How to make a difference – measures of voting power revamped*

Claus Beisbart and Luc Bovens

13th July 2009

Abstract

Voting power (i-power) measures the extent to which a vote can make a difference to the outcome of a collective decision. And a voter has the opportunity to make a difference, if the following counterfactual is true: Had the vote been different, the outcome of the collective decision would have been different. In the philosophical literature, several interpretations of counterfactuals have been suggested. One of them leads to the probability that a vote is critical and thus (if the Bernoulli model is adopted) to the Banzhaf measure. But there is arguably a more plausible interpretation of counterfactuals according to which counterfactuals trace causal connections. We provide a measure of voting power that is based on this very interpretation. The measure makes use of probabilistic causal networks. We motivate and define the measure, provide simple examples, and discuss the relation to the Banzhaf measure. We conclude by suggesting how the measure may be used for quantifying the responsibility of a voter.

1 Introduction

In this paper we propose a new measure of voting power. A measure of voting power is meant to quantify the extent to which a person's vote has the opportunity to make a difference to the outcome of a collective decision (see Felsenthal & Machover 1998, p. 35–6, e.g.). The new measure coincides with the Banzhaf measure of voting power under certain conditions, but yields markedly different results, if the conditions are violated. Our measure is based upon causal information and is thus particularly interesting for measuring causal responsibility (see Vallentyne 2008; Braham & van Hees 2009 for this topic).

*This paper is to be presented at the Voting Power in Practice Workshop at the University of Warwick, 14–16 July 2009, sponsored by The Leverhulme Trust (Grant F/07-004/AJ). `This is a preliminary version, please don't circulate.`

In some respect, our proposal further develops an idea suggested by Bovens & Beisbart (2009). Like Bovens & Beisbart (2009), we use probabilistic causal networks for measuring voting power. But unlike Bovens & Beisbart (2009), we do not calculate an average treatment effect. Rather, a different measure is suggested. Also, our motivation is different. All in all, we think, that our measure has certain advantages over the D-measure suggested by Bovens & Beisbart (2009).

The plan of the paper is as follows. Sec. 2 provides philosophical motivation for our approach. We develop the mathematical details of our approach in Sec. 3. We turn to measures of causal responsibility in Sec. 4.

2 Power and counterfactuals

In order to get started, we provide an argument that motivates the standard approach to measure voting power – it leads to the Banzhaf measure of voting power and a generalization thereof. Our claim is not that this reconstruction captures the actual train of thoughts that lead to the Banzhaf measure. Rather we provide a reconstruction in order to put our fingers on certain points that we find problematic and in order to introduce our own approach.

Our reconstruction comes in five steps.

1. Conceptual point: The voting power of a voter is the extent to which her vote can make a difference as to what the outcome of a collective decision is (Felsenthal & Machover 1998, pp. 35–6). It is agreed that the "can" here refers to opportunity or ableness, in Morriss' terms (Morriss 1987, Ch. 11).

   This point may partly be understood as a conceptual analysis; but to some extent, it simply delineates of what one is interested in.

2. A model for making a difference: It is assumed that each voter has two votes only, say yes and no, and that there are only two possible outcomes, say yes and no. Under this model, a vote makes a difference, iff the following counterfactual is true:

   CF If the vote were different/If the voter cast a different vote, the outcome of the collective decision would be different.

   The importance of conditionals for understanding power is also stressed by Morriss (1987), particularly Ch. 9. In his terms, CF is an act conditional, which is in turn a sort of manifestation conditional (see p. 62).

3. A proposal to measure extent of opportunity. We want to quantify the extent to which a vote can make a difference. This extent is usually measured in terms of the respective probability (Felsenthal & Machover 1998, p. 36 suggest that there is really no alternative).

It follows that our measure of voting power for a voter is the probability to which the counterfactual CF is true.

4. Interpretation of the counterfactual CF. For quantifying the probability of CF, an interpretation of CF is needed. In the standard approach, the truth conditions of CF are thought to be equivalent to a vote being critical. A vote i is critical regarding a bipartition B of the votes (see Felsenthal & Machover 1998, Def. 2.1.5 on p. 14), if B and a variation under which only the sign of $B(i)$ is reversed have different outcomes. In more colloquial terms, i is critical in a given voting profile, if just switching vote i and keeping the other votes, produces a different outcome.

   This analysis of CF may be underwritten in Lewisian terms (Lewis 1973). Describe possible worlds in terms of voting profiles/bipartitions. Measure the closeness of worlds by specifying the number of individual votes that are different for two worlds. Thus world w_1 has a distance of 2 to world w_2, iff exactly two individual votes are different. Then, given Lewis's analysis of counterfactuals, the above analysis of the counterfactual emerges.

   It follows that a measure of voting power will have to be the probability of criticality.

5. Specification of a probability model over voting profiles. Usually, the Bernoulli model is assumed for calculating the probability that a vote is critical. Under the Bernoulli model, each possible voting profile has the same probability (Felsenthal & Machover 1998, p. 37).

After steps 1 – 5, we end up with the Banzhaf measure of voting power. The Banzhaf measure is the probability of criticality under the Bernoulli model. When instead the most realistic probability model is used, an alternative measure of voting power arises. This measure of voting power has been used by Beisbart & Bovens (2008); it was criticized by Machover (2007); a slight modification has been suggested by Kaniovski & Leech (2009).

We will now put our fingers on some points that we find problematic in the train of thought that we have suggested underlies the standard approach.

ad 1 The conceptual point above misses the distinction between the power that a voter has on a particular occasion and generic voting power (Morriss 1987). In fact, people are mostly interested in generic voting power, but this will not always do for what follows below.

ad 3 It is not clear why the extent to which CF is true should be measured in terms of probability and what sort of probability is involved. One way of interpreting probabilities is to regard them as expressing degrees of belief or ignorance. But if this interpretation is adopted, then we should say that

what we are calculating is our best estimate of a measure of voting power, and not the measure of voting power itself.

ad 4 It has been argued that alternative interpretations of counterfactuals are more fitting, as far as our every day understanding of counterfactuals is concerned. In particular, it has been suggested that counterfactuals trace causal connections (see Hiddleston 2005 for a recent example). This point can be explained using the following example.

Consider simple majority voting with three people, A, B and C. Assume that A is an opinion leader for B. That is, after A has voted, B will simply copy A's vote. Consider now a particular occasion in which A, B and C vote no. Consider the counterfactual

CF_1 If A voted differently (viz. yes rather than no), the outcome of the decision would have been different (viz. yes rather than no).

Under the interpretation that underlies the standard approach, this counterfactual would clearly be false in our situation. But intuitively we are more inclined to reason as follows: If A voted differently, viz. yes rather than no, B would follow suit. Accordingly, at least two votes would be yes, which would be sufficient for acceptance. If this line of thought is correct, then CF_1 is true rather than false.

It is thus more fitting to interpret counterfactuals using causal relationships. This does not necessarily lead us beyond the possible world semantics. A causal interpretation of counterfactuals can be obtained, if closeness of worlds is defined appropriately (see Balke & Pearl 1994).

In the following, we provide a framework for measuring voting power. In this framework the problems that we have identified with the standard analysis, can be solved.

3 A new framework for measuring voting power

Our framework will be introduced in two ways. In Subsec. 3.1 we work through a number of examples of increasing complexity. In Subsec. 3.2 we explain our framework from scratch using notions well-known from causal modeling. Readers that are familiar with causal modeling may skip Subsec. 3.1.

3.1 A series of examples

Unless otherwise stated, our examples in this subsection share the following assumptions.

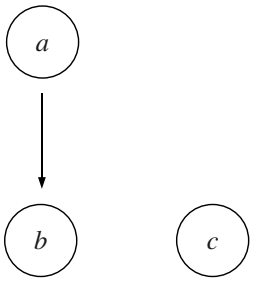

Figure 1: The causal model underlying our examples.

Assumptions 3.1.1 *There is simple majority voting with three voters that we call A, B and C. Call their votes a, b, c, respectively. If A votes yes, a is set at 1; otherwise it is set at 0. We proceed analogously for the other voters. The decision rule (in fact, any decision rule for three people) can be represented as a function D from $\{0,1\}^3$ to $\{0,1\}$. In our notation $D(a,b,c) = 1$, iff $a + b + c \geq 2$, and zero otherwise.*

We assume causal relations between the votes and depict them with arrows. We assume that there is a causal arrow from a to b, since A is an opinion leader for B. Otherwise, no causal arrows are assumed. Our causal model is visualized in Fig. 1.

3.1.1 Simple deterministic causation and full information on the voting profile

Assume first that the causal influence between a and b is particularly simple: B always copies A's vote. In this case B's vote is a function of A's vote: $b = f(a)$ with f being the identity. Assume now that a particular collective vote has been taken. In the real world, $a = a_0, b = b_0, c = c_0$. For a concrete example, assume that all voters vote no, i.e. $a_0 = b_0 = c_0 = 0$. The outcome of the decision is $D(a_0, b_0, c_0)$ (0 in the concrete example). All of this is known to us as well as the causal model. We are interested in the voting power of A on this occasion. Clearly, that voting power depends on what we are to say regarding the following counterfactual

CF_2 If A voted differently on this occasion, the outcome of the vote would be different.

Now, clearly, we have now enough information to judge whether CF_2 is true or false. No probabilities are needed. For instance, in our concrete example, we can say that, if A voted yes, B would follow suit, and the result would be different. Let us therefore assign A a power of 1 in this example.

More generally, we suggest that the voting power of A, $v(\mathrm{A})$ is calculated as follows:

$$v(\mathrm{A}) = |D(a_0, b_0, c_0) - D(1 - a_0, f(1 - a_0), c_0)| . \quad (1)$$

Let us now briefly turn to B and to the following counterfactual:

$\mathrm{CF}_{2'}$ If B voted differently on this occasion, the outcome of the vote would be different

for the same situation $(a = a_0, b = b_0, c = c_0)$.

There is no causal arrow that goes from B's vote to another vote. Thus, if B voted differently, no other voter would switch her vote. In particular, A would still cast the vote a_0. This is different from our treatment of CF_2. A and B's votes have to be treated differently in this respect, because causality is asymmetric. There is a causal arrow from a to b, but not the other way round. No causal backtracking is permitted (cf. Lewis 1979). As a consequence, we obtain for B's power

$$v(\mathrm{B}) = |D(a_0, b_0, c_0) - D(a_0, 1 - b_0, c_0)| . \quad (2)$$

Equally, the power of C is

$$v(\mathrm{B}) = |D(a_0, b_0, c_0) - D(a_0, b_0, 1 - c_0)| . \quad (3)$$

Note that B and C have power 1 just in case their vote is critical in the voting profile (a_0, b_0, c_0).

3.1.2 More complicated deterministic causation and full information on the voting profile

In case A's vote has a causal bearing on B's vote it is not realistic to assume that B always copies A's vote – that B's vote is a function of A's vote. More realistically, B may follow A in some cases, but not in others. This situation may be modeled as follows (see Balke & Pearl 1994). We define a parameter ϵ that specifies the way in which B's vote reacts to A's vote. To each value of ϵ corresponds a function f_ϵ that describes B's vote as a function of A's vote:

$$\begin{aligned} \epsilon = 0 \quad & b = f_1(a) = a , \\ \epsilon = 1 \quad & b = f_1(a) = 1 - a , \\ \epsilon = 2 \quad & b = f_1(a) = 1 , \\ \epsilon = 3 \quad & b = f_1(a) = 0 . \end{aligned}$$

For instance, in the last case, B votes no regardless of what A does.

Now, if the actual votes and the value of ϵ are known on a particular occasion, then we can clearly assign CF_2 a definite truth value. The voting power of A is thus 1 or 0:

$$v(\mathrm{A}) = |D(a_0, b_0, c_0) - D(1 - a_0, f_\epsilon(1 - a_0), c_0)| \quad (4)$$

with the appropriate ϵ.

Suppose now that the value of ϵ is not known on the particular occasion. In this situation we can only estimate the true power of A by taking the expectation value wrt ϵ. Estimates will be denoted with a tilde. We assume a probability model for ϵ. The different functions f_ϵ and the probability model for ϵ yield a *functional model* for the random variables a and b (Balke & Pearl 1994). In taking the expectation value, it is to be noted that the actual voting profile puts some constraints on the value of ϵ. For instance, if every vote is no, ϵ has to be 0 or 3. The expectation value over ϵ conditional on the information at hand will be denoted by $\mathbb{E}_{\epsilon|a_0,b_0,c_0}$. We thus obtain:

$$\tilde{v}(\mathrm{A}) = \mathbb{E}_{\epsilon|a_0,b_0,c_0} | D\left(a_0, b_0, c_0\right) - D\left(1-a_0, f_\epsilon\left(1-a_0\right), c_0\right) | \,. \tag{5}$$

For B we proceed analogously and take into account that no backtracking is allowed. We obtain:

$$\tilde{v}(\mathrm{B}) = \mathbb{E}_{\epsilon|a_0,b_0,c_0} | D\left(a_0, 1-b_0, c_0\right) - D\left(a_0, b_0, c_0\right) | \,. \tag{6}$$

Here the expectation value can in fact be neglected, because ϵ does not figure in the expression in the scope of the expectation value operator.

For a more concrete example, let us assume that the actual votes are $a_0 = b_0 = c_0 = 0$. We assume a probability model over ϵ. For instance, let $P(\epsilon = 0) = .2, P(\epsilon = 1) = .15, P(\epsilon = 0) = .25, P(\epsilon = 0) = .4$. Given the actual voting profile, ϵ can only be 0 or 3. If $\epsilon = 0$, CF_2 is true; if $\epsilon = 3$, CF_2 is false. In the first case, the power of A is 1, in the latter zero. Our estimate of A's power is thus the probability that $\epsilon = 0$, given that $\epsilon = 0$ or $\epsilon = 3$. That is, $\tilde{v}(A) = P(\epsilon = 0)/\left(P(\epsilon = 0) + P(\epsilon = 3)\right) = 1/3$. Note that the following assumption entered our calculation: The value of a has no bearing on the value of ϵ. Put differently, ϵ and a are thought to be independent. This will be explained in more detail below.

But how may we obtain the probability model over the ϵ? This probability model is constrained by the joint probability model for a and b. For instance, it can easily be shown that

$$P(b = 1|a = 1) = P(\epsilon = 0) + P(\epsilon = 2) \,. \tag{7}$$

However, in general, the probability model for ϵ is not fully determined by the joint probability model for a and b.[1] This is well-known from causal modeling (see Pearl & Verma 1991 and Balke & Pearl 1994, pp. 3–4). If you want, a one-parameter family of probability models for ϵ is compatible with the joint probability model for a and b. Additional knowledge is required for fixing this parameter and for obtaining a full functional model.

[1] There are exceptions though. For instance, if a and b are fully correlated, then the model for ϵ is uniquely determined.

Here is an example how the probability model for the votes may underdetermine the probability model for ϵ.

Since A is an opinion leader for B, the votes of A and B are positively correlated. Let $P(a = 1) = P(b = 1) = .5$ and $P(b = 1|a = 1) = P(b = 0|a = 0) = .9$. This fixes the probability model over the votes of A and B. But it does not fully determine the probability model over ϵ – as it were, different "mechanism" may account for the joint probability model. For instance, B always copies A's vote on issues of social policy ($\epsilon = 0$), where such issues arise with a probability of .85. B always votes yes on issues regarding education ($\epsilon = 2$), the probability of such issues arising being .05. B always votes no on issues regarding foreign policy, where the probability for such issues is .05 ($\epsilon = 3$). On all other issues (probability of .05) B votes exactly the other way than A ($\epsilon = 1$); . From this probability distribution over ϵ, we can derive the joint probability model specified for a and b.

But we obtain the same joint probability model for a and b, if we assume that the different issues have different probabilities, e.g. $P(\epsilon = 0) = .9$, $P(\epsilon = 1) = .1$, $P(\epsilon = 2) = 0$ and $P(\epsilon = 3) = 0$.

3.1.3 Indeterministic causation and full information on the voting profile

Suppose now that the causal relation between the votes a and b is indeterministic. That is, the value of a has a bearing on b, but there is no set of parameters such that, given specific values for these parameters, the relation between a and b is a function.

Even in this situation, we may formally introduce an ϵ parameter as before. The only difference is that this parameter does not reflect determinate conditions in the world. Rather, it is an artificial means for calculating probabilities. With its help, we can proceed as above and simply calculate the expectation value as before. This time, the result can in fact be regarded as a measure of power (not just as an estimate), because there is no possible information that would allow us to go further. The voting power is thus

$$v(\mathrm{A}) = \mathbb{E}_{\epsilon|a_0,b_0,c_0} |D(a_0, b_0, c_0) - D(1 - a_0, f_\epsilon(1 - a_0), c_0)| . \tag{8}$$

There is, however, a severe problem with this equation. As we have said before, it is not possible to extract a unique probability model for ϵ from the joint probability model over a and b. And in an indeterministic world, there is no mechanism that we could investigate in order to find the model for ϵ. Thus, in an indeterministic world, a voting power for a person on a particular occasion in which the voting profile is know cannot be specified (only bounds can be derived).

There is an alternative to working with a functional model, but we cannot go into this here.

3.1.4 Incomplete information on the voting profile

Assume now that we do not have full information on the actual voting profile. For simplicity we immediately turn to a case in which nothing is known about the actual voting profile – intermediate cases will be considered in the next subsection. The counterfactual we are interested in reads

CF_3 If A voted differently, then the outcome of the collective decision would be different.

How can we specify the extent to which CF_3 is true? How can we measure the voting power of voter A in this setting?

Note that the setting is compatible with different interpretations. We may still be interested in a single occasion, but not know what all (some) voters did. In this case, the probabilities over the votes are fittingly interpreted as degrees of (rational) belief. But if a joint probability model over all votes is assumed, we may also be interested in generic voting power. The idea might be that we have an series of collective decisions ahead, and we are interested how much power the voters have generically in view of this series. The probabilities that we use may be the limit of relative frequencies that we obtain for suitable series' of collective decisions (we will not delve into the problems of spelling out a sensible frequentist account of probabilities). In this case, a more objectivist understanding of the probabilities over the votes seems possible at least.[2]

It depends on the interpretation of the probabilities involved whether we obtain an estimate on the voting power or a measure of voting power itself. For simplicity, in the following, we will speak of voting powers.

In any case, we may proceed as follows. We go through each possible combination of the values for ϵ, a, b and c. Note that some of these combinations are not possible, since the values of a and of ϵ uniquely fix the value of b. In each particular case, we proceed as we did above, when we had information on the full voting profile, and we obtain a measure of voting power. This measure is multiplied with the respective probability for having the specific values of ϵ, a and c for the case. Finally, these results are added up.

In carrying out this algorithm, it has to be noted that ϵ, a and c are pairwise probabilistically independent. This is actually the point of introducing the variable ϵ: ϵ states how b depends on a. There is no need here to let ϵ be dependent on a. The behavior of two dependent random variables a, b is reduced to two independent random variables a and ϵ, where ϵ specifies the way b depends on a.

We illustrate our algorithm using a concrete example. We continue to use the causal model under which A is an opinion leader for B. As a consequence, the votes of A and B are positively correlated. We assume that $P(a = 1) = P(b = 1) = .5$

[2]Of course, very often, when there are objective chances, we will only have estimates of these objective chances, and in this sense, a subjective component will be present. Our point here is only that in these cases, talk of objective probabilities makes sense.

ϵ	a	b	c	o	a'	b'	c'	o'	$\lvert o - o' \rvert$	P
0	1	1	1	1	0	0	1	0	1	$.25\kappa$
0	1	1	0	1	0	0	0	0	1	$.25\kappa$
0	0	0	1	0	1	1	1	1	1	$.25\kappa$
0	0	0	0	0	1	1	0	1	1	$.25\kappa$
1	1	0	1	1	0	1	1	1	0	$.25(1.8 - 2\kappa)$
1	1	0	0	0	0	1	0	0	0	$.25(1.8 - 2\kappa)$
1	0	1	1	1	1	0	1	1	0	$.25(1.8 - 2\kappa)$
1	0	1	0	0	1	0	0	1	0	$.25(1.8 - 2\kappa)$
2	1	1	1	1	0	1	1	1	0	$.25(.9 - \kappa)$
2	1	1	0	1	0	1	0	0	1	$.25(.9 - \kappa)$
2	0	1	1	1	1	1	1	1	0	$.25(.9 - \kappa)$
2	0	1	0	0	1	1	0	1	1	$.25(.9 - \kappa)$
3	1	0	1	1	0	0	1	0	1	$.25(.9 - \kappa)$
3	1	0	0	0	0	0	0	0	0	$.25(.9 - \kappa)$
3	0	0	1	0	1	0	1	1	1	$.25(.9 - \kappa)$
3	0	0	0	0	1	0	0	0	0	$.25(.9 - \kappa)$

Table 1: An example. We consider the voting power of A and assume that no knowledge is present on the actual votes.

and that $P(b = 1|a = 1) = P(b = 0|a = 0) = .9$, as before. Let us assume that these correlations arise from a functional model for which $P(\epsilon = 0) = \kappa$, $P(\epsilon = 1) = .9 - \kappa$, $P(\epsilon = 0) = .9 - \kappa$ and $P(\epsilon = 0) = 1.8 - 2\kappa$. Here κ is a real number in the interval $[.8, .9]$. Considering the various possible values of κ, we obtain any functional model that is compatible with the joint probability model on a and b.

Since, according to our assumptions, there is no causal arrow that connects c with any of the other votes, c is probabilistically independent of a and b. We assume that $P(c = 1) = .5$.

Our calculation is summarized in Table 1. Here, a, b and c denote the votes that may be cast in the actual world; a', b' and c' denote the votes that would be cast, if A voted differently. o is the outcome that arises in the actual world for a particular voting profile o' is the counterfactual outcome. We obtain a contribution to A's power, if $o \neq o'$.

As a result we obtain that A's power is .9. Remarkably, this result is independent of the value of κ. Thus, in our example, the underdetermination of the functional model disappears – what exactly the functional model is does not matter, if it is only compatible with the joint probability model over a and b. This is not just an artifact of our particular model, but holds true more generally. For calculating voting power, this is very advantageous, because we can now calculate voting power, even if causation is indeterministic.

We can also determine the power for B and C. Calculating these powers is in fact easier than calculating the power of A, because the value of ϵ does not matter in these cases. If B switches her vote, there are no consequences on the other votes.

We have now worked through a small series of examples. Clearly, more complicated examples could be given. In particular, we could conceive of more complicated causal models. For instance, the votes of A and C may be correlated due to a common cause. Common causes can also be understood in terms of functional models. But instead of working through more complicated models, we will rather provide a general framework for dealing with voting power.

3.2 The full framework

If we are interested in the voting power that a particular person has, then we want to know to which extent a counterfactual – something like CF above – is true. We want to put a probability on a counterfactual. Depending on the details, that probability may only express our ignorance or capture objective chance present in indeterministic causation or refer to a series of votes (when we are interested in generic causation).

Balke & Pearl (1994) have suggested a method to put probabilities on counterfactuals – in their terms, they specify the probability of a counterfactual query. We suggest to use that approach in order to quantify voting power. In this section, we will first recall their account briefly. We will then use that account to quantify voting power.

A counterfactual query may be specified in three steps. First, some real world information is assumed. In a probabilistic framework, this means that values of some random variables are given; say $a = a_0$, $b = b_0$ etc. Second the antecedent of the counterfactual is specified. In a probabilistic framework, this amounts to specifying values of random variables once more. In order to make clear that the values are just *assumed* to obtain, we introduce random variables that correspond to our original random variables, but that refer to the counterfactual world, call them $\hat{a}$ etc.[3] The antecedent thus assigns values to some of these random variables. Typically some of these values will be incompatible with the values of the corresponding random variables in the real world. Finally, a certain consequence in the counterfactual world – i.e. the consequent of the counterfactual – is specified. In our examples, the consequent always concerns the outcome of a collective decision. For instance, we may be interested in the probability that the outcome under some counterfactual conditions is yes.

Roughly following Balke & Pearl (1994), we adopt the following notation. The probability of the counterfactual query:

[3]In the following we speak of one counterfactual world for simplicity. In the view of the possible world semantics it may be more appropriate to say that the second causal probabilistic network specifies a class of counterfactual worlds.

CQ Would $\hat{o} = o_2$ hold in a counterfactual world in which $\hat{b} = b_1, \hat{c} = c_1$, given that in the real world $a = a_0$, $b = b_0$?

is denoted by

$$P\left(\hat{o} = o_2 \middle| a = a_0, b = b_0; \hat{b} = b_1, \hat{c} = c_1\right) . \tag{9}$$

How can we calculate this probability?

We assume that we have a causal probabilistic model over the relevant real world random variables a, b etc. The causal relationships can be displayed in a directed acyclic graph in which the random variables are nodes. The sets of parents of random variable $a, b, \ldots$ are called $pa(a), pa(b), \ldots$. Furthermore, we have disturbance variables – they play the role that our ϵ played above. To each random variable $a, b, \ldots$ corresponds a disturbance variable $\epsilon_a, \epsilon_b, \ldots$. The idea is this: If the values of the parents of a random variable a and the value of the corresponding ϵ_a are specified, then the value of a is uniquely fixed. Put differently, there are (deterministic) functions $f_a, \ldots$ for each random variable $a, \ldots$ such that $a = f_a(pa(a), \epsilon_a), \ldots$. The ϵ_a-variable is dispensable in case a has no parent random variable. The disturbance variables are thought to follow a probability model; under that model, they are probabilistically independent. From the deterministic functions and the probability model over the disturbance variables (which form the functional model), the probability model of the random variables $a, b, \ldots$ follows.

The idea is furthermore that there is a second probabilistic causal network with the $\hat{a}, \ldots$ as nodes for the counterfactual world. The DAGs of the networks are isomorphic, if a is mapped to $\hat{a}$ etc. Moreover, we have disturbance variables $\epsilon_{\hat{a}}, \cdots$. It is assumed that the functional relationships in the counterfactual world are the same as in the real world, i.e. $\hat{a} = f_a(pa(\hat{a}), \epsilon_{\hat{a}})$ (f_a, not $f_{\hat{a}}$!). The probability model over the disturbance variables (and thus the other random variables) in the counterfactual world can be left open at this point.

Given this framework, we proceed as follows for calculating the probability $P\left(\hat{o} = o_2 | a = a_0, b = b_0; \hat{b} = b_1, \hat{c} = c_1\right)$.

1. We insert the real world information $a = a_0, b = b_0$ in the real-world network. We obtain the joint probability distribution for the disturbance variables $\epsilon_a, \ldots$, given $a = a_0, b = b_0$. Note that, in general, the disturbance variables will be stochastically dependent *given* $a = a_0, b = b_0$ (they are only independent unconditionally). The joint probability distribution that we obtain is also thought to govern $\epsilon_{\hat{a}}, \ldots$ in the following (i.e. the disturbance variables in the counterfactual world).

2. We turn to the network of the counterfactual world. We have already a probability model for the $\epsilon_{\hat{a}}, \ldots$. We now prune arrows from the parents of the variables that figure in the antecedent of the counterfactual, i.e. $\hat{b}, \hat{c}$.

The reason is that we do not want to allow for backtracking. Then we set $\hat{b}$ at b_1 and $\hat{c}$ at c_1.

3. We spread the evidence in the network for the counterfactual world. I.e. for each variable apart from $\hat{b}$ and $\hat{c}$ we infer the probability model. We obtain a probability for $\hat{o}$ being o_2. This is the probability of our counterfactual query.

How do we apply this algorithm in voting theory? We think of the votes as random variables $a, \ldots$ The outcome is modeled as another random variable, call it o. Clearly, o is a deterministic function of the votes (no disturbance variable is needed at this point). This function is just the voting rule D. We thus assume arrows from each random variable that models a vote to o. Moreover, there will in general be causal arrows between the votes. Often, there will also be correlations between votes due to common causes. For instance, two votes may often be the same, because the corresponding voters share the same political values. We model these common causes as additional random variables. All in all, the causal probabilistic network for the real world consists of votes, the outcome and common causes. Likewise, there are the hatted variables in the counterfactual world.

We can now use the above algorithm in order to put probabilities on counterfactual queries such as „Given that A and B voted yes in the real world, if A had voted no, the outcome would have been yes".

The algorithm is very general and flexible. We can apply the algorithm in settings in which we have full knowledge over the actual voting profile, but also in settings in which we do not have this full knowledge.

There is a little subtlety though in cases in which we do not know the actual outcome or A's actual vote. Let us for simplicity focus on a situation in which we know just nothing about the values of the random variables in the actual world. In such cases, we are interested in the counterfactual:

CF_4 The outcome of the vote would be different, if A voted differently.

This counterfactual cannot directly be dealt with in the Balke-Pearl algorithm, because no specific real-world information is given. The counterfactual world is only specified *relative* or *comparing* to the actual world through the "different", but it is left open how the real world is like. Our proposal here is to go through all possible cases of specific values of a and o, to calculate the probabilities of the corresponding counterfactual queries and to add them up weighing with the probabilities that a and o take the specific values. That is, we calculate the following sum

$$\sum_{i_a=0}^{1}\sum_{i_o=0}^{1} P\left(\hat{o}=1-i_a | a=i_a \wedge o=i_o; \hat{a}=1-i_a\right) \times P\left(a=i_a \wedge o=i_o\right) \ . \quad (10)$$

There are four addends in this sum. They correspond to the following cases: 1. Due to A's switching her vote from no to yes, the result changes from no to yes; 2. Due to A's switching her vote from no to yes, the result changes from yes to no; 3. Due to A's switching her vote from yes to no, the result changes from no to yes; 4. Due to A's switching her vote from yes to no, the result changes from yes to no. Here the second and the third case are a bit weired; there must be people that are "countersuggestive" to take up an expression by Grice (1989) – they tend to be influenced by A, but not quite the way one would expect for an opinion leader.

In case we know A's vote in the actual world, but not so the outcome; or in case we know the actual outcome, but no so A's vote, we proceed in an analogous manner. Only two addends need to be added in these cases.

The results that we obtain in this way are exactly identical to our results in Subsec. 3.1. In particular, going through Table 1, as we did in Subsec. 3.1.4 yields the same result as Eq. (10) and our algorithm from this subsection do.

Of course, for carrying out the above algorithm, a functional model needs to be known. In some special cases, any functional model that is compatible with the joint probability model over the votes will yield the same result, but this is not always so. Thus, when the functional model is not known, we will in general not be able to put a probability on the counterfactual – only bounds may be obtained.

For an illustration, let us briefly reconsider the concrete example of Subsec. 3.1.4 in our new framework. That is, we have no knowledge on the actual votes. We need thus calculate Eq. (10). Let us first consider the first addend

$$P\left(\hat{o}=1|a=0\wedge o=0;\hat{a}=1\right)\times P\left(a=0\wedge o=0\right)\ . \tag{11}$$

Fortunately, we need not calculate the first factor in this product $P\left(a=0\wedge o=0\right)$ at this point; let us simply call its value q. For obtaining the second factor, we consider $P\left(\epsilon_b=\epsilon_0\wedge c=c_0|a=0\wedge o=0\right)$ for $\epsilon_0,c_0\in\{0,1\}$. Obviously, in each case,

$$P\left(\epsilon=\epsilon_0\wedge c=c_0|a=0\wedge o=0\right)=P\left(\epsilon_b=\epsilon_0\wedge c=c_0\wedge a=0\wedge o=0\right)/q\ . \tag{12}$$

Now a, ϵ_b and c are pairwise independent. Moreover, concrete values for these random variables uniquely fix an outcome. Thus, for $P\left(\epsilon_b=\epsilon_0\wedge c=c_0\wedge a=0\wedge o=0\right)$, we obtain the product of the respective probabilities for a being a_0, ϵ_b being ϵ_0 and c being c_0, if the corresponding outcome o is 1, and zero otherwise. Altogether,

we have:

$$
\begin{aligned}
&P\left(\epsilon_b = 0 \wedge c = 1 | a = 0 \wedge o = 0\right) = .25 \times \kappa/q \ , \\
&P\left(\epsilon_b = 0 \wedge c = 0 | a = 0 \wedge o = 0\right) = .25 \times \kappa/q \ , \\
&P\left(\epsilon_b = 1 \wedge c = 1 | a = 0 \wedge o = 0\right) = 0 \ , \\
&P\left(\epsilon_b = 1 \wedge c = 0 | a = 0 \wedge o = 0\right) = .25 \times (1.8 - 2\kappa)/q \ , \\
&P\left(\epsilon_b = 2 \wedge c = 1 | a = 0 \wedge o = 0\right) = 0 \ , \\
&P\left(\epsilon_b = 2 \wedge c = 0 | a = 0 \wedge o = 0\right) = .25 \times (.9 - \kappa)/q \ , \\
&P\left(\epsilon_b = 3 \wedge c = 1 | a = 0 \wedge o = 0\right) = .25 \times (.9 - \kappa)/q \ , \\
&P\left(\epsilon_b = 3 \wedge c = 0 | a = 0 \wedge o = 0\right) = .25 \times (.9 - \kappa)/q \ . \qquad (13)
\end{aligned}
$$

This is also the probability model over $\epsilon_{\hat{b}}$ and $\hat{c}$. Furthermore, in the counterfactual world, we have $\hat{a} = 1$. What then is the probability that we have acceptance in the counterfactual world – i.e. that $\hat{o} = 1$? Well, first, no arrow is to be pruned. Second, since $\hat{a} = 1$ we need to make sure that at least one of $\hat{b}$ and $\hat{c}$ equals 1. Since $\hat{a} = 1$, we have $\hat{b} = 1$, iff $\epsilon_{\hat{b}}$ equals 0 or 2. Thus, from the list in Eqs. (13), only the fourth and the eighth line do not contribute. We add up the other lines and obtain for the first addend in Eq. (11)

$$
.25 \times \kappa + .25 \times \kappa + .25 \times (.9 - \kappa) + .25 \times (.9 - \kappa) = .9 \times .5 \ . \qquad (14)
$$

Since the model is completely symmetric, if all votes are switched to their opposite, we obtain $.9 \times .5$ for the addend

$$
P\left(\hat{o} = 0 | a = 1 \wedge o = 1; \hat{a} = 0\right) \times P\left(a = 1 \wedge o = 1\right) \ . \qquad (15)
$$

The remaining addends are identical once more for symmetry reasons. We focus on

$$
P\left(\hat{o} = 1 | a = 1 \wedge o = 0; \hat{a} = 0\right) \times P\left(a = 1 \wedge o = 0\right) \ . \qquad (16)
$$

This product concerns the case in which A switches from yes to no, but the outcome switches the other way round. Call $P\left(a = 1 \wedge o = 0\right) = q'$. We obtain for the conditional probabilities

$$
\begin{aligned}
&P\left(\epsilon_b = 0 \wedge c = 1 | a = 1 \wedge o = 0\right) = 0 \ , \\
&P\left(\epsilon_b = 0 \wedge c = 0 | a = 1 \wedge o = 0\right) = 0 \ , \\
&P\left(\epsilon_b = 1 \wedge c = 1 | a = 1 \wedge o = 0\right) = 0 \ , \\
&P\left(\epsilon_b = 1 \wedge c = 0 | a = 1 \wedge o = 0\right) = .25 \times (1.8 - 2\kappa)/q' \ , \\
&P\left(\epsilon_b = 2 \wedge c = 1 | a = 1 \wedge o = 0\right) = 0 \ , \\
&P\left(\epsilon_b = 2 \wedge c = 0 | a = 1 \wedge o = 0\right) = 0 \ , \\
&P\left(\epsilon_b = 3 \wedge c = 1 | a = 1 \wedge o = 0\right) = 0 \ , \\
&P\left(\epsilon_b = 3 \wedge c = 0 | a = 1 \wedge o = 0\right) = .25 \times (.9 - \kappa)/q' \ . \qquad (17)
\end{aligned}
$$

This is also the probability model over $\epsilon_{\hat{b}}$ and $\hat{c}$ in this case. Furthermore, in the counterfactual world, we have $\hat{a} = 0$. What then is the probability that we have acceptance in the counterfactual world – i.e. that $\hat{o} = 1$? Well, since $\hat{a} = 0$ we need to make sure that both of $\hat{b}$ and $\hat{c}$ equal 1. Since $\hat{a} = 0$, we have $\hat{b} = 1$, iff $\epsilon_{\hat{b}}$ equals 1 or 2. From the list in Eqs. (17), only the third and the fifth line contribute to our addend, but they are zero; thus our addend is zero.

Altogether, in Eq. (10), only two addends are not zero, but rather have the value $.5 \times .9$. Altogether, we obtain a voting power of $v_{\mathrm{A}} = .9$, as we did before. Given the fact that A is an opinion leader, this measurement of voting power seems plausible.

3.3 Relation to other measures

But how is our new measure related to other measures?

Let us first consider the probability of criticality as a measure of power. That measure arises, if we follow the steps 1 – 4 in our reconstruction in Section 2. It contains the Banzhaf measure of voting power for the special case of the Bernoulli probability model.

The following theorem can be proven.

Theorem 1 *Suppose the real-world model for the votes does not have any arrow from A's vote to another vote. It follows that* v_{A} *for the case of no knowledge on the voting profile, i.e., the sum in Eq.* (10)*, equals the probability that A is critical.*

Thus, if A's vote does not causally influence any other vote, our new measure coincides with the above measure. This is so, even if the A's vote and some other votes are correlated.

The proof of the theorem is immediately clear from the method in Subsec. 3.1.4, which we have claimed to be equivalent with evaluating Eq. (10) following the algorithm.

Let us now compare to the D-measure that Bovens & Beisbart (2009) propose. There are two conceptual differences. First, the D-measure arises by adding two weighted differences of probabilities. Thus, the idea behind the measure is the raising of probabilities. This is not so with our new measure. An important consequence is this. The D-measure may take negative values (in case a probability is lowered rather than raised). In contrast, our new measure is always positive. This also shows that our new measure is different from the D-measure.

Second, it is true that some probabilities of counterfactuals enter the D-measure, and these probabilities are based upon evaluating a causal probabilistic network. But they are not the probabilities that we evaluate here; in the probabilities entering the D-measure it is not assumed that a specific result obtains in the real world.[4]

[4]Also, Bovens & Beisbart (2009) do not work with disturbance variables. It is beyond the

4 Responsibility

In this section we consider probabilities for counterfactual queries to investigate measures of responsibility.

References

Balke, A. & Pearl, J., *Probabilistic evaluation of counterfactual queries*, in: *Proceedings on the Twelfth National Conference on Artificial Intelligence*, 1994.

Beisbart, C. & Bovens, L., *A Power Measure Analysis of Amendment 36 in Colorado*, Public Choice **124** (2008), 231 – 246.

Bovens, L. & Beisbart, C., *Measuring influence for dependent voters: A generalisation of the banzhaf measure*, Texts in Logic and Games **6** (2009).

Braham, M. & van Hees, M., *Degrees of causation*, to appear in Erkenntnis, 2009.

Felsenthal, D. S. & Machover, M., *The Measurement of Voting Power: Theory and Practice, Problems and Paradoxes*, Edward Elgar, Cheltenham, 1998.

Grice, P., *Logic and Conversation*, in: *Studies in the Ways of Words* (Grice, H. P., ed.), Harvard University Press, Cambridge (MA), 1989, pp. 1–143.

Hiddleston, E., *A causal theory of counterfactuals*, Noûs **39** (2005), 632–657.

Kaniovski, S. & Leech, D., *A Behavioural Power Index*, forthcoming in Public Choice; URL: http://serguei.kaniovski.wifo.ac.at/fileadmin/files_kaniovski/pdf/bpower.pdf, 2009.

Lewis, D., *Counterfactuals*, Blackwell, Oxford, 1973.

Lewis, D., *Counterfactual Dependence and Time's Arrow*, Noûs **13** (1979), 455 – 476, also in Lewis, D., Philosophical Papers. Volume II, Oxford University Press, New York 1986, pp. 32 – 52.

Machover, M., *Discussion Topic: Voting Power when Voters' Independence is not Assumed*, mimeo, URL: http://eprints.lse.ac.uk/2966/, 2007.

Morriss, P., *Power. A Philosophical Analysis*, Manchester University Press, Manchester, 1987, second edition 2002.

Pearl, J. & Verma, T., *A theory of inferred causation*, in: *Principles of Knowledge Representation and Reasoning: Proceedings of the Second Internationl Conference* (San Mateo, CA), Morgan Kaufmann, 1991, pp. 441 – 452.

Vallentyne, P., *Brute luck and responsibility*, Politics, Philosophy and Economics **7** (2008), 57–80.

scope of this paper whether this makes a difference.