Strategies of persuasion, manipulation and propaganda: psychological and social aspects

Michael Franke & Robert van Rooij

Abstract

Some of us believe that we can change the world for the better, and that it doesn't take many committed individuals to do so. We are not necessarlly trapped in Hardin's (1968) tragedy of the commons. *Others* believe that with enough commitment and sophisticated communication strategies they can at least change the world for the better *for themselves*, whatever their motives might be. Perhaps this belief is based on naïve and ungrounded optimism. The goal of our investigation here is to become a bit less naïve. That's why this paper asks: how can one influence the behavior of others? What is a good persuasion strategy? It is obviously of great importance to determine *what* information is best to provide and, given that our decisions might be dependent on how this information is framed (c.f. Kahnemann and Tversky, 1973), it is also important to determine *how* to formulate this information. The paper therefore reviews basic findings of decision and game theory on models of strategic communication with an emphasis on models that incorporate aspects of bounded rationality and certain psychological biases. But there is also the social issue of manipulation, concerned with determining *who we should address* so as best to promote our opinion in a larger group or society as a whole. The second half of this paper therefore looks at a novel extension of DeGroot's (1974)'s classical model of opinion dynamics that allows agents to strategically influence some agents more than others. Given the complexity of this social choice problem, we report the results of numerical simulations to sketch what counts as a good, yet resource effective social manipulation heuristics.

1 Pragmatic and social aspects of manipulation

You might be an artist, politician, banker, merchant, terrorist, or, what is likely given that you are obviously reading this, a scientist. Whatever your profession or call of heart, your career depends, whether you like it or not, in substantial part on your success at influencing the behavior and opinions of others in ways favorable to you (but not necessarily favorable to them). Those who put aside ethical considerations and aspire to be successful manipulators face two major challenges. The first challenge is the most fundamental and we shall call it pragmatic or one-to-one. It arises during the most elementary form of manipulative effort whenever a single manipulator faces a single decision maker (henceforth DM) whose opinion of behavior the former likes to influence. The one-to-one challenge is mostly, but not exclusively, about *rhetoric*, i.e., the proper use of logical arguments and other, less normatively compelling, but perhaps

even more efficiently persuasive communication strategies. But if manipulation is to be taken further, also a second challenge arises and that is social or many-to-many. Supposing that we know (roughly) *how* to exert efficient influence on individual DMs, it is another issue *who* to exert influence on in a group of DMs, so as to efficiently propagate a choice or an opinion in a society.

This paper deals with efficient strategies for manipulation at both levels. This is not only relevant for aspiring master manipulators, bur also for those who would like to brace themselves for a manipulative environment like our present-day society. In fact, one of the main themes of Section 2, which deals with the pragmatic one-to-one aspects of manipulation, is that standard models from decision and game theory predict that the DM would see through any manipulative effort and therefore neither much manipulation, nor a lot of other communication is possible. Since this verdict flies in the face of empirical evidence, we feel forced to extend our investigation to more psychologically adequate models of boundedly rational agency. Towards this end, we will review in section 3 models of (i) unawareness of the game/context model, (ii) limited step-by-step reasoning, and (iii) descriptive decision theory. We will suggest that assuming that people are only boundedly rational can help to explain why we talk so much.

Whereas Section 2 has an overview character in that it summarizes key notions and insights from the relevant literature (with an emphasis on more recent developments), Section 3 seeks to charter new territory. Section 3 formulates and explores a model of social opinion dynamics, i.e., a model of how opinions spread and develop in a population of agents, which also allows agents to choose whom to influence and whom to neglect. This way, this part of the paper addresses the social dimension of persuasion. We argue that also here *heuristics* play an important role. Since the complexity of the propaganda problem, as we will call it, is immense, and it is therefore unrealistic to assume that a truly optimal solution can always be found in real life situations, the need for simple yet efficient heuristics arises. We try to delineate in general terms what a good heuristic strategy is for social manipulation of opinions and demonstrate with a case study simulating the behavior of four concrete heuristics in different kinds of social interaction structures that (i) strategies that aim at easy targets, so to speak, are efficient on a short time scale, while strategies that aim at influential targets are efficient at a later time scale, and that (ii) it helps to play a coalitional strategy together with other likeminded manipulators, in particular so as not to get into one another's way.

When we speak of a strategy here, what we have in mind is mostly a very loose and general notion, much like the use of the word "strategy" in non-technical English, when employed by speakers merrily uninterested in any geeky meaning contrast between "strategy" and "tactic". When we talk about a 'good' strategy, we mean a communication strategy that influences other agents to act, or have an opinion, in accordance with our own preferences. This notion of communication strategy is different from the one used in other contributions to this volume.

Within game theory, the standard notion of a strategy is that of a full contingency plan that specifies at the beginning of a game which action an agent chooses whenever she might be called to act. When we discuss strategies of games in Section 2 as a formal specification of agents behavior, we do too use the term in this specific technical sense.

In general, however, we talk about strategic manipulation from a more God's-eye point of view, referring to a good strategy as what is a good general principle which, if realized in a concrete situation, would give rise to a "strategy" in the formal, game theoretic sense of the term.

Gabriel Sandu's paper in this volume appears to be closely related to our work, but there are fundamental conceptual differences between Sandu's and our approach. Sandu works in the tradition of Hintikka's game semantics. Game semantics is an approach to formal semantics that seeks to ground the concepts of truth, meaning, and validity in terms of a dialogical view of communication and of winning strategies. Although both Sandu's and our paper deal with meaning and communication, there are major differences. Whereas his paper is primarily about *semantics*, and the grounding of truth, we focus on (perhaps non-Gricean) *pragmatics*, on how agents can be influenced by means of communication. Whereas in Sandu's paper the dialogues can be very long, and the meaning of the dialogue moves is always clear, in our paper dialogues are typically short, but it is hard to determine what was meant by a dialogue move. The roles of the participants of the dialogues is very different as well. In Sandu's semantic games, there are only two participants with fixed goals: either to verify or falsify a given formula. It is common knowledge between the participants that their goals are diametrically opposed. Although when we talk about one-to-one communication, we also limit ourselves to communication games with only two participants involved, the emphasis of the later part of our paper is on influencing whole groups of agents. Equally important, the goals of our agents are not diametrically opposed. If that were common knowledge in one-to-one communications, we would predict no communication at all. In our case, communication rather presupposes common interest, and the strategy of our communicators is to persuade their hearers that their preferences are very much in common.

2 Pragmatic aspects of persuasion and manipulation

The pragmatic dimension of persuasion and manipulation chiefly concerns the use of language: although other means are conceivable, the main question is what should a manipulator *say* to elicit a response from the DM that is optimally beneficial to the manipulator. Persuasive communication of this kind is studied in rhetoric, argumentation theory, politics, law, and marketing. But more recently also pragmatics, the linguistic theory of language use, has turned its eye towards persuasive communication, especially in the form of game theoretic pragmatics. This is a very welcome development, for two main reasons. Firstly, the aforementioned can learn from pragmatics: a widely used misleading device in advertisements—a paradigmatic example of persuasion—is *false implication* (e.g. Kahane and Cavender, 1980). A certain quality is claimed for the product without explicitly asserting its uniqueness, with the intention to make you assume that only their product has that quality. Persuasion by false implication is reminiscent of the well-studied phenomenon of pragmatic *conversational implicature* (e.g. Levinson, 1983). Secondly, the study of persuasive communication *should* really be a natural part of linguistic pragmatics. The only reason why persuasion has been neglected for long is due to the fact that the prevalent theory of language use in linguis-

tics is based on the Gricean assumption of *cooperativity* (Grice, 1975). Though game theory can formalize Gricean pragmatics, its analysis of strategic persuasive communication is suitable for noncooperative situations as well. Indeed, game theory is the natural framework for studying strategic manipulative communication.

To show this, the following Sections 2.1 and 2.2 introduce the main setup of decision and game-theoretic models of one-to-one communication. Unfortunately, as we will see presently, standard game theory falsely predicts that in non-cooperative circumstances, reliable communication is extremely limited, but so is successful manipulation. This is because, according to standard theory, rational agents would basically see through any attempt of manipulation. Hence rational manipulators would not even try to exert malign influence. Section 2.3 therefore looks at a number of models in which classical assumptions, such as perfect knowledge of the decision situation or common belief in rationality, are levelled and thus create wiggle room for various manipulative strategies. In particular, we briefly cover models of (i) language use among agents who are possibly unaware of relevant details of the decision-making context, (ii) language use among agents who are limited in their depth of strategic thinking, and (iii) the impact that certain surprising features of our cognitive makeup, such as *framing effects* (Kahnemann and Tversky, 1973), have on decision making.

2.1 Decisions and information flow

On first thought it may always seem helpful to provide truthful information, and noncooperative, if not misleading to lie. But this first impression is easily seen to be wrong. For one thing, it can sometimes be helpful to lie. For another, providing truthful but incomplete information can sometimes be harmful. In other words, the story is more complicated than immediately strikes the common sense. Fortunately, intuition can be profitably aided by means of *decision* and *game theory*.

According to classical decision theory, rational DMs choose their actions so as to maximize their expected utility. Suppose that our DM is confronted with the decision problem whether to choose action a_1 or a_2, while uncertain which of the states $t_1, \ldots, t_6$ is actual:

$\mathrm{U}(a_i, t_j)$	t_1	t_2	t_3	t_4	t_5	t_6
a_1	-1	1	3	7	-1	1
a_2	2	2	2	2	2	2

If the agent considers each state equally probable, the theory of decision predicts that the agent will choose action a_2 because that has a higher expected utility than action a_1: action a_2 gives a sure outcome of 2, but a_1 only gives an expected utility of $5/3 = 1/6 \times \sum_i u(a_1, t_i)$. If t_1 is the actual state, the DM has made the right decision. This is not the case, however, if, for instance, t_3 were the actual state. It it is now helpful for the DM to recieve the false information that t_4 is the actual state: falsely believing that t_4 is actual, the DM would choose the action which is in fact best in the actual state t_3. And of course, we all make occasional use of *white lies*: communicating something that is false in the interest of tact or politeness.

More interesting than white lies, however, is providing truthful but misleading information. Suppose that the agent receives the information that states t_5 and t_6 are not the case. After updating her information state (i.e., probability function) by standard conditionalization, rationality now dictates our DM to choose action a_1 because that now has the highest expected utility: $5/2$ versus 2. Although action a_1 was perhaps the most rational action to choose, this doesn't mean that she has thus also made the right decision. Indeed, one can argue that she made the *wrong decision* if it turns out that t_1 is the actual state. One can conclude that receiving truthful information is not always helpful, and can sometimes even hurt.

Communication helps to disseminate information. In many cases, receiving truthful information is helpful: it allows one to make a better informed decision. But we have just seen that getting truthful information can be harmfull as well, at least when the information is partial and does not rule out all risk of making the wrong decision. However, because *in general* it is better to make decisions based on more information than on less, we tend to appreciate receiving information, and are happy to rely on it in making our decisions. This tendency, however, is dangerous: it can be made use of by other agents. Indeed, this tendency admits other agents to manipulate our behavior in their own advantage. Suppose, for instance, that the manipulator prefers our agent to perform a_1 instead of a_2, independently of which state actually holds. If the DM and the manipulator are both ideally rational, the informer will realize that it doesn't make sense to provide, say, information $\{t_1, t_2, t_3, t_4\}$ with misleading intention, because the DM won't fall for this and will consider information to be *incredible*. A new question comes up: how much can an agent credibly communicate in a situation like that above? This type of question is studied by economists making use of signaling games.

2.2 Signaling games and credible communication

Signaling games are the perhaps simplest non-trivial game-theoretic models of language use. They were invented by David Lewis to study the emergence of conventional semantic meaning (Lewis, 1969). For reasons of exposition, we first look at Lewisean signaling games where messages do not have a previously given conventional meaning, but then zoom in on the case where a commonly known conventional language exists.

A signaling game proceeds as follows. A sender S observes the actual state of the world $t \in T$ and chooses a message m from a set of alternatives M. In turn, R observes the sent message and chooses an action a from a given set A. The payoffs for both S and R depend in general on the state t, the sent message m and the action a chosen by the receiver. Formally, a *signaling game* is a tuple $\langle \{S, R\}, T, \Pr, M, A, \mathrm{U}_S, \mathrm{U}_R \rangle$ where $\Pr \in \Delta(T)$ is a probability distribution over T capturing the receiver's *prior beliefs* about which state is actual, and $\mathrm{U}_{S,R} : M \times A \times T \to \mathbb{R}$ are utility functions for both sender and receiver. We speak of a *cheap-talk game*, if message use does not influence utilities.[1]

It is clear to see that a signaling game embeds a classical decision problem, such as discussed in the previous section. The receiver is the DM and the sender is the manipulator. It is these structures that help us to study manipulation strategies and assess their

[1] For simplicity we assume that (i) T, M and A are finite non-empty sets, and (ii) $\Pr(t) > 0$ for all $t \in T$.

success probabilities.

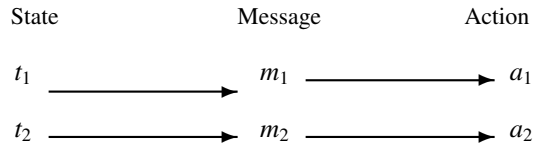
To specify player behavior, we define the notion of a *strategy*, which is now a formal object, unlike our more general use of the term, as discussed in Section 1.The sender can send a message in each possible state, and so a sender strategy, $\sigma \in M^T$ is modelled as a function from states to messages. Likewise, a receiver strategy $\rho \in A^M$ is a function from messages to actions. As usual, we say that the strategy pair $\langle \sigma^*, \rho^* \rangle$ is an equilibrium if neither player can do any better by unilateral deviation. More technically, $\langle \sigma^*, \rho^* \rangle$ is a Nash equilibrium iff for all $t \in T$:

(i) $\mathrm{U}_S(t, \sigma^*(t), \rho^*(\sigma^*(t))) \geq \mathrm{U}_S(t, \sigma(t), \rho^*(\sigma(t)))$ for all $\sigma \in M^T$, and

(ii) $\mathrm{U}_R(t, \sigma^*(t), \rho^*(\sigma^*(t))) \geq \mathrm{U}_R(t, \sigma^*(t), \rho(\sigma^*(t)))$ for all $\rho \in A^M$.

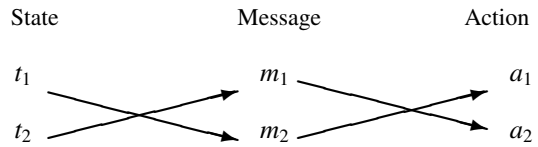
A signaling game typically has many equilibria. Suppose we limit ourselves to a cooperative signaling game with only two states $T = \{t_1, t_2\}$ that are equally probable $\Pr(t_1) = \Pr(t_2)$, two messages $M = \{m_1, m_2\}$, and two actions $A = \{a_1, a_2\}$, and where $\mathrm{U}(t_i, m_j, a_k) = 1$ if, $i = k$, and 0 otherwise, for both sender and receiver. In that case the following combination of strategies is obviously a Nash equilibrium:[2]

(1)



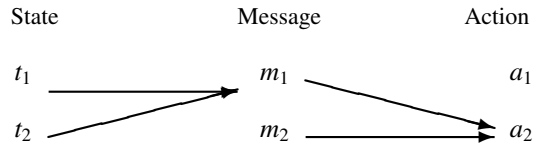
But because the messages don't have any pre-existing meaning, the following combination of strategies is an equally good equilibrium:

(2)



In both situations, the equilibria make real communication possible. Unfortunately, there are also Nash equilbria where nothing is communicated about the actual state of affairs. In case the sender's prior probability of t_2 exceeds that of t_1, for instance, the following combination is also a Nash equilibrium:

(3)



[2] A sender strategy is a function from states to messages, is here pictured by left-most arrows in the diagram. Similarly, for the receiver's strategy with the arrows on the right-hand side, leading from messages to actions.

Until now we assumed that messages don't have an *a priori* given meaning. What happens if we give up this assumption, and assume that a conventional language is already in place that can be used or abused by speakers to influence their hearers for better or worse? Formally, we model this by a semantic denotion function $[\![\cdot]\!] : M \rightarrow \mathcal{P}(T)$ such that $t \in [\![m]\!]$ iff m is true in t.[3]

Assuming that messages have a conventional meaning can help filter out unreasonable equilibria. In seminal early work, Farrell (1993) (the paper goes back to at least 1984) proposed to refine the equilibrium set for cheap-talk signaling games by a notion of message credibility, requiring that R believe what S says if it is in S's interest to speak the truth (c.f. Farrell and Rabin, 1996). Farrell's solution is rather technical and can be criticized for being impractically unrealistic, but his general idea has been picked up and refined in many subsequent contributions, as we will also see below (c.f. Myerson, 1989; Rabin, 1990; Matthews et al., 1991; Zapater, 1997; Stalnaker, 2006; Franke, 2010). Essentially, Farrell assumed that the set of available messages is infinite and expressively rich: for any given reference equilibrium and every subset $X \subseteq T$ of states, there is always a message m_X with $[\![m_X]\!] = X$ that is not used in that equilibrium.[4] Such an unused message m is called a credible neologism iff all and only the types in $[\![m]\!]$ prefer the neologism to candidate equilibrium payoffs when R responds optimally to the literal meaning of the new signal (that is $[\![m]\!]$). Formally, take an equilibrium $\langle \sigma^*, \rho^* \rangle$, and let $U_S^*(t)$ be the equilibrium payoff of type t for the sender. The types in $[\![m]\!]$ can send a *credible neologism* iff $[\![m]\!] = \{t \in T : U_S(t, BR([\![m]\!])) > U_S^*(t)\}$, where $BR([\![m]\!])$ is R's (assumed unique, for simplicity) optimal response to the prior distribution conditioned on $[\![m]\!]$. If R interprets a credible neologism literally, then some types would send the neologism and destroy the candidate equilibrium. A neologism proof equilibrium is a sequential equilibrium for which no subset of T can send a credible neologism. For example, the previous two fully revealing equilibria in (1) and (2) are neologism proof, but the pooling equilibrium in (3) is not: there is a message m^* with $[\![m^*]\!] = \{t_2\}$ which exactly the state in which m^* is true would prefer to send over the given pooling equilibrium.

Farrell defined his notion of credibility in terms of a given reference equilibrium. Yet for accounts of online pragmatic reasoning about language use, it is not always clear where such an equilibrium should come from. But in that case another reference point for pragmatic reasoning is ready-at-hand, and is a situation *without* communication entirely. So another way of thinking about $U_S^*(t)$ is just as the utility of S in t if R plays the action with the highest expected utility of R's decision problem, i.e., the utility that would ensue. In this spirit, van Rooy (2003) determines the relevance of information against the background of the DM's decision problem. Roughly speaking, it is claimed that message m is relevant w.r.t. a decision problem, if the hearer will change his action upon hearing it.[5] A message, then, is credible in case it is relevant,

[3] We assume for simplicity that for each state t there is at least one message m which is true in that state; and that no message is contradictory, i.e., there is no m for which $[\![m]\!] = \emptyset$.

[4] This *rich language assumption* might be motivated by evolutionary considerations, but is unsuitable for applications to online pragmatic reasoning about natural language, which, arguably, is not at the same time cheap and fully expressive: some things are more cumbersome to express than others, if at all (c.f. Franke, 2010).

[5] Benz (2007) criticizes this and other decision-theoretic approaches, arguing for the need to take the

and cannot be used misleadingly. Look at the following cooperative situation:

(4)

$U(t_i, a_j)$	a_1	a_2
t_1	1,1	0,0
t_2	0,0	1,1

If this were just a decision problem without possibility of communication and furthermore $Pr(t_2) > Pr(t_1)$, then R will play a_2. But that would mean that $U^*_S(t_1) = 0$, while $U^*_S(t_2) = 1$. In this scenario, message "I am of type t_1" is credible, given the revised notion, but "I am of type t_2" is not, because it is not relevant. Notice that if a speaker is of type t_2, he wouldn't say anything, but the fact that the speaker didn't say anything, if taken into account, must be interpreted as S being of type t_2 (because otherwise S would have said 'I am t_1'.) Assuming that saying nothing is saying the trivial proposition, R can conclude something more from some "messages" than is literally expressed. This is not unlike conversational implicatures (Grice, 1975).

So far we have seen that if preferences are aligned, a notion of credibility helps predict successful communication in a natural way. What about circumstances where this ideal condition is not satisfied? Look at the following table:

(5)

$U(t_i, a_j)$	a_1	a_2
t_1	1,1	0,0
t_2	1,0	0,1

In this case, both types of S want R to play a_1 and R would do so, in case he believed that S is of type t_1. However, R will not believe S's message "I am of type t_1", because if S is of type t_2 she still wants R to believe that she is of type t_1, and thus wants to mislead the receiver. Credible communication is not possible now. More in general, it can be shown that costless messages with a pre-existing meaning can be used to credibly transmit information only if it is known by the receiver that it is in the sender's interest to speak the truth.[6] If communicative manipulation is predicted to be possible at all, its succesful use is predicted to be highly restricted.

We also must acknowledge that a proper notion of messages credibility is more complicated that indicated so far. Essentially, Farrell's notion and the slight amendment we introduced above use a *forward induction* argument to show that agents can talk themselves out of an equilibrium.[a] But it seems we didn't go far enough. To show this, consider the following game where states are again assumed equiprobable:

a. perhaps reference to other papers in this volume?

(6)

$U(t_i, a_j)$	a_1	a_2	a_3	a_4
t_1	10,5	0,0	1,4.1	-1,3
t_2	0,0	10,5	1,4.1	-1,3
t_3	0,0	0,0	1,4.1	-1,6

speaker's perspective into account as is done in Benz (2006); Benz and van Rooij (2007).

[6]The most relevant game-theoretical contributions are by Farrell (1988, 1993), Rabin (1990), Matthews et al. (1991), Zapater (1997). More recently, this topic has been reconsidered from a more linguistic point of view, e.g., by Stalnaker (2006), Franke (2010) and Franke et al. (2012).

Let's suppose again that we start with a situation with only the decision problem and no communication. In this case, R responds with a_3. According to Farrell, this gives rise to two credible announcements: "I am of type t_1" and "I am of type t_2", with the obvious best responses. This is because both types t_1 and t_2 can profit from having these true messages believed: a credulous receiver will answer with actions a_1 and a_2 respectively. A speaker of type t_3 cannot make a credible statement, because revealing her identity would only lead to a payoff strictly worse than what she obtains if R plays a_3. Consequently, R should respond to no message with the same action as he did before, i.e., a_3. But once R realizes that S could have made the other statements credibly, but didn't, she will realize that the speaker must have been of type t_3 and will respond with a_4, and not with a_3. What this shows is that to account for the credibility of a message, one needs to think of higher levels of strategic sophistication of other plays: common knowledge of rationality is assumed. What this also suggests is that if either R or S do not believe in common believe in rationality, then misleading communication might again be possible. This is indeed what we will come back to presently in Section 2.3.

But before turning to that, we should address one more general case. Suppose we assume that messages not only have a semantic meaning, but that speakers also obey Grice's Maxim of Quality and do not assert falsehoods (Grice, 1975).[7] Do we predict more communication now? Milgrom and Roberts (1986) demonstrate that in such cases it is best for the DM to "assume the worst" about what S reports and that S has omitted information that would be useful. Milgrom and Roberts show that the optimal equilibrium strategy will always be the *sceptical posture*. What is more is that, in this situation, S will know that, unless he is told everything, the DM will take a stance against both the DM's own interests (had he had full information) and the interests of S. Given this, the S could as well reveal all she knows.[8] In terms of our topic, this means that when speakers might try to manipulate the beliefs of the DM by being less precise than they could be, this won't help because the DM will see through this attempt of manipulation. So, again, the conclusion is that standard economic theory predicts that manipulation by communication is impossible, a result that is very much in conflict with what we perceive daily.[9] Thus, we cannot explain the fact that in many situations we communicate even though the circumstances are not favorable. We cannot explain manipulation.

[7] It is very frequently assumed in game theoretic models of pragmatic reasoning that the sender is compelled to truthful signaling by the game model. This assumption is present, for instance, in the work of (Parikh, 1991, 2001, 2010), but also assumed by many others. As long as interlocutors are cooperative in the Gricean sense, this assumption might be innocuous enough, but, as the present considerations make clear, are too crude a simplification when we allow conflicts of interest.

[8] The argument used to prove the result is normally called the *unraveling argument*. See Franke et al. (2012) for a slightly different version.

[9] Shin (1994) proves a generalization of Milgrom and Roberts's (1986) result, claiming that there always exists a sequential equilibrium (a strengthened notion of Nash equilibrium we have not introduced here) of the persuasion game in which the sender's strategy is perfectly revealing in the sense that the sender will say exactly what he knows.

2.3 Giving up some standard assumptions

The currently most popular and successful theories of meaning and (communicative) behaviour are based on theories of ideal reasoning and rational behaviour like logic, and decision and game theory. These theories assume that we are rational beings, which is an excellent first approximation: (i) Intuitively, because when people's deviations from logical and decision theoretic predictions are pointed out, they generally agree that they have erred, (ii) empirically, because decision theory, for instance, has recently gained a lot of supporting evidence from psychology (c.f. Newel et al., 2007) and neuroscience (c.f. Glimcher and Rustichini, 2004; Glimcher et al., 2009), and (iii) conceptually, because game theory provides the conceptual and procedural tools for studying various types of social interactions, and in restricted environments it correctly predicts our behaviour (Davis and Holt, 1993).

Still, there also exists a lot of theoretical and experimental evidence that language users are not perfectly rational. Theoretical computer scientists and experimental psychologists have discovered well-known limitations of our rational behavior. Assuming Church's thesis, the theory of computation describes limitations on what we can do. If we must make decisions in real time, the theory predicts, for instance, that we cannot do so consistently. Indeed, it can hardly be doubted that we sometimes hold inconsistent beliefs, and that our decisionmaking exhibits systematic biases that are unexplained by the standard theory (Simon, 1959; Tversky and Kahnemann, 1974). Furthermore, there exists a lot of empirical evidence that standard game theory is based on some unrealistic assumptions. In this section we will discuss two of such assumptions, and indicate what might result if we give these up. First we will discuss the assumption that the game being played is common knowledge. Then we will investigate the implications of giving up the hypothesis that everybody is completely rational, and that this is common knowledge. Finally, we will discuss what happens if our choices are systematically biased. In all three cases, we will see more room for successful manipulation.

No common knowledge of game being played. In standard game theory it is usually assumed that players conceptualize the game in the same way, i.e., that it is common knowledge what game is played. But this seems like a highly idealized assumption. It is certainly the case that interlocutors occasionally operate under quite different conceptions of the context of conversation, i.e., the game they are playing. This is evidenced by misunderstandings, but also by the way we talk: cooperative speakers must not only provide information but also enough background to make clear how that information is relevant. To cater for these aspects of conversation, Franke (submitted) uses models for games with unawareness (c.f. Halpern and Rêgo, 2006; Feinberg, 2011a; Heifetz et al., 2011) to give a general model for pragmatic reasoning in situations where interlocutors may have variously diverging conceptualizations of the context of utterance relevant to the interpretation of an utterance, different beliefs about these conceptualizations, different beliefs about these beliefs and so on.[b] However, Franke (submitted) only discusses examples where interlocutors are well-behaved Gricean cooperators (Grice, 1975) with perfectly aligned interests. Looking at cases where this is not so, Feinberg (2008, 2011b) demonstrates that taking unawareness into account also provides a new rationale for communication in case of conflicting interests. Feinberg gives

b. include reference to DEL chapter and/or epistemic game theory chapter?

examples where communicating one's awareness of the set of actions which the DM can choose from might be beneficial for both parties involved. But many other examples exist (e.g. Ozbay, 2007). Here is a very simple one that nonetheless demonstrates the relevant conceptual points.

Reconsider the basic case in (5) that we looked at previously. We have two types of senders: t_1 wants his type to be revealed, and t_2 wishes to be mistaken for a type t_1. As we saw above, the message "I am of type t_1" is not credible in this case, because a sender of type t_2 would send it too. Hence, a rational DM should not believe that the actual type is t_1 when he hears that message. But if the DM is not aware that there might a type t_2 that might want to mislead him, then, although incredible from the point of view of a perfectly aware spectator, from the DM's *subjective point of view*, the message is perfectly credible. The example is (almost) entirely trivial, but the essential point nonetheless significant. If we want to mislead, but also if we wants to reliably and honestly communicate, it might be the very best thing to do to leave the DM completely in the dark as to any mischievous motivation we might pursue or, contrary to fact, might have been pursuing:

> Once, when she was littler, Sansa had even asked Mother if perhaps there hadn't been some mistake. Perhaps the grumkins had stolen her *real* sister. But Mother had only laughed and said no, Arya was her daughter and Sansa's trueborn sister, blood of their blood. Sansa could not think why Mother would want to lie about it, so she supposed it had to be true.
>
> (Martin, 1996)

This simple example also shows the importance of choosing, not only *what* to say, but also *how* to say it. (We will come back to this issue in more depth below when we look at *framing effects*.) In the context of only two possible states, the messages "I am of type t_1" and "I am not of type t_2" are equivalent. But, of course, from a persuasion perspective they are not equally good choices. The latter would make the DM aware of the type t_2, the former would not. So although contextually equivalent in terms of their extension, the requirements of efficient manipulation clearly favor the one over the other simply in terms of surface form, due to their variable effects on the awareness of the DM.

In a similar spirit, van Rooij and Franke (to appear) use differences in awareness-raising of otherwise equivalent conditionals and disjunctions to explain why there are conditional threats (7a) and promises (7b), and also disjunctive threats (7c), but, what is surprising from a logical point of view, no disjunctive promises (7d).

(7) a. If you don't give me your wallet, I'll punish you severely. threat
b. If you give me your wallet, I'll reward you splendidly. promise
c. You will give me your wallet or I'll punish you severely. threat
d. ? You will not give me your wallet or I'll reward you splendidly. threat

Sentence (7d) is most naturally read as a threat by accommodating the admittedly aberrant idea that the hearer has a strong aversion against a splendid reward. If that much accommodation is impossible, the sentence is simply pragmatically odd. The general

absence of disjunctive threats like (7d) from natural language can be explained, van Rooij and Franke argue, by noting that these are suboptimal manipulation strategies because, among other things, they raise the possibility that the speaker does *not* want the hearer to perform.[10]

These are just a few basic examples that show how reasoning about the possibility of subjective misconceptions of the context/game model affects what counts as an optimal manipulative technique. But limited awareness of the context model is not the only cognitive limitation that real life manipulators may wish to take into consideration. Limited reasoning capacity is another.

No common knowledge of rationality. A number of games can be solved by (iterated) elimination of dominated strategies. If we end up with exactly one (rationalizable) strategy for each player, this strategy combination must be a Nash equilibrium. Even though this procedure seems very appealing, it crucially depends on a very strong epistemic assumption: *common knowledge of rationality*; not only must every agent be ideally rational, everybody must also know of each other that they are rational, and they must know that they know it, and so on *ad infinitum*.[c] It is even harder to justify Nash equilibria, but also this leans heavily on this strong assumption. Unfortunately, there exists a large body of empirical evidence that the assumption of common knowledge of rationality is highly unrealistic (c.f. Camerer, 2003). Is it possible to explain deception and manipulation if we give up this assumption?

c. provide reference to other papers in this volume

Indeed, it can be argued that whenever we do see attempted deceit in real life we are sure to find at least a belief of the deceiver (whether justified or not) that the agent to be deceived has some sort of limited reasoning power that makes the deception at least conceivably successful. Some agents are more sophisticated than others, and think further ahead. To model this, one can distinguish different *strategic types* of players, often also referred to as cognitive hierarchy models within the economics literature (e.g. Camerer et al., 2004; Rogers et al., 2009) or as iterated best response models in game theoretic pragmatics (e.g. Jäger, 2008; Jäger and Ebert, 2009; Franke, 2011). A strategic type captures the level of strategic sophistication of a player and corresponds to the number of steps that the agent will compute in a sequence of iterated best responses. One can start with an unstrategic level-0 players. An unstrategic level-0 hearer (a credulous hearer), for example, takes the semantic content of the message he receives literally, and doesn't think about why a speaker used this message. Obviously, such a level-0 receiver can sometimes be manipulated by a level-1 sender. But such a sender can in turn be seen through by a level-2 receiver, etc. In general, a level-$(k+1)$ player is one who plays a best response to the behavior of a level-k player. (A *best response* is a rationally best reaction to a given belief about the behavior of all other players.) A fully sophisticated agent is a level-ω player who behaves rationally given her belief in common belief in rationality.

Using such cognitive hierarchy models, Crawford (2003), for instance, showed that in case sender and/or receiver believe that there is a possibility that the other player is

[10] Although conditional threats also might make the DM aware of the "wrong" option, these can still be efficient inducements because, according to van Rooij and Franke (to appear) the speaker can safely increase the stakes, by committing to more severe levels of punishment. If the speaker would do that for disjunctive promises, she would basically harm herself by expensive promises.

less sophisticated than he is himself, deception is possible (c.f. Crawford, 2007). Moreover, even sophisticated level-ω players can be deceived if they are not sure that their opponents are level-ω players too. Crawford assumed that messages have a specific semantic content, but did not presuppose that speakers can only say something that is true.

Building on work of Rabin (1990) and Stalnaker (2006), Franke (2010) offers a notion of *message credibility* in terms of an iterated best response model (see also Franke, 2009, Chapter 2). The general idea is that the conventional meaning of a message is a strategically non-binding *focal point* that defines the behavior of unstrategic level-0 players. For instance, for the simple game in (5), a level-0 receiver would be credulous and believe that message "I am of type t_2" is true and honest. But then a level-1 sender of type t_2 would exploit this naïve belief and also believe that her deceit is successful. Only if the receiver in fact is more sophisticated than that, would he see through the deception. Roughly speaking, a message is then considered credible iff no strategic sender type would ever like to use it falsely. In effect, this model not only provably improves on the notion of message credibility, but also explains when deceit can be (believed to be) successful.

We can conclude that (i) it is unnatural to assume common knowledge of rationality, and (ii) by giving up this assumption, we can explain much better why people communicate than standard game theory can: sometimes we communicate to manipulate others on the assumption that the others don't see it through, i.e., that we are smarter than them (whether this is justified or not).

Bounded rationality. As noted earlier, there exist a lot of experimental and theoretical evidence that we do not, and even cannot, always pick our choices in the way we should do according to the standard normative theory. Especially due to the work of Kahneman and Tversky, it is now widely accepted among psychologists that the idea that we choose by maximizing expected utility is inaccurate. Structured after the well-known Allais paradox, their famous Asian disease experiment (Tversky and Kahnemann, 1981), for instance, shows that in most people's eyes, a *certain* gain is worth more than an equally large *expected* gain. This phenomenon is known as the *certainty effect*. The fact that people buy lottery tickets and pay quite some money to insure themselves against very unlikely losses suggest that people sometimes violate the prescriptions of normative decision theory because they systematically overweight low-probability events. Experimental findings also suggest that DMs think in terms of gains and losses with respect to a reference point, rather than in terms of context-independent utilities as the standard theory assumes. This is important because people tend to be risk-averse in the domain of gains, and risk-taking in the domain of losses. This can explain why parents tend to be more persuaded to, for instance, vaccinate their children by loss-framed than by gain-framed appeals (O'Keefe and Jensen, 2007). Related to this is the so-called *status quo bias*: a preference for the status quo. This bias can be rational, due to informational or cognitive limitations. Because the outcome or utility of a decision might be very uncertain, it is in many cases a good idea to stick to a choice that worked good enough in the past. To simply maintain a current course of action also requires less mental effort, and might thus be rational. Still, evidence shows

that it really counts as a bias, because it is often irrational.

Other experiments show that not only the expected utility principle but also other fundamental principles of rational choice are frequently violated. It is standardly assumed in decision theory that preference orders are transitive and complete. Still, already May (1945) has shown that cyclic preferences were not extraordinary (violating transitivity), and Luce (1959) noted that people sometimes seem to choose one alternative over another with a certain probability (violating completeness).

What is interesting for us is that due to the fact that people don't behave as rationally as the standard normative theory prescribes, it becomes possible for smart communicators to *manipulate* them: to convince them to do something that goes against their own interest. We mentioned already the use of *false implication*. Perhaps better known is the *money pump* argument: the fact that agents with intransitive preferences can be exploited because they are willing to participate in a series of bets where they will loose for sure. Similarly, manipulators make use of *false analogies*. According to psychologists, reasoning by analogy is used by boundedly rational agents like us to reduce the evaluation of new situations by comparing them with familiar ones. Though normally a useful strategy, it can be exploited. There are many examples of this. Just to take one, in an advertisement for Chanel No. 5, a bottle of the perfume is pictured together with Nicole Kidman. The idea is that Kidman's glamour and beauty is transferred from her to the product. But perhaps the most common way to influence a DM making use of the fact that he or she does not choose in the prescribed way is by *framing*.

By necessity, a DM interprets her decision problem in a particular way. A different interpretation of the same problem may sometimes lead to a different decision. Indeed, there exists a lot of experimental evidence, that our decision making can depend a lot on how the problem is set. In standard decision theory it is assumed that decisions are made on the basis of information, and that it doesn't matter how this information is presented. It is predicted, for instance, that it doesn't matter whether you present this glass as being half full, or as half empty. The fact that it sometimes does matter is called the *framing effect*. The idea is that manipulating the way information is presented can influence decision making. Through the use of our words, and by presenting a general context around the information presented we can influence how people think about that information. For instance, it is assumed in the normative theory of decision that the relative ranking of any two actions should not vary with the addition or deletion of other irrelevant actions. The addition of a new act, which is not regarded as better than the oringinal ones, should not change a rational agent's ranking of the old actions (the *irrelevant expansion condition*). Despite its intuitive appeal, there is evidence that DMs do not always satisfy this context-independence condition. So-called *compromise effects* and *contrast-effects* are two types of violations of context-independence for which experimental evidence exist. The first is the fact that the same action is evaluated more positively when framed as intermediate in the set of actions under valuation than when it is extreme. This has the effect that an agent's choice can be manipulated by the addition or deletion of 'irrelevant' alternatives. There even exists a decision rule that violates the irrelevant expansion condition —the minimax rule— that even among decision theorists has some popularity when choices have to be made under ignorance.

Context-independence is also violated by *contrast-effects*: the fact that the same option is evaluated more favorably in the presence of similar options clearly inferior to

it than in the absence of such options. We are all familiar with this effect. Just as the same circle seems larger when it is surrounded by small circles and smaller when it is surrounded by larger ones, so it is also the case that the same product may seem more attractive in the context of less attractive alternatives, and less attractive in the context of more attractive ones. If you want to sell a product, compare it with a very similar product that is less attractive.

A frame is a reference point for all future decision making. Frames set expectations, both of oneself and of others, which can influence behavior. We are all familiar with the placebo-effect: patients experience treatment benefits based on the belief that the treatment will work. Teachers are perhaps not aware enough of the very similar effect that they tend to treat students differently based on expectations of how they will perform. And the fact that people have the tendency to behave as they believe is expected, is a well-known problem if one tries to draw objective conclusions from experiments and interviews. But at the same time as these phenomena pose problems for researchers, they offer opportunities to manipulative communicators. Using it for a good cause, doctors are advised to speak very favorably about the effects of the placebo-pill to their patients, because it will set the patient's expectations higher.

Our choice behavior is obviously effected by our limited abilities to process information. For instance, if we consider a set of alternative actions, we are normally not able to oversee all the actions at the same time. Instead, we sometimes look at them one by one, as in a list. Also this can give rise to at least two possible sources of framing effects. Firstly, the choice may depend on the order in which the alternative actions are listed. It might be that an agent tends to choose elements in the beginning of the list, a *primacy effect*, and this tendency might be made use of by a manipulating communicator. Secondly, an agent's choice may depend on the number of times a given element appears in the list, such as the tendency to favor elements that appear multiple times in the list. Indeed, our likeability of things tends to increase after repeated exposure. The *exposure effect* is one of the basic concepts behind advertising. If you see an advertisement in a magazine or a commercial on TV over and over again (though not too many times), after a time you become more likely to buy the product being advertised.

Framing effects are predicted by Kahneman and Tversky's *Prospect Theory*: a theory that implements the idea that our behavior is only boundedly rational. But if correct, it is this kind of theory that should be taken into account in any serious analysis of persuasive language use. In complex choice situations where the maximization of utility is hard to implement individuals frequently use heuristics, as Tversky and Kahnemann (1974) called them. Heuristics can be thought of as reasoning strategies allowing to reduce the complexity of a decision task. The *elimination by aspect* (EBA) procedure of Tversky is one of these heuristics. Tversky (1972) thinks of selection in terms of elimination. We first randomly select a characteristic that is relevant for the decision, and eliminate all the options not having this characteristic. The higher the utility of a characteristic is, the larger the probability of selecting this characteristic is. One repeats this procedure until one reaches a decision. This kind of decision making is natural, and less time-consuming than doing so by maximizing utility. But the EBA choice procedure sometimes results in choices that conflict with the normative theory. For one, because it leads to so-called probabilistic choices and thereby violating the completeness assumption. For another, because in some cases it gives rise to subop-

timal decisions in the majority of cases. Consider, for instance the following decision problem, where the agent has to make a choice between actions a and c, and where A, B and C are the relevant characteristics:

(8)

U(x, X)	A	B	C
a	2	2	2
b	0	5	0

According to standard theory, one should choose item a because it has the highest utility. In contrast, EBA selects b more than half of the time.

Also, EBA is not unlike some natural voting procedures in social choice theory. And just as it is well-known that social choices can be influenced in the latter case by setting the agenda (c.f. Pauly, 2005, on the German capital debate), it is equally true that someone who decides by EBA can be manipulated by presenting the relevant aspects in a smart manner.

Summary. So, why do we talk so much? Perhaps because our preferences are much aligned and participants of a conversation all profit from a larger distribution of knowledge. This would be the ideal picture, but we doubt it is the true reason behind our talking. We also talk if our preferences are not aligned. No, we talk so much, we argue, because, among others, (i) we think know better in which situation we are than others; (ii) we think we are smarter than others, or (iii) we think we can influence the probabilities and utilities of others by the way we frame their decision problems. In short, we talk and argue so much because we are and believe others to be *boundedly rational agents*.

3 Opinion dynamics & efficient propaganda

While the previous section focused exclusively at the pragmatic dimension of persuasion, investigating *what* to say and *how* to say it, there is a wider social dimension to successful manipulation as well: determining *who we should address*. In this section, we will assume that agents are all part of a social network, and we will discuss how to best propagate one's own ideas through a social network.

The model contributed by this section makes use of ideas from sociology and rational choice theory. From sociology and evolutionary game theory we will use the insight that the spread of opinion is partly determined by the relationships and interactions of the individuals within a group: the social network. The decision who to address is clearly a strategic decision and belongs to rational choice theory. The innovative part of our paper —or at least, as far as we know— is the combination of both of these perspectives.

In particular, this section presents a variant of DeGroot's classical model of opinion dynamics (DeGroot, 1974) that allows us to address the question how an agent, given his position in a social web of influenceability, should try to strategically influence others, so as to maximally promote her opinion in the relevant population. More

concretely, while DeGroot's model implicitly assumes that agents distribute their persuasion efforts equally among the neighbors in their social network, we consider a (to the best of our knowledge: new) variant where a small fraction of players is able (willing/smart enough/ ...) to re-distribute their persuasion efforts strategically. Using numerical simulations, we try to charter the terrain of more or less efficient opinion-promoting strategies and conclude that in order to successfully promote your opinion in your social network you should: (i) spread your web of influence wide (i.e., not focussing all effort on a single or few individuals), (ii) choose "easy targets" for quick success and "influential targets" for long-term success, and (iii), if possible, coordinate your efforts with other influencers so as to get out of each other's way. Which strategy works best, however, depends on the interaction structure of the population in question. The upshot of this discussion is that, even if computing the theoretically optimal strategy is out of the question for a resource-limited agent, the more an agent can exploit rudimentary or even detailed knowledge of the social structure of a population, the better she will be able to propagate her opinion.

Starting Point: The DeGroot Model. DeGroot (1974) introduced a simple model of opinion dynamics to study under which conditions a consensus can be reached among all members of the society (cf. Lehrer, 1975). DeGroot's classical model is a round-based, discrete and linear update model.[11] Opinions are considered at discrete time steps $t \in \mathbb{N}^{\geq 0}$. In the simplest case, an opinion is just a real number, representing, e.g., to what extent an agent endorses a position. For n agents in the society we consider the row vector of opinions $\mathbf{x}(t)$ with $\mathbf{x}(t)^{-1} = \langle x_1(t), \ldots, x_n(t) \rangle \in \mathbb{R}^n$ where $x_i(t)$ is the opinion of agent i at time t.[12] Each round all agents update their opinions to a weighted average of the opinions around them. Who influences whom how much is captured by *influence matrix* P, which is a (row) stochastic $n \times n$ matrix with p_{ij} the weight with which agent i takes agent j's opinion into account. DeGroot's model then considers the simple linear update in (1):

$$\mathbf{x}(t+1) = P\,\mathbf{x}(t)\,. \tag{1}$$

For illustration, suppose that the society consists of just three agents and that influences among these are given by:

$$P = \begin{pmatrix} .7 & .3 & 0 \\ .2 & .5 & .3 \\ .4 & .5 & .1 \end{pmatrix}. \tag{2}$$

The rows in this influence matrix give the proportions with which each agent updates her opinions at each time step. For instance, agent 3's opinion at time $t+1$ is obtained by taken .4 parts of agent 1's opinion at time t, .5 parts of agent 2's and .1 parts of her own opinion at time t. For instance, if the vector of opinions at time $t = 0$ is a randomly chosen $\mathbf{x}(0)^{-1} = \langle .6, .2, .9 \rangle$, then agent 3's opinion at the next time step will

[11] DeGroot's model can be considered as a simple case of Axelrod's (1997) famous model of cultural dynamics (c.f. Castellano et al., 2009, for overview)

[12] We write $\mathbf{x}(t)^{-1}$ for the transpose of the row vector $\mathbf{x}(t)$, so as not to have to write its elements vertically.

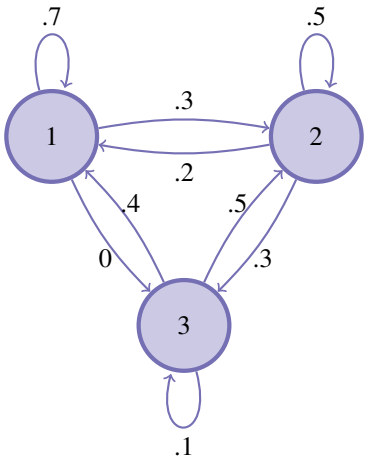


Figure 1: Influence in a society represented as a (fully connected, weighted and directed) graph.

be $.4 \times .6 + .5 \times .2 + .1 \times .9 \approx .43$. By equation (1), we compute these updates in parallel for each agent, so we obtain $\mathbf{x}(1)^{-1} \approx \langle .48, .49, .43 \rangle$, $\mathbf{x}(2)^{-1} \approx \langle .48, .47, .48 \rangle$ and so on.[13]

DeGroot's model acknowledges the social structure of the society of agents in its specification of the influence matrix P. For instance, if $p_{ij} = 0$, then agent i does not take agent j's opinion into account at all; if $p_{ii} = 1$, then agent i does not take anyone else's opinion into account; if $p_{ij} < p_{ik}$, then agent k has more influence on the opinion of agent i than agent j.

It is convenient to think of P as the adjacency matrix of a fully-connected, weighted and directed graph, as shown in Figure 1. As usual, rows specify the weights of outgoing connections, so that we need to think of a weighted edge in a graph like in Figure 1 as a specification of how much an agent (represented by a node) "cares about" or "listens to" another agent's opinion. The agents who agent i listens to, in this sense, are the influences of i:

$$I(i) = \left\{ j \mid p_{ij} > 0 \wedge i \neq j \right\} .$$

Inversely, let's call all those agents that listen to agent i as the audience of i:

$$A(i) = \left\{ j \mid p_{ji} > 0 \wedge i \neq j \right\} .$$

One more notion that will be important later should be mentioned here already. Some agents might listen more to themselves than others. Since how much agent i

[13] In this particular case, opinions converge to a consensus where everybody holds the same opinion. In his original paper DeGroot showed that, no matter what $\mathbf{x}(0)$, if P has at least one column with only positive values, then, as t goes to infinity, $\mathbf{x}(t)$ converges to a unique vector of uniform opinions, i.e., the same value for all $\mathbf{x}_i(t)$. Much subsequent research has been dedicated to finding sufficient (and necessary) conditions for opinions to converge or even to converge to a consensus (c.f. Jackson, 2008; Acemoglu and Ozdaglar, 2011, for overview). Our emphasis, however, will be different, so that we sidestep these issues.

holds on to her own opinion at each time step is given by value p_{ii}, the diagonal $\texttt{diag}(P)$ of P can be interpreted as the vector of the agents' stubbornness. For instance, in example (2) agent 1 is the most stubborn and agent 3 the least convinced of his own views, so to speak.

Strategic Promotion of Opinions. DeGroot's model is a very simple model of how opinions might spread in a society: each round each agent simply adopts the weighted average of the opinions of his influences, where the weights are given by the fixed influence matrix. More general update rules than (1) have been studied, e.g., ones that make the influence matrix dependent on time and/or the opinions held by other agents, so that we would define $\mathbf{x}(t+1) = P(t, \mathbf{x}(t))\,\mathbf{x}(t)$ (cf. Hegselmann and Krause, 2002). We are interested here in an even more liberal variation of DeGroot's model in which (some of the) agents can *strategically* determine their influence, so as to best promote their own opinion. In other terms, we are interested in opinion dynamics of the form:

$$\mathbf{x}(t+1) = P(S)\,\mathbf{x}(t)\,, \tag{3}$$

where P depends on an $n \times n$ strategy matrix S where each row S_i is a strategy of agent i and each entry S_{ij} specifies how much effort agent i invests in trying to inpose her current opinion on each agent j.

Eventually we are interested in the question when S_i is a *good* strategy S_i for a given influence matrix P, given that agent i wants to promote her opinion as much as possible in the society. But to formulate and address this question more precisely, we first must define (i) what kind of object a strategy is in this setting and (ii) how exactly the actual influence matrix $P(S)$ is computed from a given strategy S and a given influence matrix P.

Strategies. We will be rather liberal as to how agents can form their strategies: S could itself depend on time, the current opinions of others etc. We will, however, impose two general constraints on S because we want to think of strategies as allocations of persuasion effort. The first constraint is a mere technicality, requiring that $S_{ii} = 0$ for all i: agents do not invest effort into manipulating themselves. The second constraint is that each row vector S_i is a stochastic vector, i.e., $S_{ij} \geq 0$ for all i and j and $\sum_{j=1}^{n} S_{ij} = 1$ for all i. This is to make sure that strengthening one's influence on some agents comes at the expense of weakening one's influence on others. Otherwise there would be no interesting strategic considerations as to where best to exert influence. We say that S_i is a neutral strategy for P if it places equal weight on all j that i can influence, i.e., all $j \in A(i)$. We call S neutral for P, if S consists entirely of neutral strategies for P. We write S^* for the neutral strategy of an implicitly given P.

Examples of strategy matrices for the influence matrix in (2) are:

$$S = \begin{pmatrix} 0 & .9 & .1 \\ .4 & 0 & .6 \\ .5 & .5 & 0 \end{pmatrix} \qquad S' = \begin{pmatrix} 0 & .1 & .9 \\ .5 & 0 & .5 \\ 0 & 1 & 0 \end{pmatrix} \qquad S^* = \begin{pmatrix} 0 & .5 & .5 \\ .5 & 0 & .5 \\ 0 & 1 & 0 \end{pmatrix}.$$

According to strategy matrix S, agent 1 places .9 parts of her available persuasion effort on agent 2, and .1 on agent 3. Notice that since $P_{13} = 0$, agent 3 cannot influence

agent 1. Still, nothing prevents her from allocating persuasion effort to agent 1. (This would, in a sense, be irrational but technically possibile.) That also means that S_3 is *not* the neutral strategy for agent 3. The neutral strategy for agent 3 is S'_3 where all effort is allocated to the single member in agent 3's audience, namely agent 2. Matrix S' also includes the neutral strategy for agent 2, who has two members in her audience. However, since agent 1 does not play a neutral strategy in S', S' is not neutral for P, but S^* is.

Actual Influence. Intuitively speaking, we want the actual influence matrix $P(S)$ to be derived by adjusting the influence weights in P by the allocations of effort given in S. There are many ways in which this could be achieved. Our present approach is motivated by the desire to maintain a tight connection with the standard DeGroot model. We would like to think of (1) as the special case of (3) where every agent plays a neutral strategy. Concretely, we require that $P(S^*) = P$. (Remember that S^* is the neutral strategy for P.) This way, we can think of DeGroot's classical model as a description of opinion dynamics in which no agent is a strategic manipulator, in the sense that no agent deliberately tries to spread her opinion by exerting more influence on some agents than on others.

We will make one more assumption about the operation $P(S)$, which we feel is quite natural, and that is that $\texttt{diag}(P(S)) = \texttt{diag}(P)$, i.e., the agents' stubbornness should not depend on how much they or anyone else allocates persuasion effort. In other words, strategies should compete only for the resources of opinion change that are left after subtracting the agent's own stubbornness.

To accommodate these two requirements in a natural way, we define $P(S)$ with respect to a reference point formed by the neutral strategy S^*. For any given strategy matrix S, let $\overline{S}$ be the column-normalized matrix derived from S. $\overline{S}_{ij}$ is i's relative persuasion effort affecting j, when taking into account how much everybody invests in influencing j. We compare $\overline{S}$ to the relative persuasion effort $\overline{S^*}$ under the neutral strategy: call $R = \overline{S}/\overline{S^*}$ the matrix of relative net influences given strategy S.[14] The actual influence matrix $P(S) = Q$ is then defined as a reweighing of P by the relative net influences R:

$$Q_{ij} = \begin{cases} P_{ij} & \text{if } i = j \\ \frac{P_{ij}R_{ji}}{\sum_k P_{ik}R_{ki}}(1 - P_{ii}) & \text{otherwise.} \end{cases} \tag{4}$$

Here is an example illustrating the computation of actual influences. For influence matrix P and strategy matrix S we get the actual influences $P(S)$ as follows:

$$P = \begin{pmatrix} 1 & 0 & 0 \\ .2 & .5 & .3 \\ .4 & .5 & .1 \end{pmatrix} \qquad S = \begin{pmatrix} 0 & .9 & .1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \qquad P(S) \approx \begin{pmatrix} 1 & 0 & 0 \\ .27 & .5 & .23 \\ .12 & .78 & .1 \end{pmatrix}.$$

To get there we need to look at the matrix of relative persuasion effort $\overline{S}$ given by S, the neutral strategy S^* for this P and the relative persuasion effort $\overline{S^*}$ under the neutral

[14] Here and in the following, we adopt the convention that $x/0 = 0$.

strategy:

$$\overline{S} = \begin{pmatrix} 0 & 9/19 & 1/11 \\ 0 & 0 & 10/11 \\ 0 & 10/19 & 0 \end{pmatrix} \qquad S^* = \begin{pmatrix} 0 & .5 & .5 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \qquad \overline{S^*} = \begin{pmatrix} 0 & 1/3 & 1/3 \\ 0 & 0 & 2/3 \\ 0 & 2/3 & 0 \end{pmatrix}.$$

That $\overline{S^*}_{12} = 1/3$, for example, tells us that agent 1's influence of $P_{21} = 1/5$ on agent 2 comes about in the neutral case where agent 1 invests half as much effort into influencing agent 2 as agent 3 does. To see what happens when agent 1 plays a non-neutral strategy, we need to look at the matrix of relative net influences $R = \overline{S}/\overline{S^*}$, which, intuitively speaking, captures how much the actual case $\overline{S}$ deviates from the neutral case $\overline{S^*}$:

$$R = \begin{pmatrix} 0 & 27/19 & 3/11 \\ 0 & 0 & 15/11 \\ 0 & 15/19 & 0 \end{pmatrix}.$$

This derives $P(S) = Q$ by equation (4). We give only one of four non-trivial cases here:

$$\begin{aligned} Q_{21} &= \frac{P_{21}R_{12}}{P_{11}R_{11} + P_{12}R_{21} + P_{13}R_{31}}(1 - P_{22}) \\ &= \frac{2/10 \times 27/19}{1/5 \times 27/19 + 1/2 \times 0 + 3/10 \times 15/19}(1 - 1/2) \\ &\approx 0.27 \end{aligned}$$

In words, by investing 9 times as much into influencing agent 2 than into influencing agent 3, agent 1 gains effective influence of ca. $.27 - .2 = .07$ over agent 2, as compared to when she neutrally divides effort equally among her audience. At the same time, agent 1 looses effective influence of ca. $.4 - .12 = .28$ on agent 3. (This strategy might thus seem to only diminish agent 1's actual influence in the updating process. But, as we will see later on, this can still be (close to) the optimal choice in some situations.)

It remains to check that the definition in (4) indeed yields a conservative extension of the classical DeGroot-process in (1):

Fact 1. $P(\overline{S}) = P$.

Proof. Let $Q = P(\overline{S})$. Look at arbitrary Q_{ij}. If $i = j$, then trivially $Q_{ij} = P_{ij}$. If $i \neq j$, then

$$Q_{ij} = \frac{P_{ij}R_{ji}}{\sum_k P_{ik}R_{ki}}(1 - P_{ii}),$$

with $R = \overline{S^*}/\overline{S^*}$. As $S_{ii} = 0$ by definition of a strategy, we also have $R_{ii} = 0$. So we get:

$$Q_{ij} = \frac{P_{ij}R_{ji}}{\sum_{k \neq i} P_{ik}R_{ki}}(1 - P_{ii}).$$

Moreover, for every $k \neq i$, $R_{kl} = 1$ whenever $P_{lk} > 0$, otherwise $R_{kl} = 0$. Therefore:

$$Q_{ij} = \frac{P_{ij}}{\sum_{k \neq i} P_{ik}}(1 - P_{ii}) = P_{ij}.$$

□

The Propaganda Problem. The main question we are interested in is a very general one:

(9) Propaganda problem (full): Which individual strategies S_i are good or even optimal for promoting agent i's opinion in society?

This is a game problem because what is a good promotion strategy for agent i depends on what strategy all other agents play as well. As will become clear below, the complexity of the full propaganda problem is daunting. We therefore start first by asking a simpler question, namely:

(10) Propaganda problem (restricted, preliminary): Supposing that most agents behave non-strategically like agents in DeGroot's original model (call them: sheep), which (uniform) strategy should a minority of strategic players (call them: wolves) adopt so as best to promote their minority opinion in the society?

In order to address this more specific question, we will assume that initially wolves and sheep have opposing opinions: if i is a wolf, then $x_i(0) = 1$; if i is a sheep, then $x_i(0) = -1$. We could think of this as being politically right wing or left wing; or of endorsing or rejecting a proposition, etc. Sheep play a neutral strategy and are susceptible to opinion change ($P_{ii} < 1$ for sheep i). Wolves are maximally stubborn ($P_{ii} = 1$ for wolves i) and can play various strategies. (For simplicity we will assume that all wolves in a population play the same strategy.) We are then interested in ranking wolf strategies with respect to how strongly they pull the community's *average opinion* $\bar{x}(t) = {}^1/_n \times \sum_{i=1}^{n} x_i(t)$ towards the wolf opinion.

This formulation of the propaganda problem is still too vague to be of any use for categorizing good and bad strategies. We need to be more explicit at least about the number of rounds after which strategies are evaluated. Since we allow wolf strategies to vary over time and/or to depend on other features which might themselves depend on time, it might be that some strategies are good at short intervals of time and others only after many more rounds of opinion updating. In other words, the version of the propaganda problem we are interested in here is dependent on the number of rounds k. For fixed P and $\mathbf{x}(0)$, say that $\mathbf{x}(k)$ results from a sequence of strategy matrices $\left\langle S^{(1)}, \ldots, S^{(k)} \right\rangle$ if for all $0 < i \leq k$: $\mathbf{x}(i) = P(S^{(i)})\,\mathbf{x}(i-1)$.

(11) Propaganda problem (restricted, fixed P): For a fixed P, a fixed $\mathbf{x}(0)$ as described and a number of rounds $k > 0$, find a sequence of k strategy matrices $\left\langle S^{(1)}, \ldots, S^{(k)} \right\rangle$, with wolf and sheep strategies as described above, such that $\bar{\mathbf{x}}(k)$ is maximal for the $\mathbf{x}(k)$ that results from $\left\langle S^{(1)}, \ldots, S^{(k)} \right\rangle$.

What that means is that the notion of a social influencing strategy we are interested here is that of an optimal *sequence* of k strategies, not necessarily a single strategy. Finding a good strategy in this sense can be very computationally heavy, as we would like to make clear in the following by a simple example. It is therefore that, after having established a feeling for how wolf strategies influence population dynamics over time, we will rethink our notion of a social influence strategy once more, arguing that the complexity of the problem calls for heuristics that are easy to apply yet yield good, if sub-optimal, results. But first things first.

Example: Lone-Wolf Propaganda. Although simpler than the full game problem, the problem formulated in (11) is still a very complex affair. To get acquainted with the complexity of the situation, let's look first at the simplest non-trivial case of a society of three agents with one wolf and two sheep: call it a lone-wolf problem. For concreteness, let's assume that the influence matrix is the one we considered previously, where agent 1 is the wolf:

$$P = \begin{pmatrix} 1 & 0 & 0 \\ .2 & .5 & .3 \\ .4 & .5 & .1 \end{pmatrix}. \tag{5}$$

Since sheep agents 2 and 3 are assumed to play a neutral strategy, the space of feasible strategies for this lone-wolf situation can be explored with a single parameter $a \in [0; 1]$:

$$S(a) = \begin{pmatrix} 0 & \text{a} & \text{1-a} \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

We can therefore calculate:

$$\overline{S^*} = \begin{pmatrix} 0 & 1/3 & 1/3 \\ 0 & 0 & 2/3 \\ 0 & 2/3 & 0 \end{pmatrix} \qquad \overline{S(a)} = \begin{pmatrix} 0 & a/a+1 & 1-a/2-a \\ 0 & 0 & 1/2-a \\ 0 & 1/a+1 & 0 \end{pmatrix}$$

$$R = \begin{pmatrix} 0 & 3a/a+1 & 3-3a/2-a \\ 0 & 0 & 3/4-2a \\ 0 & 3/2a+2 & 0 \end{pmatrix} \qquad P(S(a)) = \begin{pmatrix} 1 & 0 & 0 \\ 4a/8a+6 & 1/2 & 3/8a+6 \\ 36-36a/65-40a & 9/26-16a & 1/10 \end{pmatrix}$$

Let's first look at the initial situation with $\mathbf{x}(0)^{-1} = \langle 1, -1, -1 \rangle$, and ask what the best wolf strategy is for boosting the average population in just one time step $k = 1$. The relevant population opinion can be computed as a function of a, using basic algebra:

$$\overline{\mathbf{x}(1)}(a) = \frac{-224a^2 + 136a - 57}{-160a^2 + 140a + 195}. \tag{6}$$

This function is plotted in Figure 2. Another chunk of basic algebra reveals that this function has a local maximum at $a = .3175$ in the relevant interval $a \in [0; 1]$. In other words, the maximal shift towards wolf opinion in one step is obtained for the wolf strategy $\langle 0, .3175, .6825 \rangle$. This, then, is an exact solution to the special case of the propaganda problem state in (11) where P is given as above and $k = 1$.

How about values $k > 1$? Let's call any k-sequence of wolf strategies that maximizes the increase in average population opinion at each time step the `greedy` strategy. Notice that the `greedy` strategy does not necessarily select the same value of a in each round because each greedy choice of a depends on the actual sheep opinions x_2 and x_3. To illustrate this, Figure 3 shows (a numerical approximation of) the `greedy` values of a for the current example as a function of all possible sheep opinions. As is quite intuitive, the plot shows that the more, say, agent 3 already bears the wolf opinion, the better it is, when greedy, to focus persuasion effort on agent 2, and vice versa.

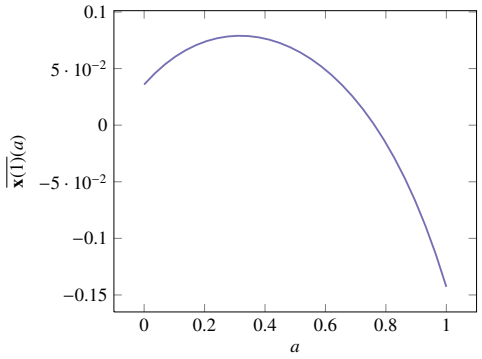


Figure 2: Population opinion after one round of updating with a strategy matrix $S(a)$ for all possible values of a, as described by the function in Equation (6).

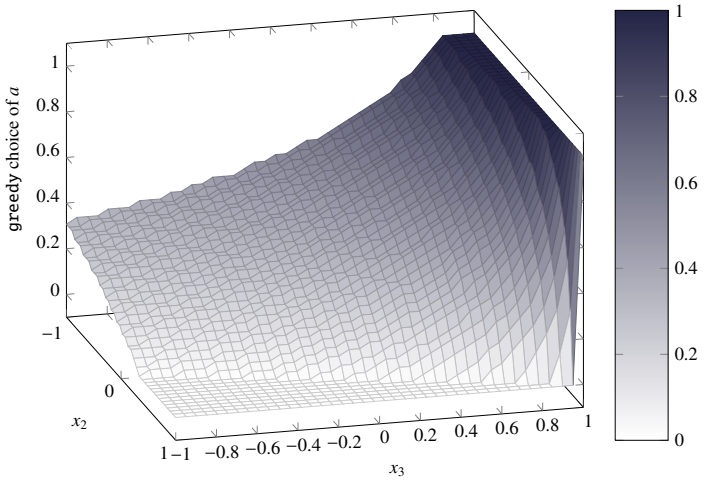


Figure 3: Dependency of the `greedy` strategy on the current sheep opinion for the lone-wolf problem given in (5). The graph plots the best choice of effort a to be allocated to persuading agent 2 for maximal increase of population opinion in one update step, as a function of all possible pairs of sheep opinions x_2 and x_3.

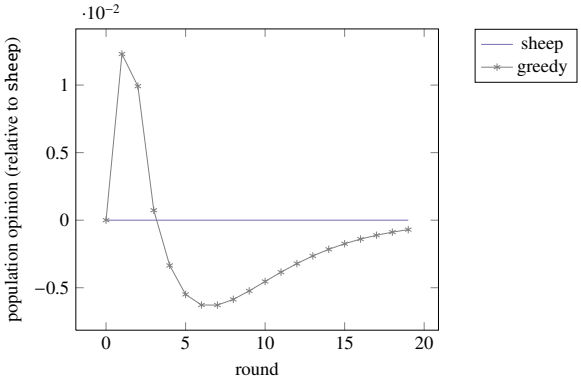


Figure 4: Temporal development of relative opinion (i.e., average population opinion relative to average population opinion under baseline strategy `sheep`) for several wolf strategies for the influence matrix in (5)

It may be tempting to hypothesize that strategy `greedy` solves the lone-wolf version of (11) for arbitrary k. But that's not so. From the fourth round onwards even playing the neutral strategy `sheep` (a constant choice of $a = 1/2$ in each round) is better than strategy `greedy`. This is shown in Figure 4, which plots the temporal development over 20 rounds of what we will call *relative opinion* for our current lone-wolf problem. Relative opinion of strategy X is the average population opinion as it develops under strategy X minus the average population opinion as it develops under baseline strategy `sheep`. Crucially, the plot shows that the relative opinion under `greedy` greedy choices falls below the baseline of non-strategic DeGroot play already very soon (after 3 rounds). This means that the influence matrix P we are looking at here provides a counterexample against the *prima facie* plausible conjecture that playing `greedy` solves the propaganda problem in (11) for all k.

The need for heuristics. Of course, it is possible to calculate a sequence of a values for any given k and P that strictly maximizes the population opinion. But, as the previous small example should have made clear, the necessary computations are so complex that it would be impractical to do so frequently under "natural circumstances", such as under time pressure or in the light of uncertainty about P, the relevant k, the current opinions in the population etc. This holds in particular when we step beyond the lone-wolf version of the propaganda problem: with several wolves the optimization problem is to find the *set* of wolf strategies that are optimal *in unison*. Mathematically speaking, for each fixed P, this is a multi-variable, non-linear, constrained optimization problem. Oftentimes this will have a unique solution, but the computational complexity of the relevant optimization problem is immense. This suggests the usefulness, if not necessity of simpler, but still efficient *heuristics*.[15] For these reasons we focus in

[15] Against this it could be argued that processes of evolution, learning and gradual optimization might have brought frequent manipulators at least close to the analytical optimum over time. But even then, it is dubious that the agents actually have the precise enough knowledge (of influence matrix P, current population opin-

the following on intuitive and simple ways of playing the social manipulation game that make, for the most part, more innocuous assumptions about agents' computational capacities and knowledge of the social facts at hand. We try to demonstrate that these heuristics are not only simple, but also lead to quite good results on average, i.e., if uniformly applied to a larger class of games.

To investigate the average impact of various strategies, we resort to numerical simulation. By generating many random influence matrices P and recording the temporal development of the population opinion under different strategies, we can compare the average success of these strategies against each other.

Towards efficient heuristics. For reasons of space, we will only look at a small sample of reasonably successful and resource efficient heuristics that also yield theoretical insights into the nature of the propaganda problem. But before going into details, a few general considerations about efficient manipulation of opinions are in order. We argue that in general for a manipulation strategies to be efficient it should: (i) not preach to the choir, (ii) target large groups, not small groups or individuals, (iii) take other manipulators into account, so as not to get into one another's way and (iv) take advantage of the social structure of society (as given by P). Let's look at all of these points in turn.

Firstly, it is obvious upon thought that any effort spent on a sheep which is already convinced, i.e., holds the wolf opinion one, is wasted.[16] A minimum standard for a rational wolf strategy would therefore be to spend no effort on audience members with opinion one (it will never get bigger than one) as long as there are audience members with opinion lower than one. All of the strategies we look at below (implicitly) conform to this requirement.

Secondly, we could make a distinction between strategies that place all effort onto just one audience member and strategies that place effort on more than one audience member (in the most extreme case that would be *all* of the non-convinced audience members). Numerical simulations show that, on average, strategies of the former kind clearly prove inferior to strategies of the latter kind. An intuitive argument why that is so is the following. For concreteness, consider the lone-wolf greedy maximization problem plotted in Figure 2. (The argument holds in general.) Since the computation of $P(S)$ relies on the *relative* net influence R, playing extreme values ($a = 0$ or $a = 1$) is usually suboptimal because the influence gained on one agent is smaller than the influence lost on the other agent. This much concerns just one round of updating, but if we look at several rounds of updating, then influencing several agents to at least some extent is beneficial, because the increase in their opinion from previous rounds will lead to more steady increase in population opinion at later rounds too. All in all, it turns out that efficient manipulation of opinions, on a short, medium and long time scale, is achieved better if the web of influence is spread wide, i.e., if many or all suitable members of the wolves' audience are targeted with at least *some* persuasion

ion, etc.) to learn to approximate the optimal strategy. Due to reasons of learnability and generalizability, what evolves or is acquired and fine-tuned by experience, too, is more likely a good heuristic.

[16] Strictly speaking, this can only happen in the limit, but this is an issue worth addressing, given (i) floating number imprecision in numerical simulations, and (ii) the general possibility (which we do not explicitly consider) of small independent fluctuations in agents' opinion dynamics.

effort. For simplicity, the strategies we consider here will therefore target all non-convinced members of each wolf's audience, but variably distribute persuasion effort among these.

Thirdly, another relevant distinction of wolf strategies is between those that are sensitive to the presence and behavior of other wolves and those that are not. The former may be expected to be more efficient, if implemented properly, but they are also more sophisticated. This is because they pose stronger requirements on the agents that are thought to implement these strategies: wolves who want to hunt in a pack should be aware of the other wolves and adapt their behavior to form an efficient coalition strategy. We will look at just one coalition strategy here, but find that, indeed, this strategy is (one of) the best from the small sample that is under scrutiny here. Surprisingly, the key to coalitional success is not to join forces, but rather to get out of each others way. Intuitively, this is because if several manipulators invest in influencing the same sheep, they thereby decrease their *relative* net influence unduly. On the other hand, if a group of wolves decides who is the main manipulator, then by purposefully investing little effort the other wolves boost the main manipulator's relative net influence.

Fourthly and finally, efficient opinion manipulation depends heavily on the social structure of the population, as given by P. We surely expect that a strategy which uses (approximate) knowledge of P in a smart way will be more effective than one that does not. The question is, of course, what features of the social structure to look at. Below we investigate two kinds of socially-aware heuristics: one that aims for sheep that can be easily influenced, and one that aims for sheep that are influential themselves. We expected that the former do better in the short run, while the latter might catch up after a while and eventually do better in the long run. This expectation is borne out, but exactly how successful a given strategy (type) is also depends on the structure of the society.

The cast. Next to strategy `sheep`, the strategies we look at here are called `influence`, `impact`, `eigenvector` and `communication`. We describe each in turn and then discuss their effectiveness, merits and weaknesses.

Strategy `influence` chooses a fixed value of a in every round, unlike the time-dependent `greedy`. Intuitively speaking, the strategy `influences` allocates effort among its audience proportional to how much influence the wolf has on each sheep: the more a member of an audience is susceptible to being influenced, the more effort is allocated to her. In effect, strategy `influence` says: "allocate effort relatively to how much you are being listened to". In our running example with P as in Equation (5) the lone wolf has an influence on (sheep) agent 2 of $P_{12} = 1/5$ and of $P_{13} = 2/5$ on agent 3. Strategy `influence` therefore chooses $a = 1/3$, because the wolf's influence over agent 2 is half as big as that over agent 3.

Strategy `impact`, in a sense, does the exact opposite of strategy `influence`. Intuitively speaking, this strategy says: "allocate effort relatively to how much your audience is being listened to". The difference between `influence` and `impact` is thus that the former favors those the wolf has big influence over, while the latter favors those that have big influence themselves. To determine influence, strategy `impact` looks at the column vector P^T_j for each agent $j \in A(i)$ in wolf i's audience. This column vec-

tor P_j^T captures how much *direct influence* agent j has. We say that sheep j has more direct influence than sheep j' if the sum of the j-th column is bigger than that of the j'-th. (Notice that the rows, but not the columns of P must sum to one, so that some agents may have more direct influence than others.) If we look at the example matrix in equation (5), for instance, agent 2 has more direct influence than agent 3. The strategy `impact` then allocates persuasion effort proportional to relative direct influence among members of an audience. In the case at hand, this would lead to a choice of $a = \frac{\sum_k P_{k2}}{\sum_k P_{k2} + \sum_k P_{k3}} = {}^5/_{12}$.

Strategy `eigenvector` is very much like `impact`, but smarter, because it looks beyond *direct* influence. Strategy `eigenvector` for wolf i also looks at how influential the audience of members of i's audience is, how influential their audience is and so on *ad infinitum*. This transitive closure of social influence of all sheep can be computed with the (right-hand) eigenvector of the matrix P^*, where P^* is obtained by removing from P all rows and columns belonging to wolves.[17,18] For our present example, the right-hand unit eigenvector of matrix

$$P' = \begin{pmatrix} .5 & .3 \\ .5 & .1 \end{pmatrix}$$

is approximately $\langle .679, .321 \rangle$. So the strategy `eigenvector` would choose a value of approximately $a = .679$ at each round.

Finally, we also looked at one coalition strategy, where wolves coordinate their behavior for better effect. Strategy `communication` is such a sophisticated coalition strategies that also integrates parts of the rationale behind strategy `influence`. Strategy `communication` works as follows. For a given target sheep i, we look at all wolves among the influences $I(i)$ of i. Each round a main manipulator is drawn from that group with a probability proportional to how much influence each potential manipulator has over i. Wolves then allocate 100 times more effort to each sheep in their audience that they were assigned main manipulator for than to those sheep some other wolf is taking on this round. Since this much time-variable coordination seems only plausible, when wolves can negotiating their strategies each round, we refer to this strategy as `communication`.

We expect strategy `influence` and `communication` to have similar temporal properties, namely to outperform baseline strategy `sheep` in early rounds of play. `Communication` is expected to be better than `influence` because it is the more sophisticated coalitional strategy. On the other hand, strategies `impact` and `eigenvector` should be better at later rounds of updating because they invest in manipulating influential or "central" agents of the society, which may be costly at first, but should pay off later on. We expect `eigenvector` to be better than `impact` because it is the more sophisticated social strategy that looks beyond *direct* influence at the *global* influence that agents have in the society.

[17] Removing wolves is necessary because wolves are the most influential players; in fact, since they are maximally stubborn, sheep would normally otherwise have zero influence under this measure.

[18] The DeGroot-process thereby gives a motivation for measures of eigenvector centrality, and related concepts such as the Google page-rank (cf. Jackson, 2008). Unfortunately, the details of this fascinating topic are off-topic in this context.

Experimental set-up. We tested these predictions by numerical simulation in two experiments, each of which assumed a different interaction structure of the society of agents. The first experiment basically assumed that the society is homogeneous, in the sense that (almost) every wolf can influence (almost) every sheep and (almost) every sheep interacts with (almost) every sheep. The second experiment assumed that the pattern of interaction is heterogeneous, in the sense that who listens to whom is given by a scale-free small-world network. The latter may be a more realistic approximation of human society, albeit still a strong abstraction from actual social interaction patterns.

Both experiments were executed as follows. We first generated a random influence matrix P, conforming to either basic interaction structure. We then ran each of the four strategies we described above on each P and recorded the population opinion at each of 100 rounds of updating.

Interaction networks. In contrast to the influence matrix P, which we can think of as the adjacency matrix of a directed and weighted graph, we model the basic interaction structure of a population, i.e., the qualitative structure that underlies P, as an undirected graph $G = \langle N, E\rangle$ where $N = \{1, \ldots, n\}$ is the set of nodes, representing the agents, and $E \subseteq N \times N$ is a reflexive and symmetric relation on N.[19] If $\langle i, j\rangle \in E$, then, intuitively speaking, i and j know each other, and either agent could in principle influence the opinion of the other. For each agent i, we consider $N(i) = \{j \in N \mid \langle i, j\rangle \in E\}$ the set of i's *neighbors*. The number of i's neighbors is called agent i's *degree* $d_i = |N(i)|$. For convenience, we will restrict attention to *connected* networks, i.e., networks all of whose nodes are connected by some sequences of transitions along E. Notice that this also rules out agents without neighbors.

For a homogeneous society, as modelled in our first experiment, we assumed that the interaction structure is given by a totally connected graph. For heterogeneous societies, we considered so-called *scale-free small-world networks* (Barabási and Albert, 1999; Albert and Barabási, 2002). These networks are characterized by three key properties which suggest them as somewhat realistic models of human societies (c.f. Jackson, 2008):

(1.) scale-free: at least some part of the distribution of degrees has a power law character (i.e., there are very few agents with many connections, and many with only a few);

(2.) small-world:

  (a.) short characteristic-path length: it takes relatively few steps to connect any two nodes of the network (more precisely, the number of steps necessary increases no more than logarithmically as the size of the network increases);

  (b.) high clustering coefficient: if j and k are neighbors of i, then its likely that j and k also interact with one another.

We generated random scale-free small-world networks using the algorithm of Holme and Kim (2002) with parameters randomly sampled from ranges suitable to produce

[19] Normally social network theory takes E to be an irreflexive relation, but here we want to include all self-connections so that it is possible for all agents to be influenced by their own opinion as well.

networks with the above mentioned properties. (We also added all self-edges to these graphs; see Footnote 19.)

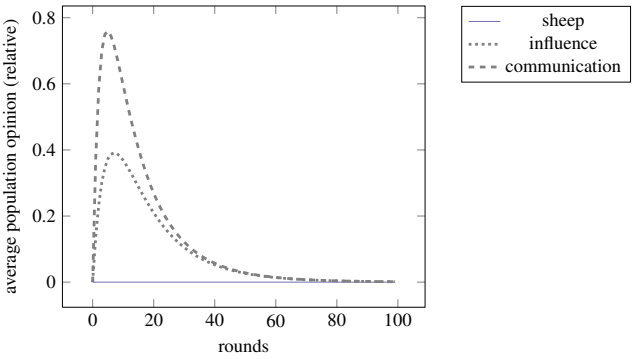
For both experiments, we generated graphs of the appropriate kind for population sizes randomly chosen between 100 and 1000. We then sampled a number of wolves averaging around 10% of the total number of agents (with a minium of 5) and randomly placed the wolves on the network. Subsequently we sampled a suitable random influence matrix P that respected the basic interaction structure, in such a way that $P_{ij} > 0$ only if $\langle i, j \rangle \in E$. In particular, for each sheep i we independently sampled a random probability distribution (using the r-Simplex algorithm) of size d_i and assigned the sampled probability values as the influence that each $j \in N(i)$ has over i. As mentioned above, we assumed that wolves are unshakably stubborn ($P_{ii} = 1$).

Results. For the most part, our experiments vindicated our expectations about the four different strategies that we tested. But there were also some interesting surprises.

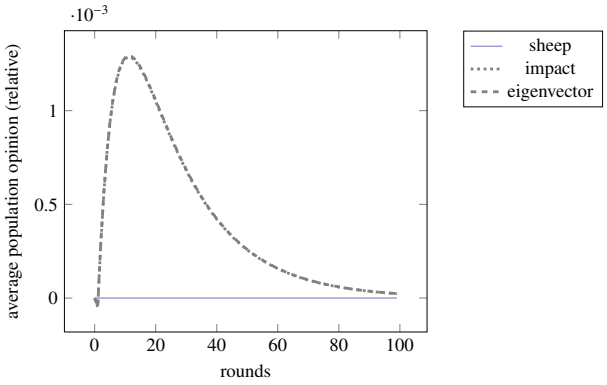
The temporal development of average relative opinions under the relevant strategies is plotted in Figure 5 for homogeneous societies and in Figure 6 for heterogeneous societies. Our general expectation that strategies `influence` and `communication` are good choices for fast success after just a few rounds of play is vindicated for both types of societies. On the other hand, our expectation that targeting influential players with strategies `impact` and `eigenvector` will be successful especially in the long run did turn out to be correct, but only for the heterogeneous society, not for the homogeneous one. As this is hard to see from Figures 5 and 6, Figure 7 zooms in on the distribution of relative opinion means at the 100^{th} round of play.

At round 100 relative means are very close together because population opinion is close to wolf opinion already for all strategies. But even though the relative opinions at the 100^{th} round are small, there are nonetheless significant differences. For homogeneous societies we find that *all* means of relative opinion at round 100 are significantly different ($p < .05$) under a paired Wilcoxon test. Crucially, the difference between `influence` and `impact` is highly significant ($V = 5050$, $p < 005$). For the heterogeneous society, the difference between `influence` and `impact` is also significant ($V = 3285$, $p < 0.01$). Only the means of `communication` and `influence` turn out not significantly different here.

Indeed, contrary to expectation, in homogeneous societies strategies preferentially targeting influenceable sheep were more successful on average for *every* $0 < k \leq 100$ than strategies preferentially targeting influential sheep. In other words, the type of basic interaction structure has a strong effect on the success of a given (type of) manipulation strategy. Although we had expected such an effect, we had not expected it to be that pronounced. Still, there is a plausible *post hoc* explanation for this observation. Since in homogeneous societies (almost) every wolf can influences (almost) every sheep, wolves playing strategies `impact` and `eigenvector` invest effort (almost) exactly alike. But that means that most of the joint effort invested in influencing the same targets is averaged out, because everybody heavily invests in these targets. In other words, especially for homogeneous societies playing a coalitional strategy where manipulators do not get into each other's way are important for success. If this explanation is correct, then a very interesting practical advice for social influencing is ready



(a) Strategies targeting influenceable sheep



(b) Strategies targeting influential sheep

Figure 5: Development of average population opinion in homogeneous societies (averaged over 100 trials). The top graph plots the results for strategies targeting influenceable sheep, while bottom graph shows strategies targeting influential sheep. Although curves are similarly shaped, notice that the y-axes are scaled differently. On the short-term strategies `influence` and `communication` are *much* better than `impact` and `eigenvector` .

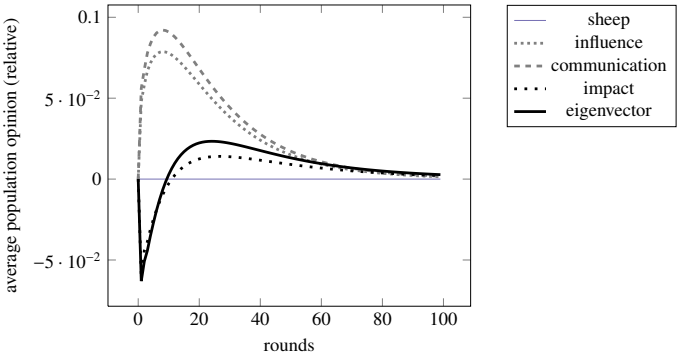
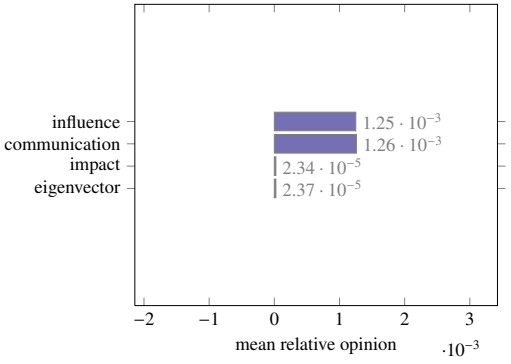
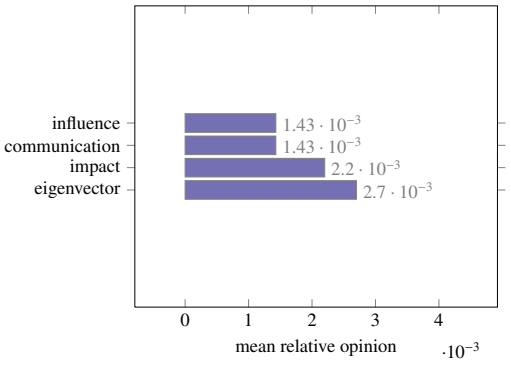


Figure 6: Development of average relative opinion in heterogeneous societies (averaged over 100 trials).



(a) homogeneous society



(b) heterogeneous society

Figure 7: Means of relative opinion at round 100.

at hand: given the ever more connected society that we live in, with steadily growing global connectedness through telecommunication and social media, it becomes more and more important for the sake of promoting one's opinion within the whole of society to team-up and join a coalition with like-minded players.

4 Conclusion

This paper investigated strategies of manipulation, both from a pragmatic and from a social point of view. We surveyed key ideas from formal choice theory and psychology to highlight what is important when a single individual wants to manipulate the choice and opinion of single DM. We also offered a novel model of strategizing opinion dynamics. Important for both pragmatic and social aspect were heuristics, albeit it in a slightly different sense here and there: in order to be a successful one-to-one manipulator, it is important to know the heuristics and biases of the agents one wishes to influence; in order to be a successful one-to-many manipulator, it may be important to use heuristics oneself.

We would like to believe that the practical usefulness of these considerations are, at least, non-negligible. Still, we admit that it is, at best, limited. Many important features of strategic manipulation have not been addressed and must be left for future work. Most strikingly, the model of social influencing given in the second part of the paper is heavily simplistic in a number of ways. We have not at all addressed the case where several manipulators with different motives compete at influencing the population opinion. We have also assumed that the pragmatic one-to-one aspect of opinion manipulation does not play a role when it comes to the social problem of opinion manipulation. Of course, there are obvious interactions: the social structure of the population will likely also affect *which* information to present to *whom* and *how* to present information to *this* or *that* guy. The future challenge, to which this paper may have made its modest contribution, lies in combining these pragmatic, psychological and social aspects into a comprehensive, yet manageable and informativ formal model of strategic manipulation.

References

Acemoglu, Daron and Asuman Ozdaglar (2011). "Opinion Dynamics and Learning in Social Networks". In: *Dynamic Games and Applications* 1.1, pp. 3–49.

Albert, Réka and Alber-László Barabási (2002). "Statistical Mechanics of Complex Networks". In: *Reviews of Modern Physics* 74.1, pp. 47–97.

Axelrod, Robert (1997). "The Dissemination of Culture: A Model with Local Convergence and Global Polarization". In: *Journal of Conflict Resolution* 41.2, pp. 203–226.

Barabási, Alber-László and Réka Albert (1999). "Emergence of Scaling in Random Networks". In: *Science* 286.5439, pp. 509–512.

Benz, Anton (2006). "Utility and Relevance of Answers". In: *Game Theory and Pragmatics*. Ed. by Anton Benz et al. Palgrave, pp. 195–219.

— (2007). “On Relevance Scale Approaches”. In: *Proceedings of Sinn und Bedeutung 11*. Ed. by Estela Puig-Waldmüller, pp. 91–105.
Benz, Anton and Robert van Rooij (2007). “Optimal Assertions and what they Implicate”. In: *Topoi* 26, pp. 63–78.
Camerer, Colin F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.
Camerer, Colin F. et al. (2004). “A Cognitive Hierarchy Model of Games”. In: *The Quarterly Journal of Economics* 119.3, pp. 861–898.
Castellano, Claudio et al. (2009). “Statistical Physics of Social Dynamics”. In: *Reviews of Modern Physics* 81, pp. 591–646.
Crawford, Vincent P. (2003). “Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions”. In: *American Economic Review* 93.1, pp. 133–149.
— (2007). “Let’s Talk It Over: Coordination Via Preplay Communication With Level-k Thinking”. Unpublished manuscript.
Davis, Douglas D. and Charles A. Holt (1993). *Experimental Economics*. Princeton: Princeton University Press.
DeGroot, Morris H. (1974). “Reaching a Consensus”. In: *Journal of the American Statistical Association* 69.345, pp. 118–121.
Farrell, Joseph (1988). “Communication, Coordination and Nash Equilibrium”. In: *Economic Letters* 27.3, pp. 209–214.
— (1993). “Meaning and Credibility in Cheap-Talk Games”. In: *Games and Economic Behavior* 5, pp. 514–531.
Farrell, Joseph and Matthew Rabin (1996). “Cheap Talk”. In: *The Journal of Economic Perspectives* 10.3, pp. 103–118.
Feinberg, Yossi (2008). “Meaningful Talk”. In: *New Perspectives on Games and Interaction*. Ed. by Krzysztof R. Apt and Robert van Rooij. Amsterdam: Amsterdam University Press, pp. 105–119.
— (2011a). “Games with Unawareness”. Unpublished manuscript, Stanford University.
— (2011b). “Strategic Communication”. Unpublished manuscript, Stanford University.
Franke, Michael (2009). “Signal to Act: Game Theory in Pragmatics”. PhD thesis. Universiteit van Amsterdam.
— (2010). “Semantic Meaning and Pragmatic Inference in Non-cooperative Conversation”. In: *Interfaces: Explorations in Logic, Language and Computation*. Ed. by Thomas Icard and Reinhard Muskens. Lecture Notes in Artificial Intelligence. Berlin, Heidelberg: Springer-Verlag, pp. 13–24.
— (2011). “Quantity Implicatures, Exhaustive Interpretation, and Rational Conversation”. In: *Semantics & Pragmatics* 4.1, pp. 1–82.
— (submitted). *Pragmatic Reasoning about Unawareness*.
Franke, Michael et al. (2012). “Relevance in Cooperation and Conflict”. In: *Journal of Logic and Computation* 22.1, pp. 23–54.
Gibbons, Robert (1992). *A Primer in Game Theory*. New York: Harvester Wheatsheaf.
Glimcher, Paul W. and Aldo Rustichini (2004). “Neuroeconomics: The Consilience of Brain and Decision”. In: *Science* 306.5695, pp. 447–452.

Glimcher, Paul W. et al., eds. (2009). *Neuroeconomics: Decision Making and the Brain*. Amsterdam: Elsevier.

Grice, Paul Herbert (1975). "Logic and Conversation". In: *Syntax and Semantics, Vol. 3, Speech Acts*. Ed. by Peter Cole and Jerry L. Morgan. Academic Press, pp. 41–58.

Halpern, Joseph Y. and Leandro Chaves Rêgo (2006). "Extensive Games with Possibly Unaware Players". In: *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 744–751.

Hardin, Garret (1968). "The Tragedy of the Commons". In: *Science* 162, pp. 1243–1248.

Hegselmann, Rainer and Ulrich Krause (2002). "Opinion Dynamics and Bounded Confidence: Models, Analysis, and Simulation". In: *Journal of Artificial Societies and Social Simulation* 5.3.

Heifetz, Aviad et al. (2011). "Dynamic Unawareness and Rationalizable Behavior". Unpublished manuscript.

Holme, Petter and Beom Jun Kim (2002). "Growing scale-free networks with tunable clustering". In: *Physical Review E* 65.2, pp. 026107–1–026107–4.

Jackson, Matthew O. (2008). *Social and Economic Networks*. Princeton University Press.

Jäger, Gerhard (2008). "Game-Theoretical Pragmatics". Unpublished manuscript, University of Bielefeld.

Jäger, Gerhard and Christian Ebert (2009). "Pragmatic Rationalizability". In: *Proceedings of Sinn und Bedeutung 13*. Ed. by Arndt Riester and Torgrim Solstad, pp. 1–15.

Kahane, Howard and Nancy Cavender (1980). *Logic and Contemporary Rhetoric*. Belmont: Wadsworth Publishing.

Kahnemann, Daniel and Amos Tversky (1973). "On the Psychology of Prediction". In: *Psychological Review* 80, pp. 237–251.

Lehrer, Keith (1975). "Social consensus and rational agnoiology". In: *Synthese* 31.1, pp. 141–160.

Levinson, Stephen C. (1983). *Pragmatics*. Cambridge, UK: Cambridge University Press.

Lewis, David (1969). *Convention. A Philosophical Study*. Harvard University Press.

Luce, Duncan R. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Martin, George R. R. (1996). *A Game of Thrones*. New York: Bantam Spectra, Random House.

Matthews, Steven A. et al. (1991). "Refining Cheap Talk Equilibria". In: *Journal of Economic Theory* 55, pp. 247–273.

May, Kenneth O. (1945). "Intransitivity, Utility and the Aggregation of Preference Patterns". In: *Econometrica* 22.1, pp. 1–13.

Milgrom, Paul and John Roberts (1986). "Relying on the Information of Interested Parties". In: *RAND Journal of Economics* 17.1, pp. 18–32.

Myerson, Roger B. (1989). "Credible Negotiation Statements and Coherent Plans". In: *Journal of Economic Theory* 48.1, pp. 264–303.

Newel, B. et al. (2007). *Straight Choices: The Psychology of Decision Making*. Princeton: Princeton University Press.

O'Keefe, Daniel J. and Jakob D. Jensen (2007). "The relative persuasiveness of gain-framed and loss-framed messages for encouraging disease prevention behaviors". In: *Journal of Health and Communication* 12, pp. 623–644.

Ozbay, Erkut Y. (2007). "Unawareness and Strategic Announcements in Games with Uncertainty". In: *Proceedings of TARK XI*. Ed. by Dov Samet, pp. 231–238.

Parikh, Prashant (1991). "Communication and Strategic Inference". In: *Linguistics and Philosophy* 473–514.14, p. 3.

— (2001). *The Use of Language*. Stanford University: CSLI Publications.

— (2010). *Language and Equilibrium*. MIT Press.

Pauly, Marc (2005). "Changing the Rules of Play". In: *Topoi* 24.2, pp. 209–220.

Rabin, Matthew (1990). "Communication between Rational Agents". In: *Journal of Economic Theory* 51, pp. 144–170.

Rogers, Brian W. et al. (2009). "Heterogeneous Quantal Response Equilibrium and Cognitive Hierarchies". In: *Journal of Economic Theory* 144.4, pp. 1440–1467.

van Rooij, Robert and Michael Franke (to appear). "Promises and Threats with Conditionals and Disjunctions". In: *Discourse and Grammar (SGG)*. Ed. by Günther Grewendorf and Thomas Ede Zimmermann. Berlin: DeGruyter.

van Rooy, Robert (2003). "Quality and Quantity of Information Exchange". In: *Journal of Logic, Language and Computation* 12, pp. 423–451.

Shin, Hyun Song (1994). "The Burden of Proof in a Game of Persuasion". In: *Journal of Economic Theory* 64.1, pp. 253–264.

Simon, Herbert A. (1959). "Theories of decision-making in economics and behavioral science". In: *American Economic Review*.

Stalnaker, Robert (2006). "Saying and Meaning, Cheap Talk and Credibility". In: *Game Theory and Pragmatics*. Ed. by Anton Benz et al. Hampshire: Palgrave MacMillan, pp. 83–100.

Tversky, Amos (1972). "Elimination by Aspects: A Theory of Choice." In: *Psychological Review* 79.4, pp. 281–299.

Tversky, Amos and Daniel Kahnemann (1974). "Judgement under Uncertainty: Heuristics and Biases". In: *Science* 185, pp. 1124–1131.

— (1981). "The Framing of Decisions and the Psychology of Choice". In: *Science* 211.4481, pp. 453–458.

Zapater, Inigo (1997). "Credible Proposals in Communication Games". In: *Journal of Economic Theory* 72, pp. 173–197.