algorithm in the previous Subsection 3.1, I want to briefly mention and discuss an application of this algorithm (in a slightly extended form) to a real problem now. This subsection is based on [13]. ELSA is the "European Law Students' Association", which is an international students organisation made up of local groups. The organisation provides internships for its members, with the basic idea that local groups talk to possible employers and negotiate possible internships, which are then provided for students of other groups. Historically, the assignment of students to internships was based on bilateral exchange — for each internship that a local group provided for a member of some other local group, one of its members was also accepted into an internship of that other group. In newer times, however, the organisation changed the procedure used to allocate internships to students; now, all provided internships go to an internal pool, and every student from every local group can apply for any internship therein — no matter how many internships his or her local group provided to the pool. The change was made because this "more open" matching system seems to fit better with the idea of a global network of students. Unfortunately, after the change was made, the number of available internships started to decrease despite the fact that an ever-increasing number of students applied for the places. The problem is likely that in the new system, local groups don't have any real incentive to provide internships other than because of their "good will"; there is no benefit to their local members if they provide more places than another group, as this no longer influences how many of its own members it can provide with internship places.

The authors of [13] assisted ELSA and tried to draft a new matching procedure that would help overcome this problem, while still keeping the allocation process more to the spirit of an internal association than reverting to the original bilateral exchange would. Here I want to briefly discuss the method used in [13] for this goal. But first, I have to give some more details about how the application and selection process works: Each student is allowed to apply for three internship places and rank them according to her own preferences; on the other hand, each employer also ranks the student applications by his preferences. Then, we have to match students to internship places in a way that tries to respect the individual preferences as much as possible — that's where the Gale-Shapley-Algorithm (see the above Subsection 3.1) comes into the game. However, we can't apply the algorithm in its vanilla form, as that would just be the current matching system — additional measures are needed to create incentives for local groups to provide more internship places. The solution suggested by [13] is to make the matching *two-stage*:

1. Assume that the local group i (with $i = 1, \ldots, N$) provided n_i places to the pool. Then apply the Gale-Shapley-Algorithm (Subsection 3.1) to match the members of this local group to $p \cdot n_i$ places, where $p \in [0, 1]$ is a parameter.

2. Match all students independently of their local groups' performances in providing internships to the remaining places, again using the rules of Gale-Shapley matching.

That is, some fraction (determined by p) of the places a group provides is "reserved" for members of that group — which gives it an incentive to provide more places. However, not all places are matched in this way — the remaining ones are open for every student to apply, thus retaining a certain amount of openness in the system. Note that by choosing $p = 1$, each local group is allowed to apply for exactly as many places as it provided — which gets us (mostly) back to the original bilateral exchange. In the other extreme with $p = 0$, all places are allocated in the open pool. Thus this new system allows to provide a graded trade-off between the two methods used up to now, such that the system can be both open to a certain degree and also encourage local groups to try hard in getting more internship places for the organisation.

4 Voting Systems and Arrow's Theorem

Of course, usually for social decisions in any society (like the formulation of laws or usage of public goods), different members have different opinions on what to do. When there are multiple alternatives, there has to be some system to decide which one the society as a whole should choose, depending on the preferences of the individual people. Examples can be things like electing a party to govern or a new president, or also what movie to watch in the cinema with your friends. A rather simple approach to solving this problem is dictatorship: A single person decides with his own preferences what should be done. This is of course not very fair by any definition — and so in modern society political decisions are usually made by some kind of democratic voting. Each member has a chance to vote for her preferred alternative (or rank the alternatives in some way), and then a *voting system* is used to decide upon the "aggregated" preferences of the voters together. This section is devoted to discussing several aspects about voting systems and their overall fairness. Note that this topic was one of the problems done at the "Woche der Modellierung mit Mathematik" in 2011, where students worked on it under guidance of Stephen Keeling of the "Institut für Mathematik und wissenschaftliches Rechnen" of the University of Graz. By discussing this topic with him beforehand, I was only made aware of Arrow's theorem and this interesting field in general. To acknowledge that, I based this section mostly on [12] as well as my discussions with him.

4.1 Arrow's Impossibility Theorem

Let's start with a famous, but unfortunately negative, result: The Impossibility Theorem by Arrow. See [3] for an original work or [14, around 6:30] as well as [12, p. 64f] for an introduction. The basic situation of voting is assumed to be as follows: We have n voters $V = \{v_1, \dots, v_n\}$ as well as k alternatives $A = \{a_1, \dots, a_k\}$ to vote for. Furthermore, assume that each voter $v \in V$ has individual preferences in the form of a strict total ordering on A, denoted by "$<_v$". Then a voting system is a function f mapping all possible preferences of individuals $(<_{v_1}, \dots, <_{v_n})$ onto another total ordering $<$ on A, which is said to be the group preference and according to which the preferred alternatives are then chosen and executed by politics. In order to achieve "fair" voting, we require certain conditions of the voting system f, namely at least the following:

Unanimity Let $a, b \in A$ be two alternatives. If every voter prefers a over b (i. e., $\forall v \in V : b <_v a$), then also the group preference should do so ($b < a$).

Independence If for two input preference sets $(<_{v_1}, \dots, <_{v_n})$ and $(\prec_{v_1}, \dots, \prec_{v_n})$, each voter ranks some alternatives $a, b \in A$ relative to each other in both sets the same (that is, $a <_v b \Leftrightarrow a \prec_v b$ for each $v \in V$), then also the resulting group preferences should have the same relative ordering of a and b: $a < b \Leftrightarrow a \prec b$. In other words, the relative ranking of two alternatives should only depend on their ordering in the individual preferences, and not on irrelevant third alternatives.

Nondictatorship There must not be any voter $v \in V$ such that $< = <_v$ always holds independent of the preferences of the other voters. (That is, no one must be allowed to determine the group preference with his or her own preferences alone for all cases.)

Then Arrow's theorem, here given without proof, states:

Theorem 5. *Given at least three alternatives ($k \geq 3$) and two voters ($n \geq 2$), no voting system f can satisfy all three conditions above.*

This may be a surprising result (since we did only require things that seem natural for a fair voting system). Furthermore, it makes clear that voting based on individual rankings, as we usually do, is fundamentally flawed if we want a fair system other than dictatorship. There are some ways to relax the conditions such that Theorem 5 no longer applies (see the end of [14] for instance), but none seems to be fully satisfactory. In Subsection 4.3 I will briefly discuss a different voting principle, which does not take total orderings of alternatives as inputs and which is also not restricted by Theorem 5.

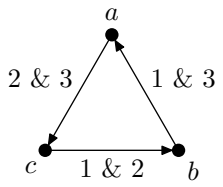

Figure 3: Paradox of voting (see Example 2), where we have a "cycle" in the majority preferences. An arrow denotes majority preferences, pointing to the preferred alternative (following the arrows leads to better alternatives). Note that this convention is the other way round as in [14], from which the example is taken.

4.2 SK Prinzip: Resistance instead of Majority

We are used to vote by majority in many situations: Each person is allowed to vote for her favourite alternative, and then the alternative with the most "positive" votes is chosen. However, according to the "paradox of voting", it is not always possible to find a majority winner. This can be demonstrated already by a rather simple example, see Example 2 (from [14, around 3:30]):

Example 2. We use the situation of Subsection 4.1 above. Then assume that we have three voters $V = \{1, 2, 3\}$ and also three alternatives $A = \{a, b, c\}$. The preferences are:

$$c <_1 b <_1 a, \;\; a <_2 c <_2 b, \;\; b <_3 a <_3 c.$$

In this case, a majority of 2 to 1 prefers a over b, but also two persons prefer b over c, and finally, again a majority is for c over a; this is shown graphically in Figure 3. Thus it is impossible to choose any alternative over another for the whole group by means of majority voting, we have a draw between all three alternatives!

Another problem with majority voting is that it may result in alternatives that are unloved by most people, simply because a very united minority pushed for them and gave them more votes than any single other alternative — even though the majority of voters disagrees with the chosen alternative. Examples for this are given on [8]. The solution proposed there is to vote with *resistances* ("negative" votes). Each voter may declare which alternative he likes *the least* rather than the most, and in the end the one with the least resistance votes wins. This is of course not fundamentally different to positive voting (and is subject to the same problem as in Example 2), but it empirically seems to produce better results in a lot of cases. This principle of resistance votes is also examined in [12, p. 67ff] in detail, including voting in multiple stages. One reason for why the outcome is often liked better by participants with resistance votes could be the following: Assume there are n alternatives to choose from, and preference for them is (approximately) uniformly distributed among voters. Then with positive votes, about $\frac{1}{n}$ of the voters will have their wish satisfied, while the majority of $\frac{n-1}{n}$ will be disappointed with the result. On the other hand, if we do resistance voting, only $\frac{1}{n}$ will not get the desired outcome, and $\frac{n-1}{n}$ will feel that democracy works out. So the point is that with more than two alternatives (with $n = 2$, of course both voting systems are equivalent anyway) more expressed preferences for *not* choosing some particular alternative can be simultaneously satisfied than expressed preferences *for* some or the other.

4.3 Fractional Voting System

While the voting so far was based on stating preferences of individuals in form of an ordering of the alternatives (as we did in Subsection 4.1 right at the beginning), the "Fractional Voting System" (see [15]) uses different input: Each voter states his or her preferences in form of a *preference distribution* for the candidates. In this way, it is not only known that some candidate is preferred over another, but also "how much". This could be done in practice, for instance, by allowing everyone to cast more than one vote and distribute the votes to the alternatives as they like (including multiple votes to a single alternative). Then, the fractional voting system determines the group preference *distribution* from the individual distributions by simply summing up all votes one candidate got from all voters. Since this is grounded on a different framework, it is not subject to Theorem 5. Again, one may state conditions that are required for a fair voting system (which I do only informally here, see [15, p. 17] for more details):

Unanimity If each voter gives the same fraction of individual preference to some alternative, also the voting system gives that fraction to it.

Independence Given the individual preference fractions, the group preference is calculated using the same function for each alternative. That is, we don't discriminate between alternatives.

Note that the axioms are similar to the corresponding ones in Subsection 4.1; only nondictatorship is missing. However, it is clear that the fractional voting system as defined above is not dictated by any one (since we sum up all votes for each candidate and thus the results *do* depend on the preferences of *each voter*). Then one gets a nice theorem, kind of opposite to Arrow's (see [15, Possibility Theorem]):

Theorem 6. *In the fractional voting system, it is possible to satisfy the conditions above. Even more, it is the unique group preference distribution that does so.*

References

[1] Allgemeines bürgerliches Gesetzbuch. Austrian law.

[2] Anonymous. *The Bible, Old and New Testaments, King James Version*, volume 10. Project Gutenberg, P.O. Box 2782, Champaign, IL 61825-2782, USA, 1989.

[3] Kenneth J. Arrow. A Difficulty in the Concept of Social Welfare. *The Journal of Political Economy*, 58(4):328–436, August 1950.

[4] asdala.de. Das väterliche Erbe. `http://www.asdala.de/raetsel/02.html#das_vaeterliche_erbe`. Online; accessed 2012-01-31.

[5] Ernst Fehr and Klaus M. Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, pages 817–868, August 1999.

[6] Martin Flechl. Hex. `http://mflechl.web.cern.ch/mflechl/hex.html`. Online; accessed 2012-01-31.

[7] Jürgen Fleiß and Ulrike Leopold-Wildburger. Inequity Aversion and Reciprocity. An Experimental Within-Subject Design Study. Submitted to *Games*, Department of Statistics and Operations Research, University of Graz.

[8] Institut für Systemisches Konsensieren. Das SK-Prinzip. `http://www.sk-prinzip.eu/`. Online; accessed 2012-02-27.

[9] D. Gale and L. S. Shapley. College Admissions and the Stability of Marriage. *American Mathematical Monthly*, pages 9–14, 1962.

[10] The Code of Hammurabi. `http://www.general-intelligence.com/library/hr.pdf`, 1915. Translated by L. W. King. Online; accessed 2012-01-31.

[11] Alastair Hudson. *Equity & Trusts*. Cavendish Publishing, London, 4th edition, 2005.

[12] Stephen L. Keeling et al. Entwicklung eines Wahlsystems. In *Woche der Modellierung mit Mathematik: Dokumentationsbroschüre*, pages 64–76, Pöllau bei Hartberg, February 2011. `http://math.uni-graz.at/modellwoche/2011/Broschuere-2011.pdf`.

[13] Ulrike Leopold-Wildburger and Dominik Stigler. Designing a 2-Stages-Deferred-Acceptance Algorithm. Working paper, Department of Statistics and Operations Research, University of Graz.

[14] H. Reiju Mihara. Arrow's Impossibility Theorem and ways out of the impossibility. `https://www.youtube.com/watch?v=RqOTigJ1xuk`. Talk by the author. Online; accessed 2012-02-27.

[15] K. K. Nambiar. Arrow's paradox and the fractional voting system. `http://www.e-atheneum.net/science/fractional_voting_screen.pdf`, November 2001. Online; accessed 2012-02-27.

[16] Alvin E. Roth. Deferred Acceptance Algorithms: History, Theory, Practice, and Open Questions. *International Journal of Game Theory*, Special Issue in Honor of David Gale on his 85$^{\text{th}}$ birthday(36):537–569, March 2008.

[17] Ian Stewart. The Riddle of the Vanishing Camel. *Scientific American*, 266(6):122, 1992.