Essays in Political Economics and Information Acquisition

by

Giovanni Reggiani

M.Sc. Economics and Social Sciences, Bocconi University (2010)
B.A. Economics and Social Sciences, Bocconi University (2008)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Author .. Signature redacted ....
Department of Economics
May 15, 2016

Certified by.. Signature redacted ....
Abhijit Banerjee
Ford Foundation International Professor of Economics
Thesis Supervisor

Certified by... Signature redacted ....
Bengt Holmstrom
Paul A. Samuelson Professor of Economics
Thesis Supervisor

Accepted by .... Signature redacted ....
Ricardo J. Caballero
~~Ford International~~ Professor of Economics
Chairman, Departmental Committee on Graduate Studies

Essays in Political Economics and Information Acquisition

by

Giovanni Reggiani

Submitted to the Department of Economics
on May 15, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis consists of three chapters respectively on optimal contracts to incentivize information acquisition, strategic voting, and conflict of interest.

The first chapter, joint work with A. Clark, studies a principal-agent problem with limited-liability where an agent is hired to acquire information and take a decision on behalf of a risk-neutral principal. The principal cannot monitor the agent's attentiveness when acquiring information and so she provides incentives with a contract that depends on the realized state of the world and the chosen decision. We build a model for this problem where the agent's cost of acquiring information is given by the average reduction in entropy. We show that the optimal contract has a linear structure: the agent receives a fixed fraction of output together with a state and decision contingent payment. The optimal contract is simple, in terms of dimensionality, and features an incentive structure analogous to that of portfolio managers in the hedge fund industry. We extend this result to problems with arbitrary utilities, a generalized form of cost functions, a participation constraint for the agent, a wealth constraints for the principal, and imperfect revelation of the state. We also show that only entropic costs can generate the separability of state and decision payments and solve for the equivalent optimal contract in a dynamic setting. Lastly we perform Monte Carlo simulations to test the robustness of our initial contract for different utilities and compare its welfare to purely linear and to unrestricted contracts.

The second chapter, joint with F. Mezzanotti, provides a lower bound for the extent of strategic voting. Voters are strategic if they switch their vote from their favorite candidate to one of the main contenders in a tossup election. High levels of strategic voting are a concern for the representativity of democracy and the allocation efficiency of government goods and services. Recent work in economics has estimated that up to 80% of voters are strategic. We use a clean quasi experiment to highlight the shortcomings of previous identification strategies, which fail to fully account for the strategic behavior of parties. In an ideal experiment we would like to observe two identical votes with exogenous variation in the party victory probability. Among world parliamentary democracies 104 have a unique Chamber, 78 have two Chambers with different functions, and only one nation has two Chambers with the same identical functions: Italy. This allows us to observe two identical votes and therefore a valid counterfactual. In addition, the majority premia are calculated at the national level for the Congress ballot and at the regional level for the Senate ballot. This provides exogenous variation in the probability of victory. Because the two Chambers have identical functions, a sincere voter should vote for the same coalition in the two ballots. A strategic voter would instead respond to regions' specific victory probabilities. We combine this intuition with a geographical Regression Discontinuity approach, which allows us to compare voters across multiple Regional boundaries. We find much smaller estimates (5%) that we interpret as a lower bound but argue that it is a credible estimate. We also reconcile our result with the literature larger estimates (35% to 80%) showing how previous estimates could have confounded strategic parties and strategic voters due to the use of a non identical vote as counterfactual.

The third chapter estimates the distortions due to conflict of interest during Berlusconi's rule over Italy. The identification is based on the efficient market hypothesis. In particular, I use electoral polls and stock market data to estimate the effect of surprising electoral outcomes, defined as the difference between actual and expected electoral results, on the stock market performance of Berlusconi's firms. I find evidence that there are substantial distortions due to conflict of interest: 6% increase in market capitalization per percentage point of a positive electoral surprise. I then match two of Berlusconi's companies operating in the same media sector but in different countries. This allows me to further test whether the extra returns are due to political distortions under different regulatory authorities. I find that the abnormal returns can be ascribed to "conflict of interest" rather than to the CEO-founder stepping down. Finally, I perform robustness tests to ensure that the cumulative abnormal returns estimates are not spurious.

Thesis Supervisor: Abhijit Banerjee
Title: Ford Foundation International Professor of Economics

Thesis Supervisor: Bengt Holmstrom
Title: Paul A. Samuelson Professor of Economics

Contents

Acknowledgements

I am infinitely grateful to my advisors Abhijit Banerjee and Bengt Holmstrom. Abhijit's deep insights and knowledge have been invaluable in guiding me through every one of my research projects in the fields of political, theoretical and development economics. I cannot thank him enough. I am also tremendously indebted for his mentorship, support, and encouragement throughout the whole graduate program. Bengt's suggestions and comments have been precious in uncovering new results and simplifying arguments. I am also extremely grateful for his encouragement and guidance on research ideas and all his support on the job market.

I would also like to thank Juuso Toikka, George-Marios Angeletos, and Ben Olken for all their feedback, which has substantially improved my research. Alberto Alesina and Bob Gibbons have been extremely generous sources of advice outside the MIT Economics Department. I have the deepest gratitude for my earlier mentors in economics: Eliana La Ferrara and Pierpaolo Battigalli. Their teaching and investment in me opened the doors to MIT and to the opportunities I have been blessed with.

I have been extremely fortunate to conduct research with an amazing group of coauthors and friends. I am grateful for all I have learned with and from Aubrey Clark, Filippo Mezzanotti, Natalia Rigol, and Reshma Hussam. I am blessed for the friendships forged during these years and for the good times I shared with Andrea Stella, Daan Struyven, David Colino, Giulia Brancaccio, Johann Blauth, Juan Passadore, Ludovica Gazze', Luigi Iovino, Matt Rognlie, Ruchir Agarwal, Sara Hernandez, Tommaso Denti, and Tilman Dette.

The love from friends around the world has been fundamental during this journey. Thanks to Nikolai and Giovanni for exploring with me realms deep below, and to Perin and Daniele for flying with me far above. My heartfelt thanks goes to Amanda, Andres, Carlos, Chiara, Davide, Danilo, Giuseppe, Laura, Luca, and Stefano for a lifetime of friendship and for always being there in good and bad times.

I thank Roberto and Silvia for their love, unconditional support, and constant encouragement. Above all, thanks to my parents Domenico and Franca, to whom I owe everything.

Chapter 1

Optimal Contracts for Information Acquisition under Entropic Costs

Essays in Political Economics and Information Acquisition

by

Giovanni Reggiani

M.Sc. Economics and Social Sciences, Bocconi University (2010)
B.A. Economics and Social Sciences, Bocconi University (2008)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Author ....
Department of Economics
May 15, 2016

Accepted by ....
Ricardo J. Caballero
Ford International Professor of Economics
Chairman, Departmental Committee on Graduate Studies

1.1 Introduction

Most employees are no longer laborers that exert effort but rather experts and data analysts that need to collect information and use their experience to take the appropriate action. The prevalence of information tasks begs the question of how information-employees ought to be incentivized and how these information contracts differ from standard effort contracts.

Demski and Sappington [1987], and more recently Zermeno [2011] and Carroll [2013], have studied this information agency problem. Following Sims [2003], rational inattention, modeled through entropy, has been used to study information problems, strategic games, bargaining, and optimal security design.

Our model bridges these two literatures, we assume that it is costly for the agent to pay attention and that the principal can incentivize the agent's information acquisition and decision. We allow the agent to be risk averse, to have limited liability, and impose a very natural restriction, borrowed from information theory, on the cost of acquiring information. The solution of the model shows that the optimal contract has a linear structure on output and state and decision payments.

This result connects our work with the literature on linear contracts. Optimal linear contracts have long been sought as a holy grail for the applied-theorist's toolbox because of their simplicity and practical prevalence. In principal-agent settings, Holmstrom and Milgrom [1987] first derived a foundation for linear contracts under the assumptions of exponential utility and normal distributions.

In this model, we build on Zermeno [2011] and assume that output is the result of an action suggested by the agent and the realization of a state of the world. Different actions might be best in different states of the world. The agent collects and processes information to tailor the action to the state. The principal observes ex-post the action that was taken and the state of the world realized and designs a contract to align the incentives of the agent with hers. Observing the state of the world and the action taken allows the principal to construct counterfactuals for all possible alternative actions. The principal will want to use these counterfactuals to reward, or punish, the agent. Consider the following example to clarify the setup and build some intuition.

Example 1. The managers of a hedge fund need to specify a contract to incentivize the analysts to process information and make the right financial investment. The analyst gives a recommendation on whether to buy or sell an asset. The trade payoff depends on the action suggested and a state of the world that is observed ex post. Assume that there are three states: low, high and medium. The outcome in the medium states is independent of the action taken, action *buy* is optimal in state *low*, and action *sell* is optimal in state *high*. The table below represents the payoff to the principal of the action a_j in state of the world θ.

$y(\theta, a)$	price low $= \theta_l$	price medium $= \theta_m$	price high $= \theta_h$
$a_1 =$ buy	1	0	-0.5
$a_2 =$ sell	-0.5	0	1

Table 1.1: Example 1: Payoffs for the Principal

We can see that the benefit of paying attention to the state of the world depends on the state itself. When prices are stable the action taken is irrelevant. Instead, in times of volatility the principals' payoff is highly dependent upon the information acquired. In light of this, it seems intuitive that the principal would like to design a contract that gives higher incentives to collect information in the states where the stakes are higher.

There are three important assumptions at work to derive our result. First, we allow the agent to choose arbitrary signals correlated with the states of the world. Second, we assume that the cost of this signal structure is the average reduction in uncertainty as measured by entropy. Lastly, we assume that the agent has log utility or risk neutral preferences with limited liability.

Intuitively, the choice of arbitrary signals together with the entropy based cost function will allow enough richness in the action space to cast the problem as a standard moral hazard where the agent chooses probabilities. This will have the benefit of making the gross utility linear in the choice variable. Entropy as a measure of uncertainty will yield separability for state and decision payments and the utility assumptions will yield the resulting linearity.

In addition to the linear structure, we draw several insights from the solution of the model. The optimal contract has a much lower dimensionality than an arbitrary incentive scheme. We interpret this lower dimensionality as greater simplicity. Less capable agents, i.e. agents with higher costs of processing information, are given low-powered incentives and the relationship between the cost of information acquisition and the strength of the incentives is monotonic. The optimal contract can be interpreted as a fully linear benchmark contract with respect to a reference action or state. We show how the structure of the optimal contract extends for arbitrary utilities and a generalized version of the cost function. We also prove that we can allow the model to have realistic features such as participation constraints for the agent or wealth constraints for the principal, and limits to the extent of state observability. Lastly, we prove an equivalent result in a dynamic setting and run Monte Carlo simulations to explore the robustness of our contract away from its assumptions and compare it with a purely linear one.

Our work relates to three main literatures. It relates to the rational inattention literature in the use of entropy, to the linear contracts literature on moral hazard and to the information acquisition literature on incentivizing experts. Although entropy had been used as a cost for acquiring information before, notably by Arrow [1985], Sims [2003] introduced entropy to model rational inattention. Recently, rational inattention has been axiomatically founded in De Oliveira et al. [2013], and applied in a variety of models from microfoundations of logit utility (Matejka et al. [2011]) to bargaining with rational inattention (Ravid [2014]) and security design (Yang [2013], Yang [2014], and Hebert [2015]). Caplin and Dean [2013] prove necessary and sufficient conditions for the agent's decision problem and test how well mutual information explains the attention choices of agents in an experimental setting. The folk-result that we use for the reduction of the signal space to the decision space was first discussed for two actions by Woodford [2001] and used by Yang [2013], Yang [2014] and Matejka et al. [2011]. This result holds for any Blackwell ordering preserving cost function (Blackwell et al. [1951],Matejka et al. [2015])[1]. We apply some of the techniques developed by Matejka et al. [2015] and Mattsson and Weibull [2002] in our dynamic extension.

We contribute to the rational inattention literature by highlighting the relationship between entropy and log utility in moral hazard problems, and by showing an exclusive property of entropy-based mutual information in agency settings not highlighted in the literature before: the separability of state and decisions payments for any utility function.

Conceptually our paper sits squarely in the expert moral hazard literature. Demski and Sappington [1987] first considered the problem of incentivizing an expert. Their setting is one without conflict between the effort choice and information collection. Zermeno [2011] gave a very general and powerful representation of the expert-agent problem and characterized the solution for a two state case under limited liability and risk neutrality. In addition, his model recognized the potential conflict between effort choice and the collection of information and provided conditions under which the conflict does not lead to contract distortions. Dang et al. [2013] analyze the optimality of debt as the security that minimizes information acquisition to ensure maximum liquidity. Based on a similar insight, Dang et al. [2014] show that banks, with their opacity, are the institutional setting that best trades off the value of information acquisition for investment selection versus the information costs due to reduced liquidity provision.

Carroll [2013] shares the model setup with Zermeno [2011] but allows an arbitrary number of actions and states, while assuming perfect observability of the state of the world and the action taken. These assumptions rule out the contract distortions that Zermeno [2011] analyzed. Carroll [2013] focuses on describing the optimal contract when one departs from common knowledge about the information gathering technology and when one assumes maxmin utility for the principal. He finds that the optimal contract features restrictions to the agent's action set with the principal artificially limiting the possibility of suggesting certain actions. Although the assumption and techniques are very different, Carroll [2013] also finds that the optimal contract features state dependent payments but not action dependent payments. Both Zermeno [2011] and Carroll [2013] assume that the agent is risk neutral with limited liability while our result holds also for a risk averse agent.

Both Holmstrom and Milgrom [1987] and Diamond [1998] used the vastness of the action set of the agent to force the principal to offer linear contracts in order to align incentives. These results necessitate of

[1] Matejka et al. [2015] recently show that a similar argument works in the dynamic setting for entropy but not for arbitrary Blackwell order preserving cost functions.

assumptions on the signals and their correlation with the state of the world. The agent is not free to choose arbitrary signals and their distribution at a cost. More recently, the optimal linear contracts literature has been reinvigorated by Carroll [2015]'s results drawing on the robustness of linear contracts to unknown information technology sets.

Similarly to Carroll [2013], we allow any finite number of states and actions. In contrast to Zermeno [2011] and Carroll [2013], we allow risk averse utilities. Nevertheless, we place more structure on the cost function. While Carroll [2013] allows arbitrary information gathering technologies, we specify an entropy based cost function. We will argue in the paper why this restriction is appropriate for an information gathering problem and why it is omnipresent in information theory.

We assume that the agent is free to choose signals arbitrarily correlated with the state space. This is reminiscent of the vastness of the choice set behind other linearity results (Chassang [2013]). But in contrast with the older literature on the optimality of linear contracts we do not need to assume a specific distribution on the state space, nor a particular signal structure. So although we maintain functional assumptions on the utility of the agent we are free from distributional assumptions about the signals structure and the states of the world.

The paper proceeds as follows. In the next section we formally present the model, then in section 1.3 we illustrate the properties of the cost function (mutual information) and some of the results that we will use in our derivations. In section 1.4, we solve the agent problem; we use this result in the solution of the principal's problem in section 1.5. There we derive our main result and discuss its interpretation and implications. In section 1.6, we present extensions of our result to other settings, and in section 1.7 we study numerically the robustness of the result for different utility functions and its performance relative to a fully linear contract. Finally, in the last section, we present the equivalent result for a dynamic setting before concluding.

1.2 Model

We start by describing the primitives of the model. The finite set Θ represents the states of the world, and $p()$ the common prior over such states. The agent chooses a decision from a finite set D. Output, $y(d, \theta)$, is determined by the realized state of the world θ and the decision d chosen by the agent. Both the state θ and the decision d are contractible and observable after the realization of the state.

The principal's action is the choice of a contract for the agent. A contract is a payment function based on the realized state and the decision taken: $b : D \times \Theta \to \mathbb{R}$. The principal will design this incentive scheme to incentivize the agent to collect information with different intensities for different states and to choose the optimal action. The principal's utility is given by the residual output after honoring the contract: $y(d, \theta) - b(d, \theta)$.

The agent takes two actions: the choice of an information-signaling structure $(IN(X,\Theta))$ at cost $C\left(IN\left(X,\Theta\right)\right)$ to obtain information about the state of the world and, after observing the signal, a decision $d \in D$. Specifically we assume that the agent can choose any mapping $d() : X \rightarrow \Delta(D)$ from the signals to the set of distributions over the decision space. We assume that the agent has wealth W and his utility is separable in total wealth and the cost of Information collection: $U(W, b(\cdot), IN(X,\Theta)) = u(W + b(d,\theta)) - C(IN(X,\Theta))$.

The timing of the model is as follows, first the principal chooses an incentive scheme $b\left(d,\theta\right)$, after observing his incentive schedule the agent chooses an information structure. Based on the information collected, the agent then takes an action; output is realized and the state of the world and the agent's action are observed by the principal. The principal pays the agent $b\left(d,\theta\right)$.

The problem of the agent is therefore one of choosing a signal X correlated with the state Θ and an action in D based on such signal. Formally the agent needs to choose an outcome space X for the signal, and a probability measure over the product space $X \times \Theta$ that is consistent with the prior $\left[p\left(\theta\right)\right]_{\theta \in \Theta}$. The agent will also pick a decision rule $\left(d\left(\cdot\right)\right)$translating, possibly probabilistically, each signal x into an action:

$$\max_{(X,\mathbb{P}) and\ d(\cdot)} \mathbb{E}_{\mathbb{P}(X\times\Theta)}\left[\mathbb{E}_{d(x)}[u\left(W + b(d\left(x\right),\theta)\right)] - \mu C\left(IN\left(X,\Theta\right)\right)\right]$$

The principal instead needs to choose an incentive scheme $b(d,\theta)$ to best align the collection of information and action of the agent to her own interests subject to the limited liability constraint:

$$\max_{(X,\mathbb{P}) and\ d(\cdot), b(\cdot,\cdot)} \mathbb{E}_{\mathbb{P}(X\times\Theta)}\left[\mathbb{E}_{d(x)}[(y(d\left(x\right),\theta) - b(d\left(x\right),\theta))]\right]$$

s.t.

$$\left(\left(X,\mathbb{P}\right), d\left(\cdot\right)\right) \in \arg\max \mathbb{E}_{\mathbb{P}(X\times\Theta)}\left[\mathbb{E}_{d(x)}[u\left(W + b(d\left(x\right),\theta)\right)] - \mu C\left(IN\left(X,\Theta\right)\right)\right]$$

and

$$b\left(d,\theta\right) \geq 0$$

We will assume that the cost of the information structure $C\left(IN\left(X,\Theta\right)\right)$ is entropy-based mutual information. The cost of a signaling structure is the extent by which the signal reduces the uncertainty about the state of the world.

In the following subsection we describe and discuss such cost function, why it is standard in the information theory literature, and why it is the most sensitive specification for our problem. In section 1.4 we show that, under mutual information, the optimal information structure and action is equivalent to the choice of a state by state probability of action: $p\left(d|\theta\right)$. We then rephrase the problem in this light before proving our main results.

1.3 Mutual Information

The agent faces two tasks, first to collect and process information through the design of an information structure and secondly to take a costless decision based on the information acquired. In this section we define mutual information as in Cover and Thomas [2012] and we detail why this specification is the most reasonable for the problem at hand.

The cost function $C()$ represents the cost of the agent's choice of an outcome space X, and the probability measure $\mathbb{P}$ over the product space $X \times \Theta$. There is, potentially, a large class of costs functions to capture the intuition that having more precise (or more useful) information should be more costly. For instance one could let $C()$ be a function of the number of outcomes in X that have positive probability, or a function of the average reduction in variance of the state of the world θ upon the observation of the signal X[2]. Entropy as a measure of uncertainty respects the Blackwell ordering but it is stronger in that it generates a complete ordering over the set.

Information theory has used for decades mutual information as a measure of cost. Intuitively, mutual information corresponds to the average reduction in "uncertainty" about the state of the world upon the observation of the signal. We will be more specific in the following about what we mean by uncertainty and average reduction, and equally importantly about how this choice relates to other possible cost functions and why it is superior for our problem.

Entropy as a measure of uncertainty

Let X be a discrete random variable defined, with abuse of notation, over an outcome space X and with probability mass function $p(x) = Pr\{X = x\}$. The entropy of the discrete variable X, denoted as $H(X)$, is defined as

$$H(X) = -\mathbb{E}_{p(x)} log\left(p(x)\right)$$

The entropy of a random variable is always non negative and positive if the random variable is non degenerate, and it does not depend on the values of the outcome space X but only on the probability distribution. It's value is finite for any discrete variable.

Shannon's theorem, shows that the minimum expected number of binary signals required to determine X is between $H(X)$ and $H(X)+1$. Entropy as a measure of uncertainty captures the expected length of the most efficient self-punctuating binary encoding. In the appendix we provide a clarifying example of this theorem and discuss alternative measures of uncertainty and their pitfalls.

[2]Gentzkow and Kamenica [2014] propose a variant of the last one.

If we consider the joint distribution of two random variable (X_1, X_2), we can derive their joint entropy from the above definition considering the two variables as a vector. Given the joint probability distribution $p(x_1, x_2)$, we define the conditional entropy $H(X_1|X_2)$ as:

$$H(X_1|X_2) = \sum_{x \in X_2} p(x_2) H(X_1|X_2 = x_2) =$$

$$= \mathbb{E}_{p(x_2)} \left[-\mathbb{E}_{p(x_1|x_2)} \left(log \left(p(x_1|x_2) \right) \right) \right]$$

The following results follow from the definition: if X_1 and X_2 are independent then $H(X_1|X_2) = H(X_1)$. That is, there is no reduction in uncertainty from observing an independent variable. Moreover $H(X_1, X_2) = H(X_1) + H(X_2|X_1)$ and $H(X_2|X_1) \leq H(X_2)$.

Conditional entropy is therefore a measure of what is the average residual uncertainty about a variable upon observing another random variable. Notice that it could well be that in some states of the world the signal provides no information, or even that it actually increases the uncertainty. Nevertheless, on average a signal reduces (at least weakly) the total uncertainty. The following example clarifies this.

Example 2. Consider the random variable X_1 and the following signal X_2 chosen to have the following joint probability

$X_2 \backslash X_1$	a	b	c	d	
γ	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$
η	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{3}$
ω	0	0	0	$\frac{1}{6}$	$\frac{1}{6}$
	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	

Table 1.2: Joint Distribution of the Outcome X_1 and the Signal X_2

It can be easily seen applying the above definition, that entropy is 0 conditional on observing ω, entropy is unchanged conditional on observing η, and increases conditional on observing γ. Nevertheless the expected conditional entropy is strictly lower than the unconditional entropy.

Therefore, the difference $H(X_1) - H(X_1|X_2)$ is always non negative, and represents the average reduction in uncertainty brought about by a signal X_2. If we think that the cost of discovering and analyzing a certain signal should be proportional to how useful it is - i.e. how much it reduces underlying uncertainty - this would be a good cost function to use.

Relative Entropy and Mutual Information

The Relative Entropy or Kullback-Leibler distance between two probability mass functions $p(x)$ and $\pi(x)$ is defined as:

$$D(p||\pi) = \sum_{x_1 \in X_1} p(x_1) \, log \left(\frac{p(x_1)}{\pi(x_1)} \right)$$

Although $D(\cdot||\cdot)$ is not a metric,[3] it represents a "distance". It can be shown that it is the average expected extra length of code that would be necessary if we were to transmit information about π using a code optimized for p. We know from the previous section that the minimum average expected length of a binary self-punctuating code transmitting about p is $H(p)$. If that same code was used to transmit information when the true distribution is π[4], the average expected length of the code would be: $H(p) + D(p||\pi)$.

We present a useful property of the Kullback-Leibler distance that we will use in later results.

Proposition 3. *Let p,q and π, η be probability distributions over a random variable X , then for any $\alpha \in (0,1)$:*

$$D(\alpha p + (1-\alpha) q||\alpha\pi + (1-\alpha)\eta) \leq \alpha D(p||\pi) + (1-\alpha) D(q||\eta)$$

Proof. In Appendix 3.7. □

Definition 4. (Mutual Information) The mutual information between two random variables X_1, X_2 with joint probability mass function $p(x_1, x_2)$ over $X_1 \times X_2$, represented as $I(X_1, X_2)$, is $I(X_1, X_2) = D(p(x_1, x_2)||p_{X_1}(x_1) p_{X_2}(x$
Where $p_{X_1}(x_1) p_{X_2}(x_2)$ are the appropriately defined marginal mass functions derived from $p(x_1, x_2)$.

The following remarks can be easily proved:

Remark 5. $I(X_1, X_2) = I(X_2, X_1)$. $I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2)$, and $I(X_1, X_2) = H(X_1) - H(X_1|X_2)$, which implies $I(X_1, X_2) \geq 0$.

So mutual information can be expressed as $I(X_1, X_2) = H(X_1) - H(X_1|X_2)$: the difference between the entropy of a random variable and its expected entropy conditional on a signal. In our example X_1 is the set of states of the world Θ and X_2 will be the set of signals. Mutual information will be the choice for the cost function for this work, it is therefore worth to return to a comparison with other possible alternatives. An alternative specification for the cost of a signal could be the difference in posterior and prior variance upon observing the signal. In the appendix we show what is the origin of the intuition that reduction in posterior variance should be a good candidate for a cost function of uncertainty. For normal variables,

[3] It does not satisfy symmetry.

[4] Remember that the two distributions share the same outcome space.

entropy coincides with a function of the variance, and mutual information therefore is exactly the log of the ratio of prior to posterior variance. So for normal variables the amount of reduction in posterior variance is mutual information. This does not apply to the discrete case. Armed with these tools we are now ready to examine the problem of the agent of choosing an optimal information structure under this cost.

1.4 Choice of the information structure: the Agent's problem

Now that we have established the shape of the cost function $C(\cdot)$ we can study the problem for the agent

$$\max_{(X,\mathbb{P}) and\ d(\cdot)} \mathbb{E}_{\mathbb{P}(X\times\Theta)}\left[\mathbb{E}_{d(x)}[u(W+b(d(x),\theta))] - \mu C(IN(X,\Theta))\right]$$

The agent can design any signal structure $(X,\mathbb{P})$ where X is a signal outcome space large enough, and $\mathbb{P}$ a joint distribution of the random variable with the state θ. The cost will be given by the mutual information between the signal X and the state of the world Θ. The choice of the function $d: X \to \Delta(D)$ from the signal space to the space of possible distributions on the decision space is costless.

To derive the optimal information structure of the agent first we argue that we can rephrase this problem in terms of choice of an experiment as defined by Blackwell and a deterministic mapping from signals to decisions. Then we further simplify the problem so that we can cast the choice of the agent directly in terms of probability distributions over the decisions conditional on the state. Lastly we show that the solution to the agent's problem is unique and derive conditions that need to hold at the optimum.

Definition 6. An experiment is a outcome space X and a $|X|\times|\Theta|$ Markov matrix $[p(x|\theta)]_{x,\theta}$. Where $p(\cdot|\theta)$ is the distribution over the set of signal outcomes X conditional on state θ.

We denote with $E(X)$ the set of experiments $m = [p(x|\theta)]_{x,\theta}$ whose available signals x belong to the set $X = 1, 2, \cdots, n$. The choice of the set of experiments $E(X)$ and the Markov matrix $m = [p(x|\theta)]_{x,\theta}$ determine the cost that the agent incurs: $I(m)$. Jointly, the two determine a joint probability distribution over $X\times\Theta$ associated with the Markov matrix and the prior. From such joint probability we can compute the total reduction in uncertainty brought about by the experiment. Because the mapping from the signal space X to the set of decisions does not affect the cost of the information structure, and the the expected utility is linear with respect to the probabilistic mapping $d(\cdot)$, we can also assume that the choice of the mapping from the signals to the decisions is deterministic so that for a signal realized x the choice is $d(x) \in D$ rather than a distribution over D.

Remark 7. For any given choice of an experiment - signal outcome set and Markov matrix $(X, [p(x|\theta)]_{x,\theta})$ - we have associated a posterior over states given by:

$$p(\theta|x) = \frac{p(x|\theta)\,p(\theta)}{\mathbb{E}_{p(\theta)}[p(x|\theta)]}$$

therefore the mutual information which represents the cost of our experiment, as shown in the previous section, is given by:

$$I(X, m) = \mathbb{E}_X\left(H\left(p(\cdot)\right)\right) - H\left(p(\cdot|x)\right)$$

So that the cost is represented by the average reduction in uncertainty weighted by the marginal over the signal outcomes. As a last point, because any signal x that does not occur with positive probability- i.e. such that the marginal over actions is zero, $\mathbb{E}_{p(\theta)}[p(x|\theta)] = p(d) = 0$ - does not affect the utility of the agent, we can without loss of generality restrict ourselves to considering only signal outcome spaces X and associated Markov experiments m where all x occur with positive probability.

The agent's problem has now been simplified to just choosing a signal outcome space whose points all occur with positive probability, a deterministic rule, and a Markov experiment in the class of experiments for such signal space:

Problem 8.

$$\max_{X, d(\cdot), m \in E(X)} \mathbb{E}_{\mathbb{P}(X \times \Theta)}[u(W + b(d(x), \theta))] - \mu I(m)$$

In the following lemmas we will work towards simplifying this problem further so that it reduces to the choice of a Markov matrix rather than a signal space and a Markov matrix over it. We start off by showing that the agent at the optimum will not choose a signal space with more signals than decisions available. We then construct an equivalent experiment signal space, decision rule and Markov matrix using the space of decisions D. Lastly we reframe the above problem in its most simple form.

Lemma 9. *Let $(X, m(), d())$ be an optimal experiment and decision rule for Problem 8, then d is injective.*

Proof. In Appendix 3.7. □

Then, because d is injective, it follows that at any solution to the agent's maximization problem, the set X of signal outcomes[5] has cardinality smaller than $|D|$.

Corollary 10. *If (X, m) and $d(\cdot)$ are a solution then $|D| \geq |X|$.*

We can now show the following Lemma that will allow us to rephrase the agent's problem in the simplest way.

[5]Defined above as those with positive probability of occurring.

Lemma 11. *The experiment and decision rule $((X,m), d(\cdot))$ is a solution to the above problem iff $((D,\widetilde{m}), id)$ is also a solution to the same problem.* [6]

$$\widetilde{p}(d'|\theta) = \begin{cases} p\left(d^{-1}(d')\,|\theta\right) & if\ d^{-1}(d') \in X \\ 0 & if\ d^{-1}(d') \notin X \end{cases}$$

Proof. By construction the Bernoulli utilities over wealth are the same. Moreover, for each $d' \in D$, we have that the posteriors coincide as well $\widetilde{p}(\cdot|d') = p\left(\cdot|d^{-1}(d')\right)$ so that $\mathbb{E}_{p(\theta)}\left[\widetilde{p}(d'|\theta)\right] = \mathbb{E}_{p(\theta)}\left[p\left(d^{-1}(d')\,|\theta\right)\right]$, and so $I(\widetilde{m}) = I(m)$. □

We are now ready to summarize the previous results and rephrase the problem as follows,

Proposition 12. *The agent's problem is equivalent to:*

$$\max_{m \in E(D)} \mathbb{E}_{p(\theta)}[\mathbb{E}_{p(d|\theta)}[u(W + b(d,\theta))]] - I(m) \tag{1.1}$$

We prove that a solution to the agent's problem exists and, unlike in our first formulation of the problem, it is unique. Of course for any unique solution to this problem we could still construct multiple valid experiments and decision rules that would be a solution to our first problem, but restricting to this class will allow us to use to our advantage the greater structure and the uniqueness of the solution

Lemma 13. *There exists a unique solution to the problem in equation 1.1.*

Proof. In Appendix 3.7. □

Proposition 14. *If m is a solution to the agent's problem in 1.1 then for any decision $d \in D$, $p(d|\theta)$ is such that either $p(d|\theta) = 0\ \forall\theta \in \Theta$ or $p(d|\theta) > 0$ for all θ and:*

$$u(W + b(d,\theta))\,p(\theta) - \mu p(\theta)\,log\left(\frac{p(d|\theta)}{p(d)}\right) = \mu p(\theta)\,log\left(\sum_{d' \in D} p(d')\,e^{\frac{u\left(W + b\left(d',\theta\right)\right)}{\mu}}\right)$$

Proof. Consider the m solution to problem 1.1, then if $p(d|\theta') > 0$ for at least some $\theta' \in \Theta\backslash\{\theta\}$, then $p(d) > 0$ and the marginal derivative of mutual information with respect to $p(d|\theta)$ is $p(\theta)\,log\left(\frac{p(d|\theta)}{p(d)}\right)$. So as $p(d|\theta) \to 0$, $\frac{\partial I}{\partial p(d|\theta)} = -\infty$. Therefore for any d, it cannot be that some of the $p(d|\theta)$ are 0 and some are not. Define $D' = \{d \in D | p(d|\theta) > 0$ for some and, by the previous argument all, $\theta\}$. Defining $m' = m_{|D'}$ as the restriction of our solution m to the decisions that are taken with positive probability, we know that m' is a solution to:

$$\max_{m \in E(D')} \mathbb{E}_{p(\theta)}[\mathbb{E}_{p(d|\theta)}[u(W + b(d,\theta))]] - I(m)$$

[6]The notation can be a little bit confusion. Here $d()$ is the decision rule for the original solution, and d' represents the the decisions/signals in the new experiment.

s.t.

$$\sum_{d' \in D'} p(d'|\theta) = 1 \; \forall \theta \in \Theta$$

$$0 \leq p(d|\theta) \leq 1 \forall d \in D', \theta \in \Theta$$

As a first observation we know from our previous discussion that for none of the d in D' the second set of constraints will bind. We can choose an open set O in $[0,1]^{|D'| \times |\Theta|}$ for which at the solution m' the objective and constraint function have continuous first derivatives in the set O. In checking the Karush-Kuhn-Tucker we only need to worry about the first set of constraints which have linearly independent gradients. Therefore the regularity conditions are satisfied and we can apply Karush-Kuhn-Tucker. The Lagrangian becomes:

$$\mathbb{E}_{p(\theta)}[\mathbb{E}_{p(d|\theta)}[u(W + b(d,\theta))]] - \mu I(m) + \sum_{\theta'} \lambda(\theta') \left(1 - \sum_{d' \in D'} p(d'|\theta)\right)$$

So we get:

$$u(W + b(d,\theta))\, p(\theta) - \mu \frac{\partial I(m)}{\partial p(d|\theta)} = \lambda(\theta) \;\; \forall d \in D' \forall \theta \in \Theta$$

$$u(W + b(d,\theta))\, p(\theta) - \mu p(\theta) \, log\left(\frac{p(d|\theta)}{p(d)}\right) = \lambda(\theta)$$

And:

$$p(d|\theta) = p(d)\, e^{\frac{u(W+b(d,\theta))}{\mu}} e^{-\frac{\lambda(\theta)}{\mu p(\theta)}}$$

Since the above equality holds for all d, we can sum over d to obtain:

$$1 = \sum_{d'} p(d'|\theta) = \sum_{d'} p(d')\, e^{\frac{u(W+b(d',\theta))}{\mu}} e^{-\frac{\lambda(\theta)}{\mu p(\theta)}}$$

So that

$$e^{\frac{\lambda(\theta)}{\mu p(\theta)}} = \sum_{d'} p(d')\, e^{\frac{u(W+b(d',\theta))}{\mu}}$$

obtaining as desired after substituting:

$$u(W + b(d,\theta))\, p(\theta) - \mu p(\theta) \, log\left(\frac{p(d|\theta)}{p(d)}\right) = \mu p(\theta) \, log\left(\sum_{d' \in D} p(d')\, e^{\frac{u(W+b(d',\theta))}{\mu}}\right)$$

and

$$p(d|\theta) = \frac{p(d)\, e^{\frac{u(W+b(d,\theta))}{\mu}}}{\sum_{d'} p(d')\, e^{\frac{u(W+b(d',\theta))}{\mu}}}$$

$\square$

1.5 The Principal's Problem

1.5.1 Main Result

Using the new formulation of the agent's problem, the principal's problem can now be rephrased in terms of the choice of an incentive scheme $b : D \times \Theta \to \mathbb{R}$ and a Markov matrix $m \in E(D)$. Moreover, the reduction to direct signals strategies for the agent is without loss of generality also for the problem faced by the principal.

From the old problem:

$$\max_{(X,\mathbb{P}) and\ d(\cdot),b(\cdot,\cdot)} \mathbb{E}_{\mathbb{P}(X\times\Theta)}\left[\mathbb{E}_{d(x)}[(y(d(x),\theta) - b(d(x),\theta))]\right]$$

s.t.

$$((X,\mathbb{P}),d(\cdot)) \in \arg\max \mathbb{E}_{\mathbb{P}(X\times\Theta)}\left[\mathbb{E}_{d(x)}[u(W + b(d(x),\theta))] - \mu C(IN(X,\Theta))\right]$$

And

$$b(d,\theta) \geq 0$$

We get:

Problem 15.

$$\max_{m\in E(D),b(\cdot,\cdot)} \mathbb{E}_{p(\theta)}[\mathbb{E}_{p(d|\theta)}[y(d,\theta) - b(d,\theta)]]$$

such that:

$$m \in \arg\max_{m} \mathbb{E}_{p(\theta)}[\mathbb{E}_{p(d|\theta)}[u(W + b(d,\theta))]] - I(m)$$

and

$$b(d,\theta) \geq 0$$

We can use Caplin and Dean [2013]'s theorem to characterize the solution to the agent's problem and use those conditions as constraints for the principal's maximization.

Remark 16. Caplin and Dean [2013] show that given the agent's information acquisition problem, an agent is rationally inattentive iff the solution $\left(B \subseteq D, [p(d)]_{d\in B}, [p(\theta|d)]_{\theta,d}\right)$ satisfies the following conditions

$$log\left(p\left(\theta|d\right)\right)-log\left(p\left(\theta\right)|d'\right)=\frac{u\left(W+b\left(d,\theta\right)\right)-u\left(W+b\left(d',\theta\right)\right)}{\mu}\ \forall d,d'\in B$$

$$\sum_{\theta}p\left(\theta|d\right)e^{\frac{u\left(W+b\left(d',\theta\right)\right)-u\left(W+b\left(d,\theta\right)\right)}{\mu}}\leq 1\ \forall d\in B,d'\in D\backslash B$$

Where $B=Supp\left(D\right)$ - i.e. the actions played with positive probability - and $\left[p\left(d\right)\right]_{d\in B},\left[p\left(\theta|d\right)\right]_{\theta,d}$ are respectively the marginal over decisions and the posterior conditional on decisions so that $\sum_d p\left(d\right)=1$, $\sum_\theta p\left(\theta|d\right)=1$, and they are consistent with the prior $\sum_d p\left(\theta|d\right)p\left(d\right)=p\left(\theta\right)$.

Theorem 17. *(Main Result: necessity) The optimal contract for problem 15 under the assumption that the agent has log utility or risk neutral preferences is:*

$$b\left(d,\theta\right)=\max\{0,K\cdot y\left(d,\theta\right)+B\left(\theta\right)+C\left(d\right)\}$$

with K=1 for risk neutral preferences and $K=\frac{1}{1+\mu}$ *for log utility.*

Proof. Let $\left(b\left(\cdot,\cdot\right),p\left(\cdot\right),\left[p\left(\cdot|d\right)\right]_{d\in D}\right)$ be a solution to the principal's problem. It is without loss to assume that $Supp\left(D\right)=D$, as otherwise we can set $b\left(d',\theta\right)=0\ \forall d'\in D\backslash Supp\left(D\right)$ and the Caplin and Dean [2013] conditions would still be satisfied as values of $p\left(\cdot|d\right)$ and p(d) for d in the $Supp\left(D\right)$ are independent of values of $b\left(d',\theta\right)$ for $d'\notin Supp\left(D\right)$. Unless the value of all $b\left(d,\theta\right)=0$ for all $d\in D$ and $\theta\in\Theta$, in which case the theorem is vacuously true, it is never possible that the inequality constraint binds as the agent utility is strictly increasing and only decisions for which there is positive support are taken.

So denoting our new set of decisions with full support B. The problem 15 of the principal can be rephrased expressing the maximization of the agent accordingly:

$$\max_{b(\cdot,\cdot),p(\cdot|\cdot),p()}\mathbb{E}_{p(\theta|d),p(d)}\left[y\left(d,\theta\right)-b\left(d,\theta\right)\right]$$

such that

$$b\left(d,\theta\right)\geq 0\ \ \left[\lambda\left(d,\theta\right)\right]$$

$$log\left(p\left(\theta|d\right)\right)-log\left(p\left(\theta|d'\right)\right)=\frac{u\left(W+b\left(d,\theta\right)\right)-u\left(W+b\left(d',\theta\right)\right)}{\mu}\ \forall\theta\in\Theta\forall d,d'\in B\ \ \left[\phi\left(\theta,d,d'\right)\right]$$

$$\sum_{\theta} p(\theta|d)\, e^{\frac{u(W+b(d',\theta))-u(W+b(d,\theta))}{\mu}} \leq 1\ \forall d \in B, d' \in D\backslash B\ [\checkmark]$$

$$0 \leq p(\theta|d) \leq 1 \forall d \in B, \theta \in \Theta\ [\checkmark]$$

$$0 \leq p(d) \leq 1 \forall d \in B\,[\checkmark]$$

$$\sum_{d} p(d) = 1\,[\alpha]$$

$$\sum_{\theta} p(\theta|d) = 1\ \forall d \in B\ [\tau(d)]$$

$$\sum_{d} p(\theta|d)\, p(d) = p(\theta)\, \forall \theta \in \Theta\, [\beta(\theta)]$$

By solving the relaxed program with full support we get rid of the three constraints marked with the checked item. Moreover it is easily checked that the constrained rank condition qualification (CRCQ) is satisfied: that is the rank is full for the matrix where columns consist of the gradients of the binding constraints in a neighborhood of the solution $(b, p(), p(\cdot|d))$ as part of the underlying euclidean space (given the finiteness of states and decisions). By KKT theorem then there exists a unique set of Lagrange multipliers vectors $[\lambda(d,\theta), \tau(d), \alpha, \beta(\theta), \phi(d,d',\theta)]$ such that $\lambda(d,\theta) \geq 0$ for all d and θ and such that it is $\lambda(d,\theta) = 0$ whenever $b(d,\theta) > 0$. Then:

$$\sum_{\theta} p(\theta|d)\,[y(d,\theta) - b(d,\theta)] + \alpha + \sum_{\theta} p(\theta|d)\,\beta(\theta) = 0\,[p(d)]$$

$$p(d)\,[y(d,\theta) - b(d,\theta)] + \sum_{d'} \phi(\theta,d,d')\,\frac{1}{p(\theta|d)} + \tau(d) + p(d)\,\beta(\theta)\ [p(\theta|d)]$$

$$-p(\theta|d)\,p(d) + \lambda(d,\theta) - \sum_{d'} \phi(d,d',\theta)\,\frac{u'(W + b(d,\theta))}{\mu} = 0\,[b(d,\theta)]$$

Solving for $\phi(d,d',\theta)$ from the last equation for $b(d,\theta)$ we get:

$$\sum_{d'} \phi(d,d',\theta) = -\frac{\mu p(\theta|d)\,p(d)}{u'(W + b(d,\theta))} + \frac{\mu\lambda(d,\theta)}{u'(W + b(d,\theta))}$$

substituting into the second equation for $p(\theta|d)$ and dividing through by $p(d)$:

$$[y(d,\theta)-b(d,\theta)]-\frac{\mu}{u'(W+b(d,\theta))}+\frac{\mu\lambda(d,\theta)}{u'(W+b(d,\theta))\,p(d)}\cdot\frac{1}{p(\theta|d)}+\frac{\tau(d)}{p(d)}+\beta(\theta)$$

Rearranging, and using the property that if $\lambda(d,\theta)>0$ then $b(d,\theta)=0$, and this implies that $y(d,\theta)+\frac{\tau(d)}{p(d)}+\beta(\theta)-\frac{\mu}{u'(W+b(d,\theta))}$ must then be negative when $\lambda(d,\theta)>0$, we get for all $\theta\in\Theta$ and all $d\in B$:

$$b(d,\theta)=max\left\{0,y(d,\theta)+\frac{\tau(d)}{p(d)}+\beta(\theta)-\frac{\mu}{u'(W+b(d,\theta))}\right\}$$

Because of our initial argument on the values of $b(d,\theta)$ outside B we have that actually for all $d\in D$ and all $\theta\in\Theta$:

$$b(d,\theta)=max\left\{0,y(d,\theta)+\frac{\tau(d)}{p(d)}+\beta(\theta)-\frac{\mu}{u'(W+b(d,\theta))}\right\}$$

For log utility and risk neutral preferences this yields:

$$b(d,\theta)=\max\{0,K\cdot y(d,\theta)+B(\theta)+C(d)\}$$

with A=1 for risk neutral preferences and $K=\frac{1}{1+\mu}$ for log utility. □

In the appendix we present an alternative proof that does not use Caplin and Dean [2013]'s theorem. The proof, derived using the result from proposition 14 as a constraint, also characterizes explicitly the state and decision payments. The optimal contract has a linear structure, the explicit dependence on output is linear and state and decision payments are separable. Output affects indirectly the fixed payments, so a change of the whole output function $y(d,\theta)$ leads to a different contract in terms of fixed payments[7]. Nevertheless, for any fixed output function, output affects linearly the incentives across decisions and states.

We now wonder how tight the result is; whether the separability of state and decision payments is true only for entropy as a measure of uncertainty. If that was the case this would allow us to characterize entropy in terms of this property for agency contracts with observable states and decisions.

Following our definition of mutual information and Caplin and Dean [2013]'s definition of a posterior-separable attention cost function, we let

$$I^g(\pi,g,p())=\mu\left(\sum_{\theta\in\Theta}\pi(\theta)\,g(\pi(\theta))-\sum_{d}p(d)\sum_{\theta\in\Theta}p(\theta|d)\,g(p(\theta|d))\right)$$

[7]For example an output function with very high values for all decisions and states might lead the principal to correct the incentive by having negative payments in all states.

and such that $\lim_{p(\theta|d)\to 0^+}\frac{\partial\sum_{\theta\in\Theta}p(\theta|d)g(p(\theta|d))}{\partial p(\theta'|d)}=-\infty$ and $\lim_{p(\theta|d)\to 1^-}\frac{\partial\sum_{\theta\in\Theta}p(\theta|d)g(p(\theta|d))}{\partial p(\theta'|d)}=\infty$

Where $\sum_{\theta\in\Theta}p(\theta|d)\,g(p(\theta|d))$ maps from $\mathbb{R}_+^{|\Theta|}$ to $\mathbb{R}$.

Theorem 18. *Under the attention cost $I^g(\pi,g,p())$, the optimal contract is:*

$$b(d,\theta)=\max\{0,y(d,\theta)+B(\theta)+C(d)-\frac{\mu}{u'(W+b(d,\theta))}\}$$

if and only if

$$g(p,d,\theta)=-log(p)+\frac{t}{p}$$

Where t is a constant. So that $\sum_{\theta\in\Theta}p(\theta|d)\,g(p(\theta|d))=t-\sum_{\theta\in\Theta}p(\theta|d)\,ln(p(\theta|d))=H(p(|d))+t$. That is for the contract to be separable in state and decision payments, the cost function must be entropy based mutual information.

Proof. By Lemma 3 in Caplin and Dean [2013], $\left([p(\cdot)]_{d\in D},[p(\theta|d)]_{\theta\in\Theta,d\in D}\right)$ is a maximizer for the agent's problem if and only if :

for some state $\overline{\theta}$ and $\forall\ d,d'\in Supp(D)$

$$p(\overline{\theta}|d)\,u(W+b(d,\overline{\theta}))+\mu p(\overline{\theta}|d)\,g(p(\overline{\theta}|d))-\mu\sum_{\theta'\neq\overline{\theta}}[p(\theta'|d)]^2\,g'(p(\theta|d))=$$

$$p(\overline{\theta}|d')\,u(W+b(d',\overline{\theta}))+\mu p(\overline{\theta}|d')\,g(p(\overline{\theta}|d'))-\mu\sum_{\theta'\neq\overline{\theta}}[p(\theta'|d')]^2\,g'(p(\theta|d'))$$

$\forall\theta\in\Theta$ and $d,d'\in Supp(D)$:

$$\mu p(\theta|d')\,g'(p(\theta|d'))-\mu g(p(\theta|d))-\mu p(\theta|d)\,g'(p(\theta|d))+\mu g(p(\theta|d'))=u(W+b(d,\theta))-u(W+b(d',\theta))$$

and using the notation for net utilities from Caplin and Dean [2013]:

$$N(p(\cdot|d))=\sum_{\theta}p(\theta|d)\,(u(W+b(d,\theta))+\mu g(p(\theta|d)))$$

we have that for some $\overline{\theta}\in\Theta$, and all $d\in Supp(D),d'\in D\backslash Supp(D)$:

$$N\left(p\left(\cdot|d'\right)\right)-\sum_{\theta\neq\overline{\theta}}\frac{\partial N\left(p\left(\cdot|d\right)\right)}{\partial p\left(\theta|d\right)}p\left(\theta|d'\right)\leq N\left(p\left(\cdot|d\right)\right)-\sum_{\theta\neq\overline{\theta}}\frac{\partial N\left(p\left(\cdot|d\right)\right)}{\partial p\left(\theta|d\right)}p\left(\theta|d\right)$$

where $p\left(\cdot|d'\right)$ is chosen such that it maximizes: $N\left(p\left(\cdot|d'\right)\right)-\sum_{\theta\neq\overline{\theta}}\frac{\partial N\left(p\left(\cdot|d\right)\right)}{\partial p\left(\theta|d\right)}p\left(\theta|d'\right)$.

Let $\left(b\left(\cdot,\cdot\right),p\left(\cdot\right),p\left(\cdot|d\right)\right)$ be a solution for the principal's problem. Since the principal has the authority to restrict the agent's available decisions it is without loss to assume $Supp\left(p\left(\cdot\right)\right)=D$, then by the Inada conditions above $Supp\left(p\left(|d\right)\right)=\Theta$ for all d$\in D$. And the solution $\left(b\left(\cdot,\cdot\right),p\left(\cdot\right),p\left(\cdot|d\right)\right)$ is for:

$$\max_{\left(b\left(\cdot,\cdot\right),p\left(\cdot\right),p\left(\cdot|d\right)\right)}\sum_{d}p\left(d\right)\sum_{\theta}p\left(\theta|d\right)\left[y\left(d,\theta\right)-b\left(d,\theta\right)\right]$$

subject to :

$$b\left(d,\theta\right)\geq 0\left[\lambda\left(d,\theta\right)\right]$$

$$p\left(\overline{\theta}|d\right)u\left(W+b\left(d,\overline{\theta}\right)\right)+\mu p\left(\overline{\theta}|d\right)g\left(p\left(\overline{\theta}|d\right)\right)-\mu\sum_{\theta'\neq\overline{\theta}}\left[p\left(\theta'|d\right)\right]^{2}g'\left(p\left(\theta|d\right)\right)=$$

$$p\left(\overline{\theta}|d'\right)u\left(W+b\left(d',\overline{\theta}\right)\right)+\mu p\left(\overline{\theta}|d'\right)g\left(p\left(\overline{\theta}|d'\right)\right)-\mu\sum_{\theta'\neq\overline{\theta}}\left[p\left(\theta'|d'\right)\right]^{2}g'\left(p\left(\theta|d'\right)\right)\ \left[\psi\left(d,d'\right)\right]$$

$$p\left(\theta|d'\right)g'\left(p\left(\theta|d'\right)\right)-g\left(p\left(\theta|d\right)\right)-p\left(\theta|d\right)g'\left(p\left(\theta|d\right)\right)+g\left(p\left(\theta|d'\right)\right)=$$

$$\frac{u\left(W+b\left(d,\theta\right)\right)-u\left(W+b\left(d',\theta\right)\right)}{\mu}\ \left[\phi\left(d,d',\theta\right)\right]$$

$$\sum_{d}p\left(d\right)=1\left[\alpha\right]$$

$$\sum_{\theta}p\left(\theta|d\right)=1\left[\tau\left(d\right)\right]$$

$$\sum_{d}p\left(d\right)p\left(\theta|d\right)=1\left[\beta\left(\theta\right)\right]$$

Substituting a well behaved function $g()$ it is possible to check that the matrix whose columns consist of the gradients of binding constraints has constant rank in a neighborhood of $\left(b\left(\cdot,\cdot\right),p\left(\cdot\right),p\left(\cdot|d\right)\right)$.

Therefore the KKT necessary conditions hold and imply that there exists unique Lagrange multipliers $\lambda(d,\theta), \psi(d,d'), \phi(d,d',\theta), \alpha, \tau(d), \beta(\theta)$ such that $\forall\theta \neq \overline{\theta}$ and each $d \in D$:

$$\sum_{\theta} p(\theta|d)\left[y(d,\theta) - b(d,\theta)\right] + \alpha + \sum_{\theta} \beta(\theta)\, p(\theta|d) = 0 \;\; [p(d)]$$

$$p(d)\left[y(d,\theta) - b(d,\theta)\right] - \mu \sum_{d'} \psi(d,d') \left(2p(\theta|d)\, g'(p(\theta|d)) + p(\theta|d)^2 g''(p(\theta|d))\right)$$

$$-\sum_{d'} \phi(d,d',\theta)\left(2g'(p(\theta|d)) + p(\theta|d)\, g''(p(\theta|d))\right) + \tau(d) + p(d)\,\beta(\theta) = 0 \;\; [p(\theta|d)]$$

$$-p(d)\, p(\theta|d) + \lambda(d,\theta) - \sum_{d'} \frac{\phi(d,d',\theta)\, u'(\overline{W} + b(d,\theta))}{\mu} = 0 \;\; [b(d,\theta)]$$

We can rearrange equation $[b(d,\theta)]$ as

$$-\frac{\mu p(d)\, p(\theta|d)}{u'(W + b(d,\theta))} + \frac{\mu\lambda(d,\theta)}{u'(W + b(d,\theta))} = \sum_{d'} \phi(d,d',\theta)$$

and substituting into the equation derived from $[p(\theta|d)]$, we get:

$$p(d)\left[y(d,\theta) - b(d,\theta)\right] - \mu \sum_{d'} \psi(d,d') \left(2p(\theta|d)\, g'(p(\theta|d)) + p(\theta|d)^2 g''(p(\theta|d))\right)$$

$$+\left(\frac{\mu p(d)\, p(\theta|d)}{u'(W + b(d,\theta))} - \frac{\mu\lambda(d,\theta)}{u'(W + b(d,\theta))}\right)\left(2g'(p(\theta|d)) + p(\theta|d)\, g''(p(\theta|d))\right) + \tau(d) + p(d)\,\beta(\theta) = 0$$

Simplifying:

$$b(d,\theta) = y(d,\theta) + \mu\left(2p(\theta|d)\, g'(p(\theta|d)) + p(\theta|d)^2 g''(p(\theta|d))\right) \cdot \left(\frac{1}{u'(W + b(d,\theta))} - \frac{\sum_{d'} \psi(d,d')}{p(d)}\right)$$

$$-\frac{\mu\lambda(d,\theta)}{u'(W + b(d,\theta))\, p(d)}\left(2g'(p(\theta|d)) + p(\theta|d)\, g''(p(\theta|d))\right) + \frac{\tau(d)}{p(d)} + \beta(\theta)$$

So that we will have separable state and action payments iff:

$$\left(2p(\theta|d)\, g'(p(\theta|d)) + p(\theta|d)^2 g''(p(\theta|d))\right) = -1 \text{ for } 0 < p(\theta|d) < 1$$

This is a second order differential equation whose solution is:

$$g(p; d, \theta) = -log(p) + \frac{t}{p}$$

$\square$

1.5.2 Discussion

Cost of information and comparative statics

Interestingly, we can perform comparative statics on the cost of acquiring information (μ). The effect is not obvious. A higher attention cost could imply that higher incentives are needed for the agent to acquire enough information. Similarly, an agent that has a small cost for acquiring information might not need strong incentives to do so. On the other hand, information acquisition will be more expensive and less effective for an incapable agent. The optimal contract, for the log utility case, gives less incentives to acquire information to an incompetent agent. The coefficient on output is $K = \frac{1}{1+\mu}$, so that highly competent agents are given high powered incentives while the opposite happens for less competent ones. Importantly, the strength of such incentives is monotonic in the cost parameter μ. The reason is both that incentives are more effective for low-cost agents, and that the higher incentives make the risk averse agents more likely to make the right choice. Intuitively, incentivizing more a risk-averse and competent agent makes him more willing to acquire information because rewards and risks are higher. The increase in risk caused by the steeper incentives is counterbalanced by his higher success rate in making the optimal choice thanks to the more extensive information collected and his competency. For an incompetent agent, this second effect does not kick in and therefore the principal is better off minimizing the extra risk that is passed to the agent.

Contract Complexity

Another important observation is the separability of the state and decision payments. This determines a significant reduction in the complexity of the contract. A generic contract $D(d, \theta)$ needs to specify one payment for each possible combination of states and decisions. Such contracts have a dimensionality of $|\Theta| \times |X|$ which can be daunting especially in circumstances where the number of states realized and possible decisions is large. The contract we derive only needs to specify $|\Theta| + |X| + 1$ terms: the state and decision payments and a coefficient on output. The reduction in complexity to write and understand such contracts is notable. Of course an important assumption in this discussion is that it is easy or costless for the agent to understand the output production function. This reduction in complexity is less compelling in situations where the agent does not know what the effect on output of different decisions and states is. The least-complexity contract would be an affine contract on output: $b(d, \theta) = max\{0, Ky(d, \theta) + A\}$. Although such a contract would be appealing from a complexity reduction point of view we next examine why such contract is not optimal. In section (1.7) we also inquire what are the welfare losses associated with the simplest linear contract versus the optimal one, and how well the reduced complexity optimal contract fares relative to the unrestricted one for different agent utilities.

State and decision payments

The simplest linear contract provides some incentives to acquire information. In the previous section we saw how this feature is tightly linked to entropy. We present here some simple counterexamples pointing to the need for state and decision payments and discuss where the separability comes from. The matrix below illustrates a principal-agent risk neutral problem with two decisions and three states. The principal would want to incentivize the agent to acquire information about states θ_1 and θ_3 in order to take respectively decision d_1 and d_3. There is another state of the world θ_2 where output is very high for the principal and where she has not any interest in acquiring information.

$y(d,\theta)$	θ_1	θ_2	θ_3
d_1	1	100	0
d_2	0	100	1

Table 1.3: Necessity of state payments

If the principal were restricted to using an affine contract of the form $b(d,\theta) = max\{0, Ky(d,\theta) + A\}$, he would face the dilemma of either providing a high coefficient on output (K) to incentivize information acquisition in states θ_1 and θ_2 while paying an unnecessarily large sum of money in state θ_2 or providing too weak incentives to acquire information in the states θ_1 and θ_2. We can improve upon this contract by letting A be a function of the state and giving the agent large negative values for $A(\theta_2)$ to compensate for the higher coefficient on output. By letting A vary with the state we allow the principal to hit the limited liability constraint in any state for at least one decision while providing insurance to the possibly risk-averse agent across states.

Consider now an identical setting with the following output function.

$y(d,\theta)$	θ_1	θ_2	θ_3
d_1	0	0	1
d_2	-100	1	0

Table 1.4: Necessity of action payments

Here the principal would want the agent to acquire information about states θ_2 and θ_3, so that actions d_1 and d_2 are chosen with positive probability. But notice that action d_2 is extremely risky for the principal and the limited liability constraint means that the principal cannot pass any part of that risk to the agent[8]. If the agent were to have a purely affine contract, $b(d,\theta) = max\{0, Ky(d,\theta) + A\}$, he would not bear such costs and he would overplay decision d_2. The principal could increase the fixed payment part of the contract (A), so that the linear incentive K would have some bite for negative values of output, but this would be costly. It can be more efficient to offer positive and negative rewards for the two decisions separately to tilt the decision of the agent towards decision d_1 while still allowing decision d_2 if the agent has high confidence that the state is θ_3.

These example matrixes seem to imply that this property is true in general so that action and state fixed effects should be separable for any cost function. This is not the case. The examples are purposefully build to isolate the two effects. In a general output function these two problems will be interconnected an an unrestricted contract $D(d,\theta)$ for an arbitrary cost function would not be separable in decision and state payments. It is the separability of entropy with respect to posteriors and the use of log as a $g()$ function in our previous section that allows to separate state and decision incentives. In addition as we will remark in the extensions section this feature only depends on the cost function and can be extended for any arbitrary agent's function.

Relation to the literature

It is interesting to note that the contract is similar to the one obtained by Carroll [2013] in a different setup. Carroll [2013] assumes that the state and decision are observable and contractable but he considers a risk neutral agent whose cost and information acquisition technology are unknown to the principal. The principal has infinite risk aversion in the form of max min preferences over the set of possible information acquisition technologies. The resulting contract is a restricted linear structure contract with state dependent payments. The contract is restricted in that the principal does not allow the agent to take all actions and the payment to the agent is given by:

$$b^{GC} = \max\{0, \alpha y(d,\theta) + B(\theta)\}$$

The origin of the linearity and fixed state payments is nevertheless different. While in Carroll [2013] the linear structure is a result of the desire for robustness stemming from the max min preferences and the unknown set of the information production technologies (see also Carroll [2015]), in our setting the separation of the

[8]Beyond a zero payoff.

state and decision payments is a result of the cost of information acquisition being entropy and therefore it allows us to assume risk aversion for the agent as well.

Applications

Obviously, this contract is not applicable to all information acquisition agency problems. In many instances the state and/or decisions taken by the agents are not observable and so they cannot be conditioned on. Consider, for instance, a doctor deciding whether to implement a type of preventive of surgery. If the treatment is not implemented we can learn the state of the world by observing subsequent health problems, but if the treatment is carried out we do not learn about the counterfactual state of the world. Nevertheless there are many domains of information acquisition agency problems where the assumption of observability and contractibility of the states and decisions is not far fetched. For instance in the Asset Management industry, we do observe the realized prices and decisions taken by an investment manager. Similarly we can observe realized states for economic forecasts or business predictions. The incentives provided to portfolio managers and proprietary traders in the hedge fund industry do resemble the optimal contract derived here. According to our conversations with traders in the industry, Portfolio Managers are offered a fraction of their profits (usually around 12.5%) in addition to a base salary and a bonus dependent on the performance of the firm and the evaluation of their specific actions. Such a contract would be consistent with log utility and significant costs of information acquisition.

Another interpretation of the contract is in terms of linear structure with respect to benchmarks. From theorem (32) by taking differences over positive incentive plans across actions and states we have the following result.

Remark 19. Assuming $b(d,\theta) > 0, b\left(d,\widehat{\theta}\right) >, b(d',\theta) > 0$ we have that the optimal contract is such that:

$$b(d,\theta) - b(d',\theta) = K\left(y(d,\theta) - y(d',\theta)\right) + C(d) - C(d')$$

$$b(d,\theta) - b\left(d,\widehat{\theta}\right) = K\left(y(d,\theta) - y\left(d,\widehat{\theta}\right)\right) + B(\theta) - B\left(\widehat{\theta}\right)$$

So that the difference in the contract payoff between any two actions is simply a linear function of the differences between the payoffs to the principal plus a constant. $\Delta_{d,d'}b(\theta) = K\Delta y_{d,d'}(\theta) + C$. This allows an easy interpretation in terms of financial benchmarks. The optimal contract is such that the incentive to the agent for choosing a financial strategy as opposed to the market benchmark (e.g. S&P500) is a linear function, for any state realized, of the difference in performance between his strategy and the S&P500. So after the agent has chosen a portfolio his payment relative to what he would have got had he chosen the benchmark is just a fraction of the difference in performance between the two for all states realized. A similar interpretation can be done for the difference in states rather than across actions. This rearranging highlights that this contracts is consistent with a linear payment with respect to a benchmark - usually the market portfolio. Such incentives tied to a benchmark are particularly common in long-only asset management and money markets.

1.6 Extensions

One might wonder how robust these results are to the specific assumptions we made about the utility function and the lack of other constraints for the agent and principal. In this section we re-examine the results of the main theorem to establish their validity outside the original context. In the process, we also clarify that the utility function plays no role in the separability of the incentives into state and decision specific payments.

In the following we consider how robust the model is for arbitrary utilities, partial revelation of states and decisions, participation constraint for the agent, and wealth constraints for the principal.

Arbitrary Utility

Our result in theorem (32) holds for risk neutral and log utility agents but as a corollary from the proof in the previous section we obtain:

Corollary 20. *The optimal contract for the Principal-Agent in problem (15) is given by:*

$$b\left(d,\theta\right)=\max\{0,y\left(d,\theta\right)+B\left(\theta\right)+C\left(d\right)-\mu\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}\}$$

This result is a direct corollary to Theorem (32) using an arbitrary differentiable utility u instead of the log or risk neutral specification. As a first observation notice that the incentives are still separable per action and state so that, if we assume that the agent knows his utility, the dimensionality of the contract is still $|\Theta|+|D|+1$ rather than $|\Theta|\times|D|$. Nevertheless this interpretation hides some of the problems with such contract. Although the optimal contract only needs to specify $|\Theta|+|D|+1$ parameters, it requires the agent being able to compute and solve a potentially complex fixed point problem. So the reduction in complexity coming from the lower dimensionality of the contract might be more than outweighed by the computational complexities of solving for the implicit contract. This naturally begs the question of how close to the optimal contract is an incentive scheme with the simple structure of theorem 32[9]. We return to this question in later sections.

[9] That is a contract whose complexity is only given by a slope parameter on output (K) and a schedule of state and decision payments:

$$b\left(d,\theta\right)=max\{0,Ky\left(d,\theta\right)+B(\theta)+C(d)\}$$

Generalized cost function

A possible criticism of mutual information is that it is costless to aggregate information so as to ignore non payoff relevant features. For instance, Woodford [2012] notices that Sims [2003]'s model of rational inattention cannot rationalize the experimental results in Shaw and Shaw [1977]. Shaw and Shaw [1977] present an experiment where subjects need to recognize one of three letters {E,T,V} shown for some milliseconds in a location at random around a circle. The testers know the true distribution over locations and letters. The agents receive the same monetary prize for guessing the correct letter independent of the location and letter. Shaw and Shaw [1977] show that when the distribution over locations in the circle is uniform the agents have indistinguishable error rates across different locations - just like Sims [2003]'s rational inattention would predict. In contrast, when the distribution is not uniform the error rates for different locations are statistically different. This piece of evidence is inconsistent with Sims [2003]'s model. According to rational inattention the agent should be able to ignore the location dimension and just acquire information about the letter. This experimental violation of rational inattention could be troublesome also for the robustness of our results. A way to generalize the cost function, so as to make it consistent with the experimental evidence, is to allow the cost of acquiring information to depend directly on the state of the world. In the mutual information section, we derived how for a distribution $p \in \Delta(\Theta)$, the minimum expected number of signals needed to communicate the state was between $H(p)$ and $H(p)+1$. To allow the cost of information acquisition to depend on the state we can introduce a state specific chance of communication disruption. Let $\alpha(\theta) > 0$ be the expected number of times that a message should be transmitted to be received in state θ. Accordingly, we get a modified entropy formula:

$$H^{\alpha}(p) = -\mathbb{E}_p\left[\alpha(\theta)\, log\left(p(\theta)\right)\right] \forall p \in \Delta(\Theta)$$

With one parameter per state to choose we can now rationalize the shape of the error rates described in Shaw and Shaw [1977]. Moreover, for the agent problem assuming that the mutual information is now defined as $I^{\alpha}(m) = \left(H^{\alpha}(p) - H^{\alpha}(p(\cdot|d))\right)$ we get the following result:

Corollary 21. *The optimal contract assuming that the cost function is given by I^{α} is:*

$$b(d,\theta) = \max\{0, y(d,\theta) + \widehat{B}(\theta) + \mu\alpha(\theta)\widehat{C}_1(d) + \widehat{C}_2(d)\}$$

when the agent has risk neutral preferences and,

$$b(d,\theta) = \max\{0, \frac{y(d,\theta)}{1+\mu\alpha(\theta)} + \widehat{B}(\theta) + \frac{\mu\alpha(\theta)}{1+\mu\alpha(\theta)}\widehat{C}_1(d) + \widehat{C}_2(d)\}$$

If the agent has log utility

Proof. In Appendix 3.7. □

Notice that to accommodate the generalized cost function that would rationalize the experiment by Shaw and Shaw [1977], we lose the separability of state and decision payments as we have to depart from entropy based cost functions.

Participation constraint

The optimal contract is virtually unchanged by the requirement that the principal satisfies a participation constraint for the agent. The state and decision specific constant will change accordingly but the structure is identical.

Proposition 22. *The optimal contract for Problem (15) under a participation constraint for the principal is:*

$$b(d,\theta) = \max\{0, Ay(d,\theta) + B'(\theta) + C'(d)\}$$

where $A=1$ and $A=\frac{1}{1+\mu}$ for risk neutral and log utility respectively. Similarly for the problem with arbitrary utility $u()$, the optimal contract is $b(d,\theta) = \max\{0, y(d,\theta) + B'(\theta) + C'(d) - \mu\frac{1}{u'(W+b(d,\theta))}\}$.

Proof. In Appendix 3.7. □

Wealth constraint for the principal

Similarly we can analyze the result of assuming that the there is a wealth constraint for the Principal that limits the payments that it can make to the agent. Assume that the principal still is risk neutral and has state and decision contingent positive wealth $w(i,\theta) \geq 0$.

Proposition 23. *The optimal contract for the Principal agent problem (15)under limited liability and wealth constraints for the principal is:*

$$b(d,\theta) = \max\{0, min\{y(d,\theta) + B(\theta) + C(d) - \mu\frac{1}{u'(W+b(d,\theta))}, w(d,\theta)\}\}$$

For the log case and risk neutrality this becomes respectively:

$$b(d,\theta) = \max\{0, min\{Ky(d,\theta) + B(\theta) + C(d), w(d,\theta)\}\},$$

with $K = \frac{1}{1+\mu}$ for log utility and $K = 1$ for risk neutral preferences.

Proof. In Appendix 3.7. □

Partially revealing states and decisions

Another important assumption of our theorem has been the fact that both the state and the decision are assumed observable and contractible. Zermeno [2011]'s paper studies the optimal contract for a two state decision making problem where the agent has risk neutral utility and where the state and decision are not always revealed.

An example of a situation in which his setting is particularly compelling is that of a doctor-patient relationship deciding whether to carry out a preventive surgery:

$y(d,\theta)$	Necessary	Unnecessary
Perform surgery	0	0
Don't Perform	-10	1

Table 1.5: Patient's payoffs

In this situation we only learn about whether the preventive surgery was necessary if the doctor decides not to treat the patient. Zermeno [2011]'s point is that the optimal contract in such situation might optimally bias the agent towards inefficient decision making relative to the information available. In particular, Zermeno [2011] shows that the optimal decision rule is tilted towards decisions that reveal more information (in our case not performing the surgery).

Definition 24. Let $\Theta(d)$ denote a partition of the states corresponding to decision d. And $\theta(d)$, the element in the partition $\Theta(d)$ that contains θ.

Then the following result holds:

Proposition 25. *Suppose for all $d \in D$ $y(d,\theta) = y(d,\theta')$ for all θ and $\theta' \in \theta(d)$ and that it is required for the feasibility of the contract that for all $d \in D$ $b(d,\theta) = b(d,\theta')$ for all θ and $\theta' \in \theta(d)$. Then the optimal contract has the form:*

$$b(d,\theta) = \max\{0, Ky(d,\theta) + B(\theta(d)) + C(d)\}$$

Proof. In Appendix .

□

1.7 Robustness and Simulations

We showed that if the agent has limited liability and risk neutral or log-utility preferences, the optimal contract has a linear structure with respect to output.

$$b(d,\theta) = max\{0, Ky(d,\theta) + B(\theta) + C(d)\}$$

A natural question is how well this contract would fare for different utilities. Given the prevalence of linear contracts, we also wonder if the contractual complications introduced by state and decision payments are justified by the optimal contract performance relative to the purely linear one.

1.7.1 Monte Carlo Simulation

To address these concerns we resort to Monte Carlo numerical simulations in MATLAB. We build an architecture of functions and methods to evaluate what is the surplus to the principal of various contracts under changing utilities, costs or parameters.

To construct the measure of the surplus we start by the observation that the value to the principal of a contract b^* will depend on the utility function of the agent ($u()$) which will determine how effective the incentives are, as well as the shape of the cost function ($C()$), the output function $y(\cdot,\cdot)$, and other parameters of the problem ($\alpha = \{\mu, p..etc\}$). We can express the value for the principal of a contract b^* as the utils obtained by the principal under such contract $V^P(U(), C(), y(\cdot,\cdot), \alpha, b^*)$. We can then compare the value of different contracts using the same functional forms and parameters. The risk neutrality of the principal ensures that we can interpret these utils as dollars, but the comparison of welfare across different contracts would not be particularly illuminating absent a benchmark from which to reference the gains of each contract type. As a benchmark, we choose the welfare that the principal would have obtained not incentivizing any information acquisition and requesting the agent to pick the ex-ante optimal decision. We denote such contract[10] with b^{EA}. For a given functional form $u()$, cost function $C()$, output matrix $y(\cdot,\cdot)$ and parameters α, we have that the gain of contract b^* is:

$$G_{b^*}(u(), C(), y(\cdot,\cdot), \alpha) = V^P(u(), C(), y(\cdot,\cdot), \alpha, b^*) - V^P\left(u(), C(), y(\cdot,\cdot), \alpha, b^{EA}\right)$$

Before describing the results and the algorithms we use to obtain them, it is helpful to have a broad overview of how our Monte Carlo simulation works. We select random matrixes $y(\cdot,\cdot)$ from a distribution (uniform here), for each of these matrixes we ran a loop of problems changing some parameters[11]; for each such problem

[10] For instance, such a contract would specify a positive infinitesimal incentive for the ex ante optimal decision on all states and zero dollars in any other decision.

[11] For example, here we will vary the agent's CRRA Utility functions through the CRRA coefficient.

we ran a separate optimization for each particular contract and compute the optimizing ex ante action and associated value to the principal. We store each of these welfare gains, for each contract, within the loop for each utility function, within the Monte Carlo simulated random matrixes. A graphical illustration is provided below.

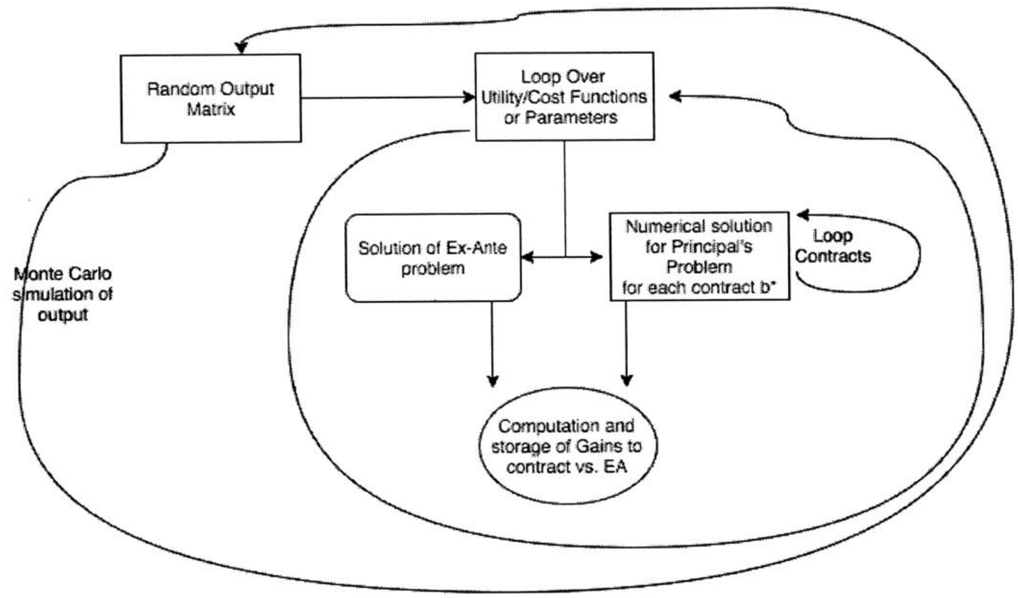


Figure 1-1: Graphical Overview of the simulation

The numerical solution to the principal's problem for each contract is the important piece. We reframe the optimization problem as one where the principal chooses simultaneously the matrix of conditional probabilities $[p(d|\theta)]_{d,\theta}$ and the contract with the agent $[b(d,\theta)]_{d,\theta}$ subject to the first order conditions of the agent's optimization. This poses a significant challenge for the numerical optimization even for a relatively low dimensional problem (e.g. three actions and six states). The difficulties arise because of the $N \cdot M$ non linear constraints that the algorithm needs to evaluate in addition to $2(N \cdot M) + M$ linear constraints. Linear constraints are fast to evaluate numerically within MATLAB whereas nonlinear constraints are quite daunting. To solve this, we rewrite all the $N \cdot M$ first order conditions for the agent's problem as a unique vectorized matrix. This allows us to have a scalable algorithm that can be used to solve problems of arbitrary size for N actions and M states and to reduce the computing time for the solution by a factor of 40.

We apply the `fmincon` local optimization package within MATLAB using, after several tests with `sqp`, the interior-point algorithm. Before implementing our maximization using this faster local optimizer, we test using `gsearch` and the state of the art global maximizer `knitro` that the solutions found by fmincon coincide with the global ones. Ultimately we choose `fmincon` over `knitro` because of the faster running time and because the server on which we can more efficiently run this problem parallelized does not have a `knitro` license[12]. Lastly to improve the performance of the algorithm we parallelize the jobs at the level of the random output matrix loops. The computing time is reduced by a factor of the number of cores used.

1.7.2 Results

In the following we study how robust the optimal contract is to changes in the utility function within the space of CRRA utilities. For a set of randomly generated matrixes and for a sequence of CRRA utility functions, we compare three contracts: the unrestricted optimal contract ($b\,(d,\theta)$), the linear contract with state and decision payments that is optimal for log and risk neutral preferences ($b\,(d,\theta) = max\{0, Ky\,(d,\theta) + B(\theta) + C(d)\}$), and a simple linear contract ($b\,(d,\theta) = max\{0, A + Ky(d,\theta)\}$). For each of these we compute the surplus and graph their distribution with respect to the random output matrixes.

We present three graphs to illustrate the following three points from the simulations: the surplus of the purely linear contract can be negative and in general quite far from the surplus of both the unrestricted contract and the state and decision payments linear contract, this negative surplus is more likely to occur when the agent risk aversion is higher, the purely linear contract can perform quite well relative to the action and state payments contract for low CRRA-coefficients.

The simulations presented here have been obtained using the following parameters:

Parameter	Value
Utility	CRRA
CRRA-coefficient	0.142,1(log),1.42
Wealth of the Agent	1
Limited Liability	Yes
Cost Of Attention (μ)	1
Outside Participation Constraint	0
Distribution of the Output matrix	3by6,Non symmetric Uniform$[0,1]$ over a symmetric base
Cost Function	Entropy

Table 1.6: Simulations Parameters

The first of the following graphs plots the distribution for the three contracts for different random output matrixes when the agent has risk aversion around 0.142. There are three observations to be made: first the total surplus generated by these contracts is positive and big relative to higher CRRA-coefficients (higher risk aversion means less value from risky payoffs and less collection of information overall), second the

[12] Artelys Knitro is the most advanced optimization solver for non linear constraints but is privately developed outside MATLAB.

unrestricted contract performs substantially better than the the other two. Third, although the decision and state payments linear-contract does better than the purely linear one, the two are quite close.

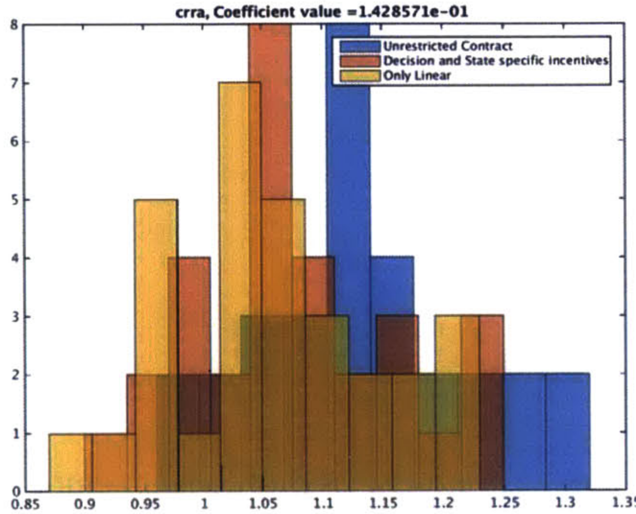


Figure 1-2: Surplus for the three contracts - CRRA 0.142

The distribution of the surplus for the log case in turn highlights that, as we proved in our theorem, the contract with state and decision payments is optimal so that its welfare distribution coincides with the unrestricted contract one. Moreover notice that the linear contract achieves negative payoffs in many random output matrixes. This is because unlike the other two contracts the linear contract cannot have decision specific payments and therefore the principal might not be able to incentivize no information collection through the choice of the ex ante optimal decision. It can then happen that the surplus from the ex ante optimal choice contract (b^{EA}) is higher than the one from the linear contract. Moreover as the risk aversion increases the optimal contract is more likely to incentive the ex-ante optimal action to avoid the agent bearing any risk.

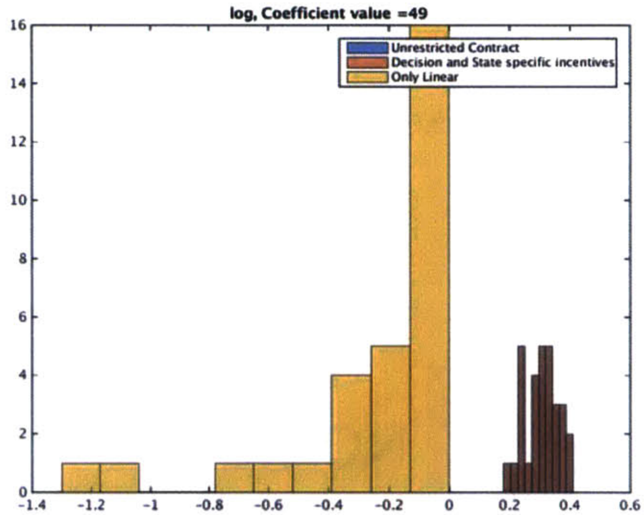


Figure 1-3: Surplus for the three contracts - CRRA 1

This is what the last histogram shows. The distribution of the unrestricted contract and decision and state payments linear contract has shifted closer to 0 (the surplus that would result if one were to incentive only the ex-ante optimal action). This is because as the risk aversion has become higher the incentives for the ex ante optimal action increase and the collection of information decreases.

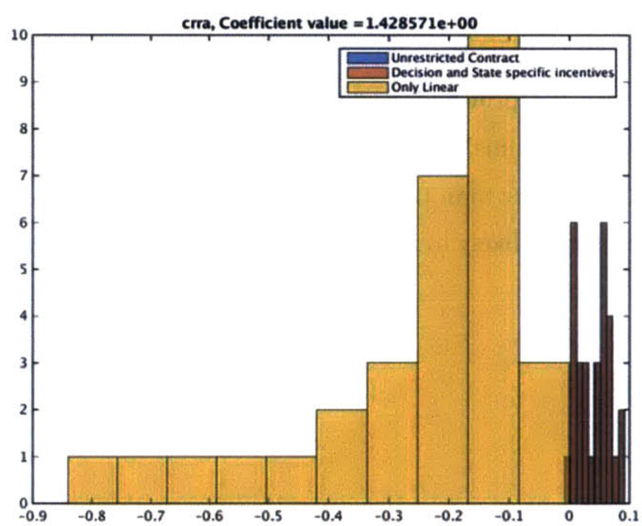


Figure 1-4: Surplus for the three contracts

Primitives	
Θ	Finite states
D	Finite Decisions
$\pi \in \Delta(\Theta^T)$	Prior over all states/times
$y : \bigcup_{t \in T} (D^t \times \Theta^t) \to \mathbb{R}$	Output function
$[y(d^t, \theta^t) - b(d^t, \theta^t)]$	Principal's flow utility
$u : \bigcup_{t \in T} (b(d^t, \theta^t)) \to \mathbb{R}$	Agent's flow utility
δ_p, δ_a	Principal and Agent discount factor
$b : \bigcup_{t \in T} (D^t \times \Theta^t) \to \mathbb{R}$	Contract
X	Signal space s.t. $\lvert X\rvert \geq \lvert D\rvert$
$f : \bigcup_{t \in T} (X^{t-1} \times \Theta^T)$	experiment
$\sigma : \bigcup_{t \in T} X^t \to D$	Decision Rule

Table 1.7: Dynamic Model

1.8 Dynamics

We consider a dynamic extension of the model with multiple periods. The agent takes a decision at each time and can acquire information about the states of the world today through the construction of an information structure. The principal wants to incentivize the agent to acquire information and chose the optimal decision dynamically.

Let time be finite, $t \in \{0, 1, 2, ..T\} = T < \infty$. We denote a variable at time t as d_t and the history of that variable, up to and including time t, as d^t. There is a finite set Θ of states. The agent needs to choose a decision from a finite set D after choosing an information structure given by a Signal outcome space X and and experiment p that respects the prior over all states over time π. The agent has a utility u over the path of decisions and states determined by the contract function $b(d^t, \theta^t)$. This dynamic setup is borrowed from Matejka et al. [2015], the table below summarizes the primitives of the model:

Matejka et al. [2015] show in Lemma 1 that the problem of the agent of choosing an experiment f and a decision σ over time:

$$\max_{f,\sigma} \mathbb{E}\left[\sum_{t=1}^{T} \delta_a^{(t)} \left(u_t\left(b\left(\sigma^t\left(x^t\right), \theta^t\right)\right) - \mu I\left(\theta^t, x_t | x_{t-1}\right)\right)\right]$$

can be simplified to the choice of a stochastic choice rule $p\left(d^t|\theta^t, d^{t-1}\right)$ and a predisposition function[13] $p\left(d^t|d^{t-1}, \theta^{t-1}\right)$. This result is the equivalent of the static reduction of signals to the decision space. They then show that the problem can be reduced to a sequence of static problems.

We depart from Matejka et al. [2015] in making the further assumption that the states and decisions are observed after their realization each period. This allows us to perfectly observe the history of states θ^t. Like Matejka et al. [2015] we assume that the agent can learn through the experiment about the current state of

[13]Which is just what we called in the static case the marginal over decisions.

the world and the past but not the future. The assumption that the agent cannot learn about the future and the fact that the state is perfectly revealed makes the prior over states conditional on a history (d^t, θ^t) identical to the original conditional prior $\pi\left(|\theta^t\right)$. We can then solve the problem backwards from the end of time obtaining that the dynamic contract has the same shape of our static one.

Theorem 26. *The optimal contract in the dynamic problem is:*

$$b\left(d^t, \theta^t\right) = \max\{0, K\left(y\left(d^t, \theta^t\right) + \delta_p v_p\left(d^t, \theta^t\right) + B\left(d^{t-1}, \theta^t\right) + C\left(d^t, \theta^{t-1}\right)\right)\}$$

with $K = 1$ for the risk neutral preferences and $K = \frac{1}{1+\mu}$ when the agent has log utility.

Proof. We can recast the agent problem recursively as:

$$\max_{p(\cdot|d^0,\theta^0),p(\cdot|(d^0,\theta^0),d_1)} \sum_{d_1} p\left(d_1|d^0, \theta^0\right) \sum_{\theta_1} p\left(\theta_1|\left(d^0, \theta^0\right), d^1\right) \widehat{u}\left(d^1, \theta^1\right) - \mu I\left(p\left(\right), p\left(|\right)|d^0, \theta^0\right)$$

where

$$\widehat{u}\left(d^1, \theta^1\right) = u\left(b\left(d^1, \theta^1\right)\right) + \delta_a v_a\left(d^1, \theta^1\right)$$

with $v_a\left(d^1, \theta^1\right)$ being defined as the continuation value for history $\left(d^1, \theta^1\right)$:

$$\max_{p(\cdot|d^1,\theta^1),p(\cdot|(d^1,\theta^1),d_2)} \sum_{d_2} p\left(d_2|d^1, \theta^1\right) \sum_{\theta_2} p\left(\theta_2|\left(d^1, \theta^1\right), d_2\right) \widehat{u}\left(d^2, \theta^2\right) - \mu I\left(p\left(\right), p\left(|\right)|d^1, \theta^1\right)$$

We can define recursively the remaining values up to T. Similarly, under the assumption that the principal does not have the ability to commit, we can recast the principal's problem as:

$$\max_{p(\cdot|d^0,\theta^0),p(\cdot|(d^0,\theta^0),d_1),b(d^1,\theta^1)} \sum_{d_1} p\left(d_1|d^0, \theta^0\right) \sum_{\theta_1} p\left(\theta_1|\left(d^0, \theta^0\right), d^1\right) \left(y\left(d^1, \theta^1\right) + \delta_p v_p\left(d^1, \theta^1\right) - b\left(d^1.\theta^1\right)\right)$$

subject to the maximization of the agent and with $v_p\left(d^1, \theta^1\right)$ being defined as the continuation value for the principal after history $\left(d^1, \theta^1\right)$:

$$\max_{p(\cdot|d^1,\theta^1),p(\cdot|(d^1,\theta^1),d_2),b(d^2,\theta^2)} \sum_{d_2} p\left(d_2|d^1, \theta^1\right) \sum_{\theta_2} p\left(\theta_2|\left(d^1, \theta^1\right), d^2\right) \left(y\left(d^2, \theta^2\right) + \delta_p v_p\left(d^2, \theta^2\right) - b\left(d^2.\theta^2\right)\right)$$

and recursively up to T.

Because the state and decision are observed at the end of each period, for any history (d^t, θ^t) and because we restricted the acquisition of information to the past and present, at any history (d^t, θ^t) the agent's prior belief over the states is given by his prior conditioned on the node given by the decision and state $\pi(\cdot|\theta^t, d^t)$. Therefore, the continuation payoffs are independent of any information acquired in the past and present. So the principal's problem can be solved at time T , and then by backward induction. At time T the problem is identical to a static one and so is the optimal contract and at each step we can solve the static problem with payoffs $\widehat{u}$ for the agent and $y + \delta_p v_p$ for the principal. With the result following from theorem 17.

A feature of the dynamic contract is the strength of the incentives on a path with high terminal value. Suppose that the tree of decisions leads for a certain state and decision path to a very high payoff, then the principal will incentivize for the full extent of those gains the agent along the path. This might seem surprising because if the cost of acquiring information is low, there are multiple steps before the realization of the high payoff, and the discount factor is close to 1 then the total incentives over time can be several times the payoff for the principal. This is only true in gross terms. Although the gross incentives can be several times the total final payoff, state and decision dependent negative payments can counterbalance the high incentives ensuring that the net incentives are lower than the final principal's payoff similarly to what we saw in the static case for output matrixes with high values. □

1.9 Conclusions

We have analyzed a principal-agent setting for an information collection problem. The agent needs to acquire information and take a decision for the principal. In line with information theory, we have modeled the cost of information acquisition through entropy-based mutual information. We approached the problem in two steps, first we derived the optimal signal and information structure under mutual information. And then, we used this insight to simplify the problem of the principal. We found that the optimal contract for a risk neutral or log utility agent with limited liability is linear in output with fixed state and decision payments. This implies that the optimal contract is linear with respect to any benchmark action or state. We interpret this feature as consistent with portfolio benchmarks in the long-only asset management industry and discuss how the optimal contract resembles the incentives that portfolio managers are given in the hedge fund industry. We also highlight the simplicity, in terms of reduced dimensionality of the optimal contract. We proved that entropy is the unique cost function for which such separability of state and decision payments occurs and therefore the unique cost function for which contracts are dimensionally simple. We then extended our result to different utilities, cost functions and constraints on the principal and the agent. We concluded by presenting an equivalent dynamic result and analyzing the robustness of our contract relative to a fully linear contract without state and decision payments. Although derived in a different setting and with different techniques our contract is similar to Carroll [2013]: both feature a linear coefficient on output and fixed state effects. In future research, we plan to use the numerical simulations to assess the robustness of these two contracts to different cost functions and utilities and to study their relative performance. We also plan to expand the dynamic section to situations where states are correlated through time but not fully revealed.

Chapter 2

Counting Votes Right: Strategic Voters versus Strategic Parties

Essays in Political Economics and Information Acquisition

by

Giovanni Reggiani

M.Sc. Economics and Social Sciences, Bocconi University (2010)
B.A. Economics and Social Sciences, Bocconi University (2008)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Author ....

Department of Economics
May 15, 2016

Accepted by ....

Ricardo J. Caballero
Ford International Professor of Economics
Chairman, Departmental Committee on Graduate Studies

2.1 Introduction

In the political economy literature, strategic voters are defined as citizens that would switch to their preferred contender party from their ideologically favorite one to avoid wasting their vote when a district is pivotal and too close to call. A careful estimation of the extent of strategic voting is the main objective of this paper.

A high fraction of strategic voters would be detrimental both for the representativity of the political system and for the level of competition among parties. Consider the case of a fully strategic population of voters uniformly distributed between a radically left and right parties. The great majority of these voters would prefer a centrist party to be in power but none of them would switch their vote from their preferred extremist party to the centrist because they would anticipate that this would give victory to the opposite wing. In this example we see how strategic voting can prevent external competition against status quo parties, and therefore limit moderating political competition as well as the candidates quality enhancement. It could well be that the majority of the population is exasperated with the extremism of the two main parties but yet, because of their strategic behavior, refuses to vote for a third centrist party that better represents their views (Duverger [1959]).

It's important to note that the representativeness of the elected parties is not a philosophical concern about Democracy, rather an economic one. Governments are responsible for the level, allocation, and quality of public goods, transfers and services for a share of between 30% to 50% of GDP in developed democracies. Therefore if, due to strategic voting, representative democracy is not an effective way to convey the preferences of the citizens, the efficiency of the allocation of government spending is a first order economic concern as both parties that receive votes in equilibrium could make decisions inconsistent with citizens' preferences.

Because of these concerns, understanding the extent of strategic voting is a central question to evaluate the functioning of modern democracy. Recent work in economics has estimated the fraction of strategic voters to be 70% (Kawai and Watanabe [2013a]). If a majority of voters are indeed strategic, there is no easy fix with a change of electoral rules, Satterthwaite [1975] showed that for electoral systems to be strategy proof they need to be either dictatorial or non deterministic. In the present work, we apply a new geographical RD methodology (Pinkovskiy [2013]) and a conceptual insight on the importance of comparing two identical votes to derive much smaller estimates (5%) that reassure us on the fact that strategic voting is not a first order concern for democracy and for the economic efficiency of government services.

We argue that previous larger estimates are due to the use of non identical votes which makes controlling for strategic behavior by parties increasingly difficult. Indeed, the main parties also have incentives to behave strategically in a tossup district; via enhanced allocation of resources to the district, selection of more appealing candidates, or shaping policy towards the constituents of such district. As a result even sincere supporters for minority parties may vote for one of the two main contenders not because they are strategic but because the strategic behavior of parties has accommodated their preferences.

Our identification strategy exploits the uniqueness of the perfect Italian bicameralism and a recent temporary change in the electoral law, which asymmetrically assigned the majority premium between the Congress and the Senate, to observe two identical votes and to have variation in the pivotality and significance of the vote. We consider the difference in the share of votes of the top two Coalitions across the two Chambers in each municipality as the outcome variable. These two Coalitions are the only true contenders for the electoral victory and therefore the only two that could be positively affected by strategic voting. By considering the within Municipality difference in votes across the two Chambers we remove any factor that affects voters actions symmetrically in the two Chambers. We then test whether our outcome variable is affected by the regional changes in electoral strategic incentives. If we consider the municipalities on the border, we can exploit multiple geographical discontinuities with different treatments by using a two-step geographic RD estimator recently developed by Pinkovskiy [2013]. In light of this new approach and the two identical votes setup, we find significant but very small estimates, consistent with the important contribution of strategic parties to total misalligned voters.

To understand the role and necessity of each part of our empirical strategy let us think about the ideal test we would like to run. In an ideal randomized controlled trial, we would like to observe how the vote of the same individual changes when we change her beliefs about the probability of being pivotal. We argue that our empirical strategy is as close as possible to this ideal benchmark. The Italian Constitution perfect bicameralism, i.e. the fact that the two Chambers perform the same functions, allows us to observe two identical votes. The fact that the majority premiums are determined at the Regional level (Senate) and the National level (Congress) allows us to have exogenous variation in the probability of being pivotal. In addition, the difference between the outcome variable in Senate and Congress allows us to absorb district fixed effects such as higher campaigning by national leaders or advertisement, and the difference across regions allows us to use variation in the pivotality beliefs of voters. The regression discontinuity helps to guarantee that the distribution of costs for voters is similar and to mitigate possible ecological fallacies (see model). Lastly, the fact that electoral lists are closed and long - impeding the knowability of the candidates - allows us to attenuate the concern of strategic parties driving the results through candidate selection. This is discussed more in detail in section §2.5.

Our contribution is twofold; we conceptually highlight the importance of the distinction between strategic voters and strategic parties providing theoretical and empirical arguments on its importance, and we provide estimates on the extent of strategic voting that are much lower than those previously suggested (5% versus more than 30% Spenkuch [2012] and 75% Kawai and Watanabe [2013a]). We discuss in section §3.2 these works, their identification assumptions and how their much larger estimates might be due to a joint estimation of strategic parties and strategic voters. We prudently interpret our results themselves as a lower bound. We believe that the unique institutional setting (see section 2.2.1) and lower estimates are consistent with a lower bound close to the true value.

In the next section we provide the institutional details relevant for our strategy, and in section §2.3 we present a simple model of our empirical framework. In section §2.4 we illustrate the data, the empirical strategy, and why we should use geographical RD to properly test the predictions of our model. In section 2.4.5 we present results from the two step estimator and in section 2.4.4 we study the case of Lombardy and Emilia Romagna where the starkest incentives were at play. We discuss possible concerns and the meaning of our estimates in section §2.5. Finally, before concluding we discuss the previous literature on strategic voting, previous estimates, and the discrepancy between our estimates and previous results.

2.2 Background Information

2.2.1 The Parliament

After the end of the Second World War, and the experience of Fascism, the authors of the Italian Constitution chose a *perfect bicameralism* to prevent future dictatorships. They prescribed the existence of two chambers (Congress and Senate) with identical powers and functions. Any law needs to be initiated by one of the two Chambers and approved by both with no exceptions. Similarly, the executive needs to have the approval of both chambers to remain in power. The only difference between the two Chambers is their size and active and passive electorate. Citizens need to be 18 to vote for Congress and 25 to vote for Senate, while citizens need to be 25 to run for Congress and 40 to run for Senate; Congress has 630 members and the Senate has 315 members.

2.2.2 Electoral Law

Akin to gerrymandering in the Anglo-Saxon world, changes in electoral laws have been a constant of Italian politics. Parties in power change rules to improve their odds or make government harder for their opponents. In this section we explain the Italian electoral law, why it has recently been ruled unconstitutional, and how it facilitates our test. The law n. 270 approved on the 21^{st} December of 2005 has been the Italian electoral law for the elections of 2006, 2008 and 2013. It was ruled unconstitutional by the constitutional court on December 4^{th} 2013. The electoral law has not been known as n.270, but rather as "Pig Crap" ("Porcellum") ever since its writer used such a nickname in an interview to define his own work. [1] Historically seats have been assigned at the national level. The new electoral law kept the majority premium at the national level

[1] Silvio Berlusconi was in power in 2005 and was fairing badly in survey polls. He expected to lose to the center left under the previous electoral law. The opposition party argued that Silvio Berlusconi requested to write an electoral law that would make harder the creation of a Government for the Center-Left. "Pig crap" is an expression used sometimes in Italian to indicate a dirty trick, and hence the law's nickname. The dirty trick is easily explained. Historically, the center left and Berlusconi are very closed in the nationwide support but their geographical distribution is quite different. Most of the support for the center left is concentrated in the regions in center Italy (Emilia, Tuscany and Umbria) where it usually wins with extremely high margins; Berlusconi instead usually wins the remaining regions though with smaller margins.

for congress and made it region-based for the Senate. The electoral law is fully proportional with majority premiums for both chambers. Parties need to select themselves into coalitions and indicate the name of the coalition leader as well as subscribe a program. The coalition that gets most votes at the national level in the Congress ballot receives a full house majority regardless of its electoral weight. For the Senate there is a hefty premium Region by Region to the coalition with a plurality of votes in each region[2]. The "Pig Crap" law also abolished preference votes: parties choose their candidates and people vote closed lists without ranking their preferences. This gave enormous power to parties and more leverage to their whips that could use threats of future blacklisting to bargain with members of parliament.

These two provisions were the explicit motivation for the Supreme court ruling the law as unconstitutional. But, as we have already hinted in the introduction, these two features are the foundation of our identification strategy.

Even if the same electoral law was in place since 2006, the 2013 election was the only one valid for our identification. Up until 2013, the Italian political landscape has been dominated by two coalitions: Center-Right and the Center-Left. In 2006 there were only two coalitions: the Center-Left, led by Prodi, that spanned from the very far left to the center, and the Center-Right, headed by Berlusconi, that spanned from the very far right to the center. Together they got more than 99% of the votes so there were no third coalitions. Without a third coalition there cannot be strategic incentives. In 2008, there were two additional small coalitions[3]. However, Berlusconi was widely expected to win everywhere (both Congress and Senate) and no important regions were actually toss-up nor pivotal. So again there were no incentives to vote strategically. In 2013 instead, many big regions (Lombardy, Veneto, Sicily and to a lesser degree Campania and Puglia) were tossups for the first time. The Congress was thought to be won safely by the Center-Left but the Senate was uncertain. There were two other parties - Monti and 5Stars - that could not win the majority premium in any region nor the national congress but whose voters could potentially affect the regional vote and therefore the final outcome in the Senate. The 2013 elections provided a unique case of varying expectations at the national and regional level for big pivotal, tossups regions.

2.3 A Simplified Framework

In this section we consider a simplified theoretical representation of the Italian Elections in 2013, the purpose of this section is to clarify why we use as an outcome variable the difference between the two chambers of the sum of the percentage votes of the two main parties and to explain why RD is necessary to obtain correct estimates.

Our model will follow the following abstraction:

[2]But no further premium if the coalition would have already reached the super majority

[3]One on the very far left and one in the center. Both ended up below 6%, and only one made his way into the Parliament

- Each individual casts two votes, one for each Chamber;
- The majority premium is given at the National level for Congress and at the regional level for the Senate.
- There are three parties: A,B,C. Parties A and B are likely contenders in some regions. Party C is never a contender for victory.

This parsimonious setting will be enough to characterize the prediction that we will test. Parties A, B should receive relative more votes in Senate than in Congress when we compare tossup regions to non tossup regions.

2.3.1 A numerical Example:

Let each individual have a party ranking denoted by the vector r. She casts two votes, has utility 0.1 from voting her favorite candidate and utility of 0.5 if she is the pivotal voter and gets a less disliked candidate elected. Assume that there is no uncertainty about turnout and preferences and only two regions: Emilia and Lombardia. Table 2.1 presents voters types per region. Remember that one premium is determined at the regional level (Lombardy and Emilia in the above example) and one at the national level. Under common knowledge, we solve this example by considering the only possible equilibrium if all voters are strategic.

Table 2.1: Numerical Example Ranking Types per Region

types of voters	1	2	3	4
Rank1	A	B	C	C
Rank2	C	C	A	B
Rank3	B	A	B	A
Lombardy	4	4	0	1
Emilia	8	1	1	1
Total	12	5	1	2

Each voter supporting either Party A or B will cast both votes for her favorite party. Types 3 and 4 instead may face a strategic dilemma. Their favorite party (C) cannot win so they may be tempted to cast a strategic vote. First, consider the Congress vote: since they know that the election is not close their dominant strategy is to cast the vote for C irrespective of their region of residence. In the case of Senate, the premium is given at the Regional level, so incentives are different for voters in different regions. The type 4 voter in Lombardy knows that she will be pivotal. Therefore her optimal strategy is to vote for B in the Senate election. Type 3 and 4 in Emilia instead know they will not be pivotal in the Senate election so they will vote C as they did in the congress ballot. Therefore the electoral outcome would be:

Table 2.2: Summary of Results for Numerical Example

	Congress L	Senate L	Congress E	Senate E
Party A	4	4	8	8
Party B	4	5	1	1
Party C	1	0	1	1
A+B	8	9	9	9
$\Delta\left(A+B\right)$	-1		0	

Denote by $\Delta\left(A+B\right) = \left(Votes_A^{Congress} + Votes_B^{Congress}\right) - \left(Votes_A^{Senate} + Votes_B^{Senate}\right)$. In this specific example we have $\Delta\left(A+B\right)_{Lombardy} < 0$ and $\Delta\left(A+B\right)_{Emilia} = 0$. More generally the prediction that we will be testing is: $\Delta\left(A+B\right)_{Lombardy} - \Delta\left(A+B\right)_{Emilia} < 0$.Here, we see that $\Delta\left(A+B\right)_{Lombardy} - \Delta\left(A+B\right)_{Emilia} = -1$.

This negative sign reflects the fact that strategic voters are relatively more likely to switch to one of the top two parties in toss-up regions versus non toss-up regions.This simple prediction is generalized in the following Section and is at the core of our empirical specification.

2.3.2 A general set up

There are three parties (A,B,C) and two ballots (Senate and Congress). The majority premium is at the national level for Congress and at the Regional level for the Senate. C is never a likely winner neither in the national election nor in any region. We assume that there is common knowledge of the probability of being a pivotal voter at the national level $\pi(\alpha,\beta)$ and at the regional level $\pi^j\left(\alpha_j,\beta_j\right)$.[4]These probabilities are derived from the distance in polls (α) and the size of the electorate in the region (β). [5]

Each voter ι is characterized by a vector ranking her preferences r over the three parties (e.g. A,C,B or B,C,A etc), and a region j to which she belongs. We summarize the characteristics of voter ι with a tuple $\left(r, j, \delta_\iota\right)$. Each voter casts two votes a_c and a_s, one for the Chamber and one for the Senate. She has "warm glow" utility δ_ι from choosing her first ranked party and a different utility value for the various possible rankings of parties (to simplify we assume she only cares about the winner of that ballot). Therefore the utility when choosing a_c for the congress elections and a_s for the senate elections is:

$$U_\iota^C = \delta_\iota \mathbb{I}\left(a_c = r\left(1\right)\right) + \sum_{z=A,B,C} \nu_\iota^z \mathbb{I}\left(Chamberwinner = \ j\right)$$

[4]Allowing these probabilities to depend on individuals' priors $\pi_\iota(\alpha,\beta)$ would not change our results (as long as we assume common priors) but would introduce unnecessary complication.

[5]Notice that it is obvious that lower alpha (more contested elections) means higher probability of being pivotal all else constant, while the same is not true for β. On one side lower population means that there are fewer seats awarded in the senate and so the overall probability of being pivotal in the senate goes down, but the fact that there are fewer voters also makes it more likely to be precisely the pivotal voter for constant α.

$$U_\iota^S = \delta_\iota \mathbb{I}(a_s = r(1)) + \sum_{z=A,B,C} \nu_\iota^z \mathbb{I}(Senatewinner = j)$$

Each agent has a different taste for his utilitarian pleasure from coherence (δ_ι)[6] We assume that it is distributed across regions with region-dependent distribution F_j over support $[0, K]$. We also assume that the ranking types (r) are identically distributed across regions and that ν_ι is the same for voters with the same ranking. That is $\nu_k^l = \nu_z^l \forall l \in \{A, B, C\}$ as long as $r_k(m) = r_z(m)\ \forall m \in \{1, 2, 3\}$ [7]

We now show how the optimal strategies of such rational (strategic) voters would depend on the poll distance and the population of the region.

Remark 27. Voters whose favorite candidate is A or B, will cast a ballot for such candidate in both the Senate and the Congress irrespective of public information and their regional residence (j).

Proof. See Proof. □

It remains to show how the C supporters vote.

Proposition 28. *In every region j, and election ballot M - where N stands for national ballot and r for regional ballot from region r - there exists a cutoff $\overline{\delta}^{M,j}$ such that all C supporters s.t. $\delta_\iota \leq \overline{\delta}^{M,j}$ vote strategically $a_M = r(2)$ and all C supporters such that $\delta_\iota > \overline{\delta}^{M,j}$ will vote sincerely $a_M = C$.*

Proof. See Proof. □

Now remember that in our initial discussion we highlighted how smaller α_js correspond to region j being more tossup. The effects of β instead depend on the underlying model used to describe $\pi(\alpha, \beta)$.

Proposition 29. *Under identical benefits distribution F across regions, if a region l is relatively more tossup than a region j $(\alpha_l < \alpha_j)$there will be relatively more misaligned voters, i.e.:*

$$\left(F_l\left(\overline{\delta}^{N,l}\right) - F_l\left(\overline{\delta}^{l,l}\right)\right) - \left(F_j\left(\overline{\delta}^{N,j}\right) - F_j\left(\overline{\delta}^{j,j}\right)\right) < 0 \tag{2.1}$$

Proof. See Proof. □

Remark 30. Note that this result depends on assuming identical distributions (F) across regions. An easy counterexample to the above theorem would be letting region j be more tossup than region l but assuming that the distribution of the benefits (δ) is a Dirac measure concentrated on K while the distribution of F_j is uniform.

[6]We do not actually need to have a behavioral assumption. Given that the system is proportional, it is enough that voters prefer to be a larger minority rather than a smaller minority.

[7]Again this assumption is not necessary but simplifies the argument.

From this little formal set up we can be pretty clear about what we should observe in the data:

- We should observe that the parties A,B receive more votes in the Senate relative to the Congress for the tossup regions relative to non tossup
- We should observe that A,B receive more votes when α is lower.
- Unclear predictions for comparative statics on β.

Our model predicts that all voters playing strategically will be supporters of party C. These are not the total strategic voters but just the misaligned voters. Strategic voters' estimates should be adjusted for the size of party C.

As highlighted before, our results might not hold if F differs across regions. It is reasonable to think that F is not constant across the country, but it is reasonable to believe that F is constant in municipalities within few km. In this case, geographic Regression Discontinuity would consistently estimate strategic voting.

2.4 Data and Empirical strategy

2.4.1 Data

We merge three municipalities data-sets on voting, social-economic background, and geographic coordinates. Furthermore, we use the most recent electoral polls , published by television broadcast Sky, to measure the level of strategic incentives at the regional level.

Voting data are provided by the Historical Office of the Italian Department of State ("Ministero degli Interni") and contain municipality level votes by party for both Senate and Congress. We aggregate the party votes into coalitions because this is the level at which the strategic incentives operate. We use electoral documents to classify parties into coalitions. Coalitions are the same for Senate and Congress across Italy and they are officially defined before the election. The parties within a coalition can vary across regions, but for any Region they are the same between Senate and Congress. For instance, the Center-Right coalition contains some regional parties that appeal to local pride. Since our identification looks at the shift of votes across coalitions between Congress and Senate at the Municipality level, this heterogeneity is not a concern.

Italy is divided in 20 Regions and 8092 Municipalities. It is worth stressing that municipalities do not correspond to electoral districts, meaning that all the municipalities within a Region face exactly the same type of voting process. We drop from our analysis two regions: Valle d'Aosta and Trentino-Alto Adige. The reason for this choice is that the electoral rule in these regions is different from the standard one[8] and

[8] Valle d'Aosta (VA) and Trentino Alto Adige (TA) are two autonomous regions with a special statute. Therefore, the Constitution allows them higher legislative flexibility.

that local parties representing the interests of linguistic minorities play an important role in these regions. Dropping these two regions brings the number of municipalities used in the analysis to around 7700.

We collect various demographic and economic data from the last issue of the "Atlante dei Comuni". This is published by ISTAT, the national Bureau of Statistics, and contains information at municipal level. ISTAT also provides us with cartographic information for all the Italian municipalities. We measure the distance of a municipality to the local regional borders using geo-coded information. The procedure we employ is the following. First, we use the coordinates of the border of each municipality to determine its centroid. Then, for each municipality, we define as the distance of the municipality to the border as the airline distance to the closest point of the border. We also compile manually a list of municipalities right at the border.

For poll data, we resume to official sources. The Italian government established a web-site[9] where every media company is required to publish any public electoral poll. Using the web-site, we identified the poll that: (a) was closest in time to the election date; (b) had data on the intention to vote at regional level for the Senate ballot. While there are many pools in the period before the election, very few cover something different than the national result or a small subset of regions. The previous criteria led us to select the poll produced by the marketing company Tecne' for the TV broadcast Sky, one of the 3 biggest Italian TV group and subsidiary of the multinational group News Corporation[10]. The results of the poll are provided in the appendix. Using this data, we construct an index to capture how tossup Region j was before the election:

$$Tossup_j = -|CenterLeft_j - CenterRight_j|$$

Where $CenterLeft_j$ and $CenterRight_j$ are the expected share of votes at the regional level for the two main coalitions in the Senate ballot. In other words, this measures how close the top-two coalitions are expected to be right before the election. The closest the index is to zero, the more toss-up a region is and therefore the more we should expect voters to engage in strategic voting, as predicted by the model. Notice that the index is always negative, therefore the higher the index, the more tossup the region.

2.4.2 Introduction to the Empirical Strategy

In an ideal randomized experiment on strategic voting, we would observe how the vote of the same individual changes under different beliefs about the probability of being pivotal. While not quite identical to an actual randomized experiment, the Italian Electoral system in 2013 had some features that made it very well suited to answer this question. First of all, Italy is a rare example of perfect bicameralism, where the two elected Chambers have exactly the same institutional role and they differ only on their size and the rules governing

[9] http://www.sondaggipoliticoelettorali.it/

[10] The poll is based on surveys run around February 11th, each poll is stratified at the regional level by the socioeconomic and geographical residence. Since two Regions had missing info, we fill the gap with the equivalent poll produced the week before by the same company, Tecne', for Sky.

active and passive electorate. A consequence of this perfect bicameralism is that any voter should have the same exact preference ranking across coalitions in the two Chambers. Comparing the share of votes going to each of the two Chambers we difference out the true underlying preference of the voters. More generally, the difference between the two reflects only factors that have an asymmetric effect across the two Chambers. Secondly, the "Pig Crap" law is characterized by a wide level of heterogeneity in strategic incentives across the two Chambers. While the seats in the Congress are assigned at national level, the seats in the Senate are assigned based on the electoral results at the regional level. We focus here on the incentives generated by the large majority premium assigned to the coalition receiving the largest share of vote in each region[11]. While the strategic incentive for the Congress is constant across the whole country, the incentive for Senate changes region by region, depending on the level of closeness of the two-contenders coalitions, which is measured by our $Tossup_j$ index. We exploit this geographical heterogeneity in our empirical strategy.

A simple model of strategic voting predicts that supporters of non contenders in toss-up Regions, will be relatively more likely to switch to one of the top two coalitions in the Senate ballot. This is the prediction of Proposition 29 of our model. The condition in Proposition 29 translates into the following linear equation, where we would expect the parameter δ to be negative:

$$\Delta^{C-S}_{\%_{ij}} = \alpha + \delta\left(TossUp_j\right) + \beta X_{ij} + \epsilon_{ij}$$

$\Delta^{C-S}_{\%_{ij}}$ is the difference in the sum of votes for the two main coalitions between Congress and Senate, $TossUp_j$ is an index that measures the level of closeness of the parties in the pre-electoral poll, higher values of the index imply more closeness, and X_{ij} is a set of covariates at the municipality level. Under the assumption of no heterogeneity in preferences, the parameter δ is a consistent estimator for the the share of misaligned voters for a given level of closeness in the electoral race. However, the assumption of no heterogeneity in preferences is very strong. If it fails, the least squares estimator for δ is consistent if and only if the set of covariates X_{ij} controls for all the observables and unobservables heterogeneity across municipalities. Here, we employ a geographical Regression Discontinuity setting to relax the identification assumption and provide evidence on strategic voting with high internal validity.

The intuition behind the Regression Discontinuity framework is the following. Consider comparing two adjacent cities, call them A and B, that are separated by a Regional border. Given that the population lives just a few minutes apart and given that there is full mobility of factors and people across Regions, we expect these two cities to be identical, both in observables and unobservable characteristics. However, the two cities crucially differ in the expectation regarding the results for the Senate race at regional level. Our test looks at how the difference in votes for the contenders' coalitions change as a function of the ex-ante perception on regional tossupness.

[11]For Congress, the (same) majority premium is assigned at national level

While we show that municipalities close to the border are similar in observables characteristics, Regional borders could be associated with discontinuity in other relevant dimensions not captured by our set of covariates. For instance, Municipalities across the border differ in the identity of the belonging Region. Italian Regions have an important role in public good provision and more broadly in the local economy. It follows that Regional institutions may be an important determinant of political orientation. However, this is not a concern for us: since both Chambers have exactly the same role in the political decision making and given that we focus on the difference in voting across the two, local specific fixed effects would not be a concern for our results[12]. The only threat to our identification comes from factors that affect the votes across the two Chambers and that are correlated with the Toss-up index presented above. Strategic behavior by parties is the main confounding factor we have in mind. In the last part of the paper, we argue that this concern is very unlikely to be first-order here, because of the electoral institutional features. Our estimates could be interpreted as a lower bound because the function that relates the number of strategic voters to the tossupness of the election could be non linear. By estimating the average effect over treatments of differing intensity the non linearity could lead to an underestimation of the percentage of strategic voters.

In the next Section, we discuss more in detail the empirical framework. After that, we focus on the most relevant case of difference in strategic incentives, the border between Lombardy and Emilia-Romagna. This example will help build intuition about the empirical framework as well as providing a plausible bound for our estimates. In the end, we present our main results and discuss robustness tests.

2.4.3 The Regression Discontinuity test at the National Level

In the setting of our analysis, there are a total of 27 Regional borders, for a total of 54 border-sides. Our empirical framework differs from the standard Regression Discontinuity framework because we have different boundaries where the treatment changes discontinuously but in potentially border specific ways. In the basic case, a forcing variable (distance from the border here) defines one relevant discontinuity and the test focuses on studying how the outcome of interest discontinuously changes across it. Here, we need to develop a test that is generalizable to the 27 borders. A notable example in the literature is Black [1999]. She is interested in studying the effect of school quality on house prices, using school district boundaries in Massachusetts to identify the relevant causal effect. While her framework is particularly valuable because of its simplicity, it may lack in flexibility for a case where it is not possible to use observations exactly on the discontinuity. We therefore use a more general two-step framework, developed by Pinkovskiy [2013]. As a robustness, we present our results also using a framework equivalent to Black [1999] and Dube et al. [2010], and we show that the results are unchanged and, if anything, statistically stronger. [13]

[12] Two people with the same underlying preferences might have different political tastes at the national level in two regions or even just two municipalities that have different policies, but such preference would be reflected in the same (potentially different between the two people) party vote at the Congress and Senate.

[13] A strategy similar to Dube et al. [2010] has been recently used in Naidu [2012], where fixed effects per pairs of counties are added to isolate the effect of state level law changes.

The intuition behind the Pinkovskiy's procedure is simple. Different borders may differ in their size, density of municipalities and dependence of the outcome on the forcing variable. Therefore pooling together the whole set of municipalities at the border may not be the cleanest procedure. One way to look into this is to aggregate at border-side level the information. This is what we do here. In a first stage, we estimate the conditional expectation at the border of our outcome variable $\Delta^{C-S}_{\%_{ij}}$ separately for each border-sides set of municipalities. In the second stage, we use these estimates as outcomes in a cross-sectional regression on the level of closeness in the regional race, $Tossup_j$. In practice, we start by estimating the following equation for each border-side separately:

$$\Delta^{C-S}_{\%_{ij}} = \tilde{\Delta}^{C-S}_{\%_j} + \rho(d_{ij}) + \epsilon_{ij}$$

This is estimated over the set of all municipalities that are within a bandwidth of B km from the relevant border, assuming ρ as linear. Notice that $\tilde{\Delta}^{C-S}_{\%_j}$ is simply the constant of the least-squares estimator and it estimates the conditional expectation of the outcome variable at the border ($d = 0$). Since we estimate a different function per each border-side, we allow total flexibility on the conditional expectation across border-sides. Then we use these estimates in a second stage as outcome. In particular we estimate:

$$\tilde{\Delta}^{C-S}_{\%_j} = \alpha + \delta Tossup_j + \beta \tilde{X}_j + \epsilon_j$$

where $\tilde{X}_j$ are the conditional expectation of the standard covariates at the border[14]. The observations are weighted by the number of voters at the border-side. Given the definition of the variables, the theory predicts that δ should be negative in presence of strategic voting. The standard errors in this model are clustered at the border level. In the result section, we discuss some relevant specification robustness, such as allowing for border specific fixed effects.

This framework as any other Regression Discontinuity requires two important assumptions. First, we need that every relevant factor different from the main outcome is a smooth function of distance across the discontinuity. In the result section, we show this is actually the case for a set of important covariates. This is not surprising, since the sets of municipalities that are compared are usually very close and regional border do not determine any relevant change in labor markets institutions, credit or infrastructures.

Second, we need to make sure that our results are not simply driven by sorting of citizens across the border. Lee and Lemieux [2009] argues that sorting is probably the first order concern around geographical Regression Discontinuity. People can choose where to live based on their own preferences and characteristics. If the endogeneity of the choice is related to the mechanism we are testing, then our estimates could be biased. Luckily, we can strongly reject this criticism. In this context, it is highly implausible that the location choice can be related to voting behavior as described in our model. In Italy, voters need to vote in the Municipality

[14] We have simply run the same exact first stage procedure on the covariates

where they are resident. Changing the Municipality of residence is costly and time consuming[15] and therefore it is highly unlikely that anyone would undertake this process for the small and uncertain gains from strategic voting.

It is important to make the following point about inferences. Pinkovskiy [2013] argues that the standard type of asymptotic, where data are independently generated with the number of observations going to infinite, may not be particularly well suited for the case of geographical Regression Discontinuity. He proposes a new estimator for the variance under infill asymptotic. In the standard asymptotic, the domain from where observations are drawn is thought to go to infinite: with infill asymptotic instead we have that resolution of our data increases to infinity. In our setting, this would be equivalent to have data on votes for infinitely smaller municipalities. He shows that, when infill data are used and errors are correlated, the standard variance estimator is too conservative. While we believe the analysis presented by Pinkovskiy [2013] is of great interest and deserves more exploration in the future, in our work we decided to use the standard White estimator for the variance. There are two, related reasons why we make this choice. First, the typical standard errors are overly conservative and therefore they would prudently bias us against finding any statistical significance. Secondly, the actual properties of the estimator of the variance proposed by Pinkovskiy [2013] are not well known in finite sample.[16]

2.4.4 A case study example: Lombardy vs. Emilia-Romagna

We start the presentation of the results, by looking at the border between Lombardy and Emilia-Romagna. This case study is important for multiple reasons.

First of all, it helps exemplifying the intuition behind the Regression Discontinuity approach. Secondly, this border represents the cleanest example of discontinuity in strategic incentives, because it compares one almost perfectly toss-up Region (Lombardy) with one where the electoral result was completely uncontested (Emilia-Romagna). According to the most recent polls before the election, in Lombardy Center-Left and Center-Right were expected to be tied. Elections were expected to be decided by thousands of votes and the strategic importance of this area was particularly stressed by media and politicians. Instead, Emilia-Romagna was, together with Tuscany, the least contested Region in Italy. The Center-Left was expected to lead the race by at least 20 percentage points. Historically, the communist party first and the center-left coalition afterwards had always won the election in this area post World War II. Lombardy was not only more contested, but, because larger than Emilia-Romagna in population, its victory was more determinant for the final outcome. All in all, the expected value of voting strategically was characterized by a large jump

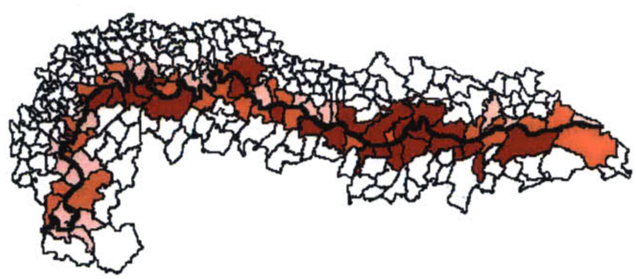
[15]Voters need to apply to the new Municipality much in advance than the Election, providing a proof of residence in the new Municipality. Local police then need to validate the information provided by inspecting the new residence. The whole process may take weeks, if not months, and it requires filing multiple forms and paying some fees

[16]In applying its theoretical framework to his empirical problem, Pinkovskiy [2013] develops a routine where he uses either the White or its own variance estimator depending which one is smaller. The idea is that the variance he develops is always smaller than the standard White estimator in asymptotic, but this may not be in finite sample. That's why choosing White standard errors is more conservative in our setting.

across the border. In the end, this case is particularly interesting because it provides an upper bound for the size of strategic voting.

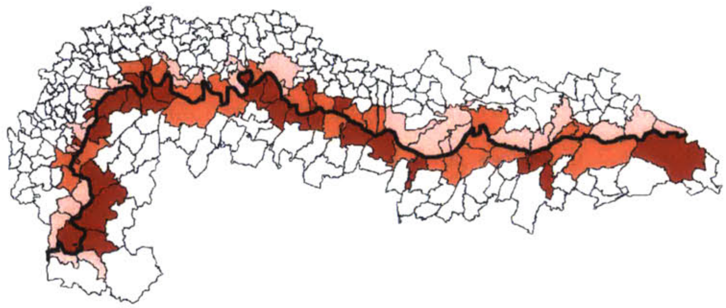
Since we are comparing the most contested with the least contested Region, the probability of being pivotal is characterized by a sharp discontinuity at this border. However on other dimensions the two Regions appear to be quite similar at the border. Lombardy and Emilia-Romagna are the two richest Regions of Italy, among the richest in Europe. While quite different in many instances, along the border their differences shrink. Consider for instance, the subset of Municipalities which are exactly on the border between Lombardy and Emilia-Romagna. This set of 83 Municipalities is the closest you can get to an ideal Regression Discontinuity (figure -7). By construction, every Municipality at one side of the border is contiguous with at least another Municipality in the (counter-factual) Region. Consider the heat map presented in figure -8, where the running variable is income per person. Income does not appear to be characterized by any discontinuity across the border. In fact, if anything, the spatial distribution of income seems to be characterized by clusters along the border, with Municipalities with similar income grouping together across the border. This pattern is not unique of the Municipalities along the border, but the results do not change when considering wider bandwidth around the border line (look figure figure -9 for same with a 10km bandwidth). Furthermore, results are qualitative similar when considering other outcomes.

Figure 2-1: Distribution of Income by Municipality, Lombardy vs. Emilia-Romagna



Notes: this map contains all the Municipalities in Lombardy or Emilia-Romagna, that are within 10km from the other Region border. Colored are the Municipalities that are on the border. A darker color signal higher level of income per capita in that Municipality. Data on income are provided by ISTAT. The black line is the border between the two Regions.

Figure 2-2: Distribution of the outcome by Municipality, Lombardy vs. Emilia-Romagna



Notes: this map contains all the Municipalities in Lombardy or Emilia-Romagna, that are within 10km from the other Region border. Colored are the Municipalities that are on the border. A darker color signal higher level of the outcome variable, which is the difference in the share of votes going to one of the top-two Coalitions between Congress and Senate at Municipality level. Data on income are provided by Italian Department of State. The black line is the border between the two Regions.

However, the story is different when we look at our outcome variable $\Delta^{C-S}_{\%_{ij}}$ as input in the heat map (figure -9). Here instead, the discontinuity is quite evident. The Municipalities in the north side of the border appear to have on average a larger share of voters voting for the two contenders coalitions in Senate than in Congress. The result is confirmed when looking at different bandwidths.

The story at this point is clear: Municipalities at the two borders differ by the level of incentives for being strategic, but they tend to be very homogenous along other dimensions. While the case of Lombardy versus Emilia-Romagna cannot be generalized further, the bottom line is confirmed by further tests in the national sample. As we will show, in fact, the level of pivotality of regions does not appear to be correlated with values of covariates at the discontinuity.

The results presented graphically can also be confirmed in a formal specification. We consider four samples. We start considering what we think is the closest to the ideal RD setting, which is the case where we analyze only the behavior of Municipalities right at the border. We then consider the set of Municipalities whose centroids are at 10,15, and 20 km from the border. Results will be both qualitative and quantitative similar across the different samples. We run a simple linear local regression, where we study how being in the more toss-up Region (Lombardy) affects the behavior of voters in close-by Municipalities[17].

The results can be found in table 7. Crossing the border of Emilia-Romagna and Lombardy implies a drop in the outcome variable of about 1.5%. Later we discuss how to recover the extent of strategic voting from these estimates, under quite mild assumptions. In this case, it is only worthy to point out that these estimates imply a maximum extent of strategic voting around 5% once we correct for misaligned voters. While still substantial and potentially relevant, this is very far from previous work. In Column (1) we present the

[17]The equation of interest is the following:

$$\Delta^{C-S}_{\%_{ij}} = \alpha + \delta Lombardy + \rho(d_{ijk}) + \epsilon_{ij}$$

where d_{ijk} is the distance of Municipality i to the border between j and k, and $\rho$is assumed to be linear and different at the two sides of the border. Observations are weighted by number of voters.

simple coefficient with the border regression, where no correction for distance and covariates is applied. In Column (2) to (4), we subsequently add controls and distance functions[18]. Results are stable. Notice that in Regression Discontinuity, adding controls is not required for identification. We add them here mostly for robustness and to gauge the precision of our estimates. In the end, between Columns (5) to (7), we present the results for the different distances, and in particular for all the Municipalities within 10,15, and 20 km from the border. Again, the results are not statistically different from each other and in line with what expected. In the end, in the appendix we present a formal test for the conjecture of balancing of the covariates (see table 8). For all the sub-set of data, we do not find any statistical differences across the border in relevant outcome variables.

2.4.5 Results and Robustness

We now generalize our previous test to the whole country, by employing the two-step procedure develop by Pinkovskiy [2013] and presented above. We start by showing that our $Toss\text{-}up_j$ index does not systematically predict differences in other covariates at the border - the extension of the test we already performed for Emilia Romagna and Lombardy. Results in this direction can be found in the appendix (see table 10). In particular, we test whether the conditional mean at the border of the covariates appears to be correlated with the toss-up score. Of all variables, only the size of the population appears to be significantly correlated; only when considering a bandwidth of 20km and only at the 10% level. So the test confirms the balancing assumption for the national sample.

We estimate our two-step estimator in three sub-samples, looking at bandwidth of 10,15, and 20 km from the border. We show these results in table 9, where we report every specification both with and without controls. As expected, the introduction of the controls does not substantially change the value of the estimated δ but it reduces standard errors. The specification confirms the results provided for Lombardy, where the more toss-up sides of the borders tend to have a larger share of voters acting strategic than their counterparts. In the next Section, we discuss more in detail the interpretation of the magnitude of these coefficients.

One concern with this procedure is that, the density of Municipalities close to the border may be particularly low in some specific borders. While this is not a big concern when considering a 20km mile bandwidth, it can be a problem with the 10km border. For instance, with the 10km bandwidth we were forced to drop two borders (Emilia-Romagna vs. Piemonte and Marche vs. Lazio) in the first-stage. We try to address this concern in two ways. We start by expanding the bandwidth in our first stage up to 30km from the border, in order to test the sensitivity of the model to number of observation in the first-stage. Our results are always in line with our main specification. Then, we implement a non-parametric bootstrap, in order to test

[18]Notice that the border regression is the closest you can get to the ideal RD. Here controlling for distance makes little sense, since the function of distance do not really discriminate between Municipalities that are closest or further from the border-since they all are on the border but rather between Municipalities whose centroid is closest or further from the border. Since we do not know the population distribution within the Municipality territory, this type of discrimination would be arbitrary.

whether our conclusions may be somehow driven by small sample bias. For every iteration i, we randomly draw with replacement N observations from the sample of N Municipalities within B km from the border. This is done completely independently for each of the 54 border-sides. Then, we use this sample to estimate the $\delta^{(i)}$ following the usual procedure. Using the empirical distribution of $\{\delta^{(i)}\}_{i=1}^{1000}$, we estimate confidence interval for the parameter. Again, all results are confirmed. All these tests can be found, as well as other robustness, in the appendix of the paper (see table 13).

Our estimates δ exploits the cross-sectional variation in the level of the toss-up index and $\Delta^{C-S}_{\%_{ij}}$, while creating a balance sample of homogenous areas. Alternatively, we may instead consider a different estimator, where we exploit only within border variation through the introduction of a border-specific fixed effect[19]. This estimator compares how, within the same border, differences in the tossup level affect the share of misaligned voters. If our primary specification is correctly specified, then we would expect this test to substantially confirm our previous results. This is what we find in the data. In the new fixed effects specification, the magnitude of δ is around 20% larger, and still highly significant. While larger, a formal test reveals that the two parameters are not statistically different between each other.

As a concluding robustness to our results, we present a simple one-step equivalent of our two step estimator. In particular, we pool together all the Municipalities that are within a distance B to the border for the whole 27 borders and we test whether being in a more toss-up Region affects the voters' strategic behavior. This is very similar to the methodology used by Black [1999] and Dube et al. [2010][20]. While our two-step estimator is more flexible in controlling for the differential effect of distance across different borders, their estimator is more parsimonious. If the relationship between distance and outcomes is not particularly heterogeneous across border-sides, we expect this specification to produce estimates close to the one in the two-step estimator. The results presented in table table 13 confirm our previous findings. If anything, these estimates seem to be even smaller than the one with the two-step estimator.

2.5 Discussion

In this section we show how we can recover from our previous estimates the extent of strategic voters. Furthermore, we discuss the role of strategic parties in our setting and in the interpretation of the results.

So far we estimated the effect of the level of electoral contestability on the size of misaligned voters[21]. Because citizens from the main parties face no strategic incentives, the total misaligned voters represents the amount of strategic voters within secondary coalitions. To get back an estimate for the total strategic

[19] The specification is $\tilde{\Delta}^{C-S}_{\%_j} = \alpha_{b(j)} + \delta Tossup_j + \beta \tilde{X}_j + \epsilon_j$, where $\alpha_{b(j)}$ is a border fixed effect, which is the same level of clustering of our standard errors.

[20] We estimate $\Delta^{C-S}_{\%_{ij}} = \alpha + \delta Toss - up_j + \rho(d_{ij}) + \epsilon_{ij}$ over the whole set of Municipalities within a bandwidth B to the border

[21] Voters that voted for a different Coalition in Senate relative to Congress

voters we need to scale back those estimates by the total fraction of votes obtained in the congress ballot by secondary coalitions. For instance, across the border between Emilia-Romagna and Lombardy we found a jump of around 1.5% of misaligned voters across the borders. Since the voters of non-top two Coalitions were only around one third of total voters in this area, this translates into an upper bound of voters that are strategic in the area slightly below 5%[22]. When considering the nationally pooled Regression Discontinuity, the intuition behind the result is the same, while the procedure to obtain them is slightly different. Consider a point estimate of our parameter δ of 5, which is coherent with the results in the border with 15km bandwidth. This is telling us that an increase of 1 percentage point in the closeness between the two parties implies an increase in misaligned votes of about 0.05 percentage points. Comparing a situation where we expect to have no voters being strategic (one of the contender Coalition is expected to lead the race by 20 percentage points), with one where the incentive is maximum (the two Coalitions are perfectly tied), we obtain an estimate of voters that are misaligned of about 1%. This again translates in a share of voters that are strategic of about 3%. Using the whole distribution of the coefficients estimated across all the specifications, our results are coherent with a share of strategic voters between 1% and 5%. The maximum estimate of total strategic voting is found in the regression discontinuity between Lombardy and Emilia Romagna. This is not at all surprising as these were the Regions where the relative electoral distance was greatest. Our pooled national regression is nevertheless robust to the exclusion of the Lombardy and Emilia Romagna border.

In section §3.2, we discuss how previous tests could potentially be estimating the joint equilibrium of strategic parties and strategic voters. Obviously such a critique could be turned against our estimates as well. Strategic parties and strategic voters go in the same direction: more tossup regions are more subject to strategicity. We argue that the natural experiment we use is less likely to suffer from this problem. In our setting, we can compare two equivalent votes. Both Chambers have the exact same institutional role, and the electoral system for the two is identical, with exactly the same choice set of coalitions available. Keeping the actions of political parties constant, a sincere voter would always vote for the same coalition in both Chambers. However, political parties can take actions to influence voters' behavior. In particular, a party chooses how to optimally distribute resources (advertisement, campaign funds,...) and how to allocate candidates across districts.

By looking at the difference in votes across the two Chambers at Municipal level, we implicitly impose a fixed effect at the municipal level. Therefore, all the actions that can be taken by parties to influence voters at city level, such as advertising or campaigning by one of the top coalition leaders, will influence both Chambers in equal extent and therefore its effect will be differentiated out by our estimator.

The only concern left for strategic parties is the assignment of candidates across Regions and across the two Chambers. In particular, the top two coalitions could place the highest quality candidates asymmetrically across Chambers and Regions, with the highest quality politicians being systematically placed in the Senate

[22]This is of course under the assumption that the share of potentially strategic voters is uniformly distributed across parties. A common assumption in the literature.

in toss-up Regions. It is worth stressing that differences in quality across Regions is not a threat to our identification, but rather differences in relative quality between the two Chambers across Regions. We argue that this is not a concern in our setting. In Italy's 2013 election voters choose a party, each of which presents a long, closed list of candidates, rather than a specific politician, like in uni-nominal systems. In order to be influenced by the actual quality of politicians, voters should both be able to understand the identity of the marginal candidates for each party and care about the relative quality of the elected officials in the difference across the two regions between the chamber and senate lists. We argue that the ability of assessing the marginal candidate was greatly impaired with the "Pig Crap" electoral system. This is not a mere matter of scholar judgment but is also certified by the official motivation of the the Italian Constitutional Court. In fact, in December 2013 the Italian Constitutional Court ruled unconstitutional the "Pig Crap" law precisely because of the "unknowability"[23]of the candidates. According to their ruling, it was impossible for a voter to understand who was the candidate they were voting for. In particular, some features are particular relevant, such as : (1) Candidates are presented in very long closed lists with no possibility of preference selection; (2) same top candidates were in multiple districts with ex post party selection of the seat 3) large premiums. [24]

One could still argue that voters would not vote fully rationally for their unknowable marginal candidate but rather for the top few on the list or a weighted average of the first candidates. These cases do not worry us as the same people are top candidates in neighboring regions or in all regions (e.g. Berlusconi for the center right both in Emilia and Lombardy's Senate), and because we collect and show observable quality metrics across regions and parties in the two main tossup regions (Emilia-Romagna and Lombardia). In a separate technical appendix we detail the behavioral rules that would be necessary for strategic parties to still be the driver of the estimated 5%.

2.6 Strategic parties or strategic voters?

We have shown how our methodology gives us very low estimates for strategic voting. This is in contrast with recent literature on the extent of strategic voting. In this section we review such literature and discuss

[23]"né con altri caratterizzati da circoscrizioni elettorali di dimensioni territorialmente ridotte, nelle quali il numero dei candidati da eleggere sia talmente esiguo da garantire l'effettiva conoscibilità degli stessi e con essa l'effettività della scelta e la libertà del voto (al pari di quanto accade nel caso dei collegi uninominali)". Authors' translation "[the unconstitutional law cannot be defended by comparing it with other] systems characterized by electoral seats that are small, in which the number of candidates is so limited in the short list to warranty the effective Knowability of the candidates and with it the freedom of choice and vote."

[24]Understanding who is the marginal candidate is not a trivial exercise. First, you must know the total number of Senators elected in the Region. This number is updated overtime by the Government to reflect changes in population. In particular, the number of Senators elected in Lombardy and Emilia-Romagna changed in 2013 for the first time in more than ten years, increasing for both by about 5%. Once you know the total number of seats assigned, you need to form expectations about the shares of votes received by each coalition and also about the shares of votes received by each party within each coalition (e.g. Center-Right has more than five parties in the coalition). Any party presents his own list of candidates and seats are first assigned to coalitions, and then parties split the seats of the coalition. Two factors make the problem even more complex: (a) The problem is naturally discrete and there are relatively small parties.Therefore even a small changes in vote by a small parties may determine a jump in the number of elected officer by the party; (b) It is not uncommon that some elected officers reject the election because of incompatibility with other roles or because the same candidate was elected in another electoral district: this creates all another level of uncertainty in the problem. On top of all this for tossup regions it is effectively impossible to determine who the marginal candidate is.. precisely because they are tossups. So the premium could go to any of the coalitions.

the discrepancy.

Two recent papers try to move beyond indirect tests on the existence of strategic voting to provide estimates of its relevance. Kawai and Watanabe [2013b] use a structural model to conclude that as much as 85%, and at least 60% of voters are strategic. A result that we find quite big compared to previous surveys. Spenkuch [2012] estimates the number of strategic voting somewhere above 30%. Relative to this literature our contribution is to show both theoretically and numerically how these high estimates might easily be due to incorrect identification due to parties acting strategically and provide much lower estimates from a cleaner natural experiment (5%). Besides these two very recent works that test the extent of strategic voting, the richness of the empirical literature confirms the importance of the topic. Laboratory experiments support tactical behavior (Duffy and Tavits [2008a]) and pivotal voter models (Levine and Palfrey [2007a]), but it remains to be seen whether this is something relevant in big scale elections and when people are casting real votes. Alvarez and Nagler [2000] and Ganser and Veuger [2012] use surveys to argue that strategic voting occurs also in real elections, but their estimates vary depending on the type of surveys used. Indirect tests of strategic voting that do not estimate its extent are provided by Coate et al. [2008a] and Fujiwara [2011]. Coate et al. [2008a], observe very big winning margins that are inconsistent with a positive cost of voting and the pivotal model. Indeed with positive cost of voting and pivotal behavior some voters should abstain from going to the polls in equilibrium. This evidence is inconsistent with a large proportion of voters being strategic, if we assume that the cost of voting is substantial. Fujiwara [2011] instead finds indirect support for the existence of strategic voting by testing Duverger's law (Duverger [1959]).

Strategic voters vote differently when subject to different pivotality likelihoods (Myerson and Weber [1993b]). Instead, parties behave strategically when in tossup districts they choose candidates that are more appealing to marginal voters or otherwise alter their behavior (marketing and campaigning) in response to the likelihood that the district will be contested. It is clear that strategic parties and strategic voters go hand in hand. When a district is too close to call, the main parties have incentives to candidate their more appealing public figures and voters are more likely to be pivotal and therefore behave strategically. In this context and without further identification refinement, an estimate of the extent of strategic voting combines together the strategicity effect of both voters and parties. We show that controlling for all unobservable and observable district and candidate characteristics does not help isolating the effect of strategic voters. In particular, we construct a very simple example where an identification strategy comparing candidates' votes with lists' votes would overestimate strategic voters by 25 percentage points[25]. We also show how failing to explicitly model and account for strategic parties in a structural estimate can lead to even larger overestimates. [26] Through this analysis, we highlight that a test isolating the extent of strategic voting needs to (a) compare two identical votes; and (b) control that other strategic players are not able to take actions that are asymmetric across the two ballots.

[25]Precisely the difference between our estimates and Spenkuch [2012]

[26]The difference between our estimates and Kawai and Watanabe [2013a] is 60-78 percentage points.

Spenkuch [2012] compares "list votes" awarded in a proportional way nationally with "uninominal votes" awarded in a "first past the post" way at the same administrative unit. Because the two votes have different electoral rules (proportional and uni-nominal), big parties have an extra incentive to place the best candidates in the uni-nominal ballot and the opposite incentives hold for small parties. Therefore even fully sincere voters that care about candidates characteristics would desert the small parties for the big ones. Importantly, controlling for observable or unobservable characteristics of candidates or districts would not eliminate this concern. Consider comparing two different votes across different units of observation, where the sets of candidates or characteristics of the district are constant. In particular, Spenkuch [2012]'s specification is:

$$v^{C}_{k,r} = \chi_k + \lambda v^{L}_{k,r} + \varepsilon_{k,r}$$

Where $v^{C}_{k,r}$ is the percentage votes for the candidate in the uninominal election in district k and precinct r , $v^{L}_{k,r}$ the percentage votes for the list of the candidate in the national election and χ_k a fixed effect for the district-candidate.

If the two votes are not identical, regressing one vote against the other will attribute proportional variation across units of observations within a district to strategic voters and only absolute average changes to fixed effects (candidate or districts). In other words, it assumes that better candidates will be rewarded for their superior quality with the same percentage points effects within district, rather than proportionally. This is a very strong assumption that is probably implausible as the following example will clarify.

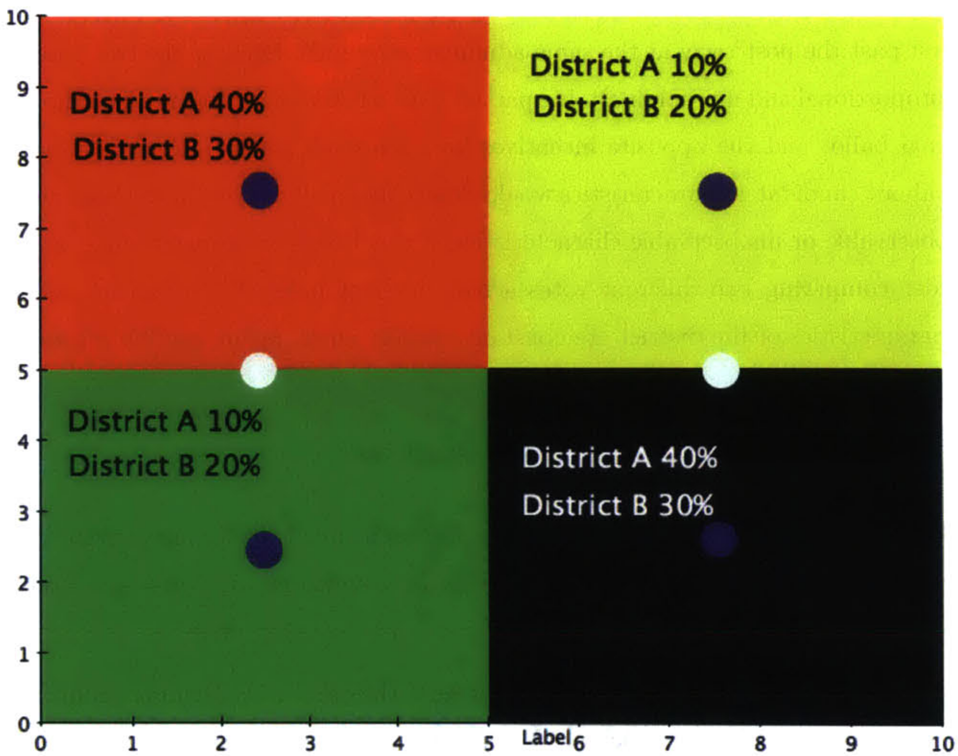


Figure 2-3: Counter example

Example 31. Assume that all voters are sincere, so that there are no strategic voters at all, and that their preferences can be represented on a bi-dimensional ideological plane as in Figure 2-3.[27] The voters are sincere and vote whomever party (for the list election) or candidate (for the uninominal first past the post mandate) is closer to their ideology. The ideology of each voter is represented by her (x, y) coordinates in the plane. The density of voters is uniform over each square and the total mass of voters in each square is as represented in Figure 2-3. We have only two types of districts $\{1, 2\}$ and within the districts two types of sub-districts $\{A, B\}$ with some small measurement noise. Now let us assume that for the uni-nominal elections the NE and the SW parties strategically choose their candidates to be in the white dots position. Then the electoral support would be as represented in Table 2.3. Now for simplicity assume that there are a total 100 sub-districts of type A and 100 sub-districts of type B, with 50 of each type in each district $\{1, 2\}$. Creating a simulated data, with the same averages per groups as previously specified, and running the same regression as Spenkuch [2012][28] gives right away a 0.75 as the statistically significant estimate of sincere voters. Obviously by choosing arbitrarily the numbers we can construct examples of any magnitude.[29]

[27] Say that the x axis represents economic ideology and the y axis social ideology

[28] Spenkuch's specification is:

$$v^C_{k,r} = \chi_k + \lambda v^L_{k,r} + \varepsilon_{k,r}$$

Where $v^C_{k,r}$ are the votes for the candidate in the uni-nominal election in district k and precinct r , $v^L_{k,r}$ the votes for the list of the candidate in the national election and χ_k a fixed effect for the district.

[29] In particular the 25% estimate of strategic voting does not depend on the initial values chosen (that can be arbitrary) for the sub-district nor the districts. It only depends on the fact that the distribution of voters within the quadrant implies

Table 2.3: Votes in each Election and district when all voters are sincere

	Red	Yellow	Green	Black
List Votes Sub-District A	40%	10%	10%	40%
Uninominal Sub-District A	42.5%	7.5%	7.5%	42.5%
List Votes Sub-District B	30%	20%	20%	30%
Uninominal Sub-District B	35%	15%	15%	35%

Nevertheless, structural estimates that do not fully model the strategic incentives of parties may still produce biased estimates. For instance, Kawai and Watanabe [2013a] model structurally the preferences of voters as depending on a set of ideological and candidates' quality parameters, but they do not directly model parties' incentive problem. Then, they use data on uninominal vote from the last Japanese election to estimate the parameter of interest. [30]

The implicit assumption behind not modeling the strategic incentives of parties is that parties are not strategic in their choice of candidates - conditional on a few observable characteristics (party affiliation, an indicator for the home municipality of the candidate, and whether the candidate has held an office before). After having estimated the structural parameters of the model conditional on the candidates' and district characteristics, they infer the level of strategic voting by observing the degree of variation on predicted close (ideally identical) Municipalities. Under the assumption of parties not being strategic, there is a unique equilibrium for a set of parameters' values if voters are fully sincere. To the extent that similar districts have different result such multiplicity can be attributed to strategic voters. Their size is then backed up by considering extreme values beliefs potentially different across districts and identical within. The problem is twofold: if parties are truly strategic, they are likely to place their candidates based on many more observable and unobservable characteristics. More importantly, strategic parties can generate disparate outcomes for identical districts even under fully sincere voters and unique equilibrium. For instance, if there are substitution forces that lead parties to strategically choose their candidates to exploit initial advantages or disregard lost battle grounds. Therefore, even under fully sincere voters the inequality estimator would load the multiplicity in outcomes generated by diverse single equilibrium due to strategic parties as strategic voters induced multiple equilibria.[31]

Our identification strategy does not suffer from the same problems. In our setting we can compare two equivalent votes. Both Chambers have the exact same institutional role, and the electoral system for the two is identical, with exactly the same choice set of coalitions available. Keeping the actions of political parties

that 25% of them would vote sincerely for a candidate with a closer ideology. Strategic parties get fully attributed to strategic voting.

[30]In particular the utility function they assume is: $u\left(x_k, z_{il}, \theta^{PREF}\right) = -\left(\theta^{ID} x_k - \theta^{POS} z_i^{POS}\right)^2 + \theta^{QLTY} z_{il}^{QLTY}$. The z_i characteristics of candidates include the party affiliation, an indicator for the home municipality of the candidate, and whether the candidate has held an office before.

[31]As a side-note, there would be another conceptual issue if one were to explicitly model strategic parties. Because there are complementarity in beliefs over a district and the choice of good candidates, the resulting bounds would be even larger.

constant, in this context a sincere voter would always vote for the same coalition in both Chambers. By looking at the difference in votes across the two Chambers at Municipal level, we implicitly impose a fixed effect at Municipal level. Therefore, all the actions that can be taken by parties to influence voters at city level, such as advertising or campaigning by one of the top coalition leaders, will influence both Chambers in equal extent and therefore its effect will be differentiated out by our estimator. On the other hand, one could argue that the average of polls is only a noise measure of the underlying beliefs of voters. If voters' beliefs are dispersed around the electoral poll mean as they probably are, then it would be more conservative to interpret our estimates as a lower bound. In fact, strategic voters whose beliefs are far away from the mean will not behave strategically leading to an underestimation of the true effect.

2.7 Conclusions

We showed that at most 1.5% of voters were misaligned during the 2013 Italian Elections. This corresponds to a total of under 5% of strategic voters across the whole population. In the context of the recent literature on the extent of strategic voting, our estimate is considerably lower than the share determined by other works (at least 30% and up to 85%), but still substantial and statistically significant. We discussed how our lower estimates can be reconciled with the previous literature. In particular, we believe that our lower estimates reflect a neater separation of strategic parties from strategic voters. We built simple examples to show numerically how previous identification strategies would confound these two concepts and therefore exaggerate the extent of strategic voting. In contrast, we use the peculiar features of the Italian electoral law and institutional system to better separate the two in reduced form.

We also build a model to justify our empirical strategy and the use of regression discontinuity. To build our geographical regression discontinuity estimates we use novel techniques (Pinkovskiy [2013]) to pool together multiple geographical regression discontinuities and assess its robustness with bootstrap methods. As a further robustness we also try alternative methods used in the literature to pool together cross border RD (Dube et al. [2010]). Lastly, we discussed why it is prudent to interpret our estimates as a lower bound since the underlying function relating the closeness of parties in percentage points to the percent of strategic voters could well be non linear. Future avenues of research would ideally assess the external validity of our results and provide an even cleaner separation of strategic parties and strategic voters. Our contribution in the debate has been to highlight the difference between strategic voting and strategic parties, while ascertaining much lower effects, and showing how easily these two different concepts can be confounded even in apparently sound identification strategies.

Chapter 3

The Costs of Conflict of Interest: Stock Market Performance, Excess Returns and Political Proximity

Essays in Political Economics and Information Acquisition

by

Giovanni Reggiani

M.Sc. Economics and Social Sciences, Bocconi University (2010)
B.A. Economics and Social Sciences, Bocconi University (2008)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Author . . . .
Department of Economics
May 15, 2016

Accepted by . . . .
Ricardo J. Caballero
Ford International Professor of Economics
Chairman, Departmental Committee on Graduate Studies

3.1 Introduction:

When owners of big firms enter politics and gain control over the legislature there is a substantial risk of market distortions. An estimate of the market distortions brought about by conflict of interest is particularly important at a time when many tycoons are entering politics in the developed world: Berlusconi has been in politics for over 20 years, Donald Trump is now running for the Republican nomination, and Gina Hope Rinehart[1], has been rumored to run for prime minister.

We perform a case study of Berlusconi's rule over Italy and estimate the effects of the distortions due to conflict of interest. Our identification strategy is based on the Efficient Market Hypothesis (from now on EMH), the assumption that information is embedded quickly enough in stock prices. Under this assumption, surprising electoral outcomes should not affect Berlusconi's firms more than other firms in the same sector unless there are substantial conflicts of interest. We use the difference between polls and electoral outcomes to measure the surprise of each election. We infer from the observation of abnormal returns for Berlusconi's firms around elections the market price for political distortions. We find large effects: 6% increase of market capitalization per percentage point of positive surprise.

A possible concern is that the abnormal returns of Berlusconi's firms do not reflect distortions due to his entry in politics but rather the market concern that he might be jailed if not in office and the belief that his leadership brings significant value for his firms. There is ample evidence of negative abnormal returns on stocks when CEOs step down or leaders pass control over to their relatives Perez-Gonzalez [2002]. We find that this is unlikely to be the case because the two firms that are outside the Italian government's reach, Tele5 and Mediolanum bank, do not experience positive abnormal returns for positive electoral surprises.[2]

Although we can be confident that there are no issues of reverse causality - i.e. the surprising electoral outcomes are not caused by the abnormal returns of his stocks - we cannot disentangle whether the firms' extra returns are due to Berlusconi bending markets and the law in his favor or to the markets pricing the risk of revenge-legislation by his political opponents. Anecdotal evidence[3] and a recent paper by DellaVigna et al. [2013] suggests the former. Regardless of the channel, the estimated effect can be interpreted as "distortions" due to conflict of interest and entry in politics. The magnitude of the effects - 6% per percentage point of surprise on Mediaset and 4% on Mondadori - and their significance suggests that conflict of interest laws preventing people with large interests in big firms from running for office could be efficient.

The structure of the paper is the following: in section §3.2 we detail how the present work relates to previous literature and its contribution, in section §3.3 we describe the background and discuss the data, in section §3.4

[1] The richest Australian and owner of most Australian mines

[2] In addition, at the time, Berlusconi was not in any operative role within his companies.

[3] There are many claims and ample anecdotal evidence that Berlusconi used his power to benefit his companies. Some of the most controversial decisions included: subsidizing the adoption of the cable technology, expelling profitable and successful journalists from the state owned RAI, and asking entrepreneurs not to run advertisements on newspapers criticizing his government.

we illustrate the empirical strategy and the details of the relevant efficient market hypothesis. Section 3.5 contains the main results, section §3.6 provides some robustness checks and section §3.7 concludes.

3.2 Literature Review:

Previous literature has documented the causal effect of political connections on stock market performance only for developing countries. Research focusing on developed countries fails to find proof of similar effects - suggesting that the checks and balances in place in developed countries might be sufficient without the need to explicitly ban tycoons from running for office. Our paper fits in between these findings: it shows how even for developed countries conflict of interest can cause significant market distortions.

There is ample evidence that in developing non-democracit countries politically connected firms experience abnormal returns (Fisman [2001], Ferguson and Voth [2008], Faccio et al. [2006]). For developed countries, we have evidence on soft political connections (donations and board sitting) being associated with extra returns (Goldman et al. [2009], Jayachandran [2006]). But it is harder to argue that these associations in the developed world are necessarily due to crony capitalism and politicians favoring connected firms. Indeed, a correlation between abnormal stock returns and political victory is not evidence of market distortions. For example, imagine that a party is relatively more prone to use war as a diplomatic tool, and that a defense firm, anticipating more orders from its victory, elects a party representative to its board in order to better tailor its products to the party's ambitions. Suppose that the firm is listed in the stock market. In case of victory, we would probably observe extra returns but this would not be a sign of corruption or distortions.

Indeed, Fisman et al. [2006] failed to find any evidence that the personal ties to Dick Cheney were beneficial to his former firm or to those closely connected to him. Fisman et al. [2006] argue that this is evidence of how well the checks and balances work in a developed and free society like the US. Faccio [2006] finds significant effects for personal ties using a sample of developing and developed countries but the estimated effects are much smaller than what we find (4% total vs. 6% per percentage of surprise) and it is not clear how much of the results are due to the developing countries subsample. In developing countries, the channels for the extra returns are driven by procurement contracts (Goldman et al. [2009])), and easier lending (both Li et al. [2008] and Claessens et al. [2008]).

Although our paper is related to Fisman et al. [2006] and their case study on Dick Cheney and Suharto (Fisman [2001]), our strategy is quite different. Unlike these, we do not use surprises on health; rather we are able to estimate the electoral surprise using the difference between weighted averages of polls and electoral outcomes.

This allows us to have a more precise idea of the time window when the surprise arrives, and a more precise quantitative estimate of the surprise.

The test and estimates are obtained under the hypothesis of informational market efficiency. We detail the empirical implication of different hypothesis on efficiency in section §3.4. Suffice it to say here that we can also use the structure of the paper to carry out the reverse exercise; i.e. assume that Berlusconi will favor his firms in power[4] and carry out a test of the degree of informational efficiency in financial markets. This would relate our paper to the financial literature that tests informational efficiency - originated by Fama et al. [1969]. Under the political assumption of conflict of interest, we find that markets are more than semi-strongly efficient, which is consistent with their findings.

More importantly our paper is directly complementary to recent work by DellaVigna et al. [2013]. In this work the authors use precise data on Media advertisement in Italy, viewers surveys and changes in government to estimate the indirect bribing that occurs when Berlusconi is in power. An important theoretical insight of the paper is that in the presence of conflict of interest we should worry not only about direct corruption and direct favoritism by the ruler, but also about indirect market distortions. Other firms might direct their business towards the firm connected with the politican in fear of retaliatory government action. Our work and DellaVigna et al. [2013] complement each other in the following way: through stock markets and EMH we can value all conflicts of interest and therefore also provide a more accurate estimate of how much of the increased revenue translates into higher profits and capitalization. In addition, DellaVigna et al. [2013] document that the indirect bribing channel is probably the main driver of the value of conflict of interest in developed countries. Our estimates are consistent with those of DellaVigna et al. [2013]. DellaVigna et al. [2013] find extra profits for Mediaset equivalent to 20% of the market capitalization over 9 years. We find that a value of 6% extra returns per percentage point of electoral surprise.

3.3 Background and Data:

3.3.1 *Berlusconi and his companies*:

Mr. Berlusconi is among the wealthiest men in the world. He made his fortune constructing residential units outside Milan and, later, setting up the first non-state-owned TV company in Italy. In the early years, his political connections were of paramount importance in preserving and expanding his TV stations. Some observers claim that his subsequent entry in politics was triggered by the political scandal of Tangentopoli and a fear of losing his status of political protégé.

He entered politics in 1993, ran for prime minister (1994) and won the elections. His government did not last long and he lost the 1996 elections against Romano Prodi. He managed to return to power in 2001 and held it until 2006 when Romano Prodi defeated him for the second time. After a short term as head of the opposition, he won again snap elections in 2008 and held power until late 2011 when a political crisis triggered by sex scandals coupled with the crises of Italy's public finances forced him to resign.

[4]Or equivalently that his detractors will damage his firms when not in power.

Ticker	Name	Market Share	Sector	Under IT Gov Regulation Control
MS	Medisaset	40%	Media- Italy	Yes
TL5	Tele Cinco	44%	Media- Spain	No
MON	Mondadori	27.1%	Publishing	Partly
MED	Mediolanum	5%	Retail Banking	No

Table 3.1: Summary of Berlusconi's Firms and their regulatory bodies.

When he entered politics in 1994 he already owned the listed companies that he controls today: Mediaset (from now on MS), Tele5 (TL5), Mondadori (MON) and Mediolanum (MED). Some of these companies were already listed (e.g. MON), while others were listed during his political tenure (MS and TL5, 1996 and 2004 respectively). There are two relevant metrics for our study on which these companies differ:

- whether the Italian prime minister can rewrite the regulation for the industry in which the firms operate
- the market share

We summarize this facts in table 3.1:

The firms businesses can be summarized as follows:

- MS controls approximately half of the Italian generalist TV market. The other half is controlled by the state owned company RAI. Not only does the prime minister have access to regulation for this sector but he/she also appoints the board of RAI, controlling therefore indirectly the main competitor's actions.
- MON is a publishing company. The government can set subsidies for publishing companies and partially determine the regulatory boards.
- MED has less than 5% market share in retail banking. Its regulation is subject to Bank of Italy, a body over which the government has no control.

For all these companies, I collect data from the Bloomberg Platform on daily Total returns (correcting for dividends payouts), volatility, and their industry associated indexes at the European level (MXMD0EU for the media/publishing companies and BEFINC for MED). I also download the same data for the Italian Stock index (FTSEMIB).

3.3.2 *Elections:*

Since Berlusconi's companies have been listed, there have been 8 relevant elections for our study: two European Union elections (2004, 2009), three national political elections (2001, 2006, 2008), and two regional elections (2005, 2010) and one local election with national effects (2011, Milan mayor).

For our identification strategy, we limit to post 2000 elections to benefit from a "silence of polls" law which allows us to construct a surprise metric. In the following we explain why it is appropriate to include elections other than the national political elections and the criterion for including only certain regions and the city of Milan in the regional and local elections while we postpone the discussion of the "silence law" to the next subsection.

Clearly national political elections are highly relevant since they determine the prime minister and consequently control over RAI and regulation and subsidies affecting all companies except MED[5]. Nevertheless one might wonder why we include the rest of the elections since they have no direct effect on the offices and regulatory seats that affect the profitability of the companies.

In Italy governments are formed by coalitions of parties which are very sensitive to the oscillations of their support given by the outcomes of these elections. [6]In a nutshell, due to a fragmented parliamentary system, European and regional elections carry a national political meaning that goes well beyond their institutional one. Politically, what matters is the results in a few big and too close to call regions (Piedmont, Lazio, Puglia and very recently Campania). Indeed the rest of the regions are either too small or the outcome is already known in advance and the turnout less reflective of the support for the national government.

We also include the 2011 local election of Milan because it was an instance of high politicization of a local election. After a disastrous round of local elections and in the midst of a sex scandal with an underage prostitute, Berlusconi put his face on the ballot of a city that represented for a long time the heart of his party support. He asked to consider that election as a referendum on his person and campaigned fiercely for the city.

For all these elections I collected data on the electoral outcome for Berlusconi's coalition and his opponents directly from the official ministerial website.

3.3.3 *The Italian law on polls:*

In many countries it is legal to publish polls up to voting day. In Italy, instead, a recent law [7] prohibits the publishing or diffusion of electoral polls during the 15 days preceding an election. The law also establishes that all results must be published on an official government website together with the indication of the payer, methodology used to form the sample, sample size, date of the poll, and questions asked.

If this particular law was not in place, we could have measured the electoral surprise exactly by looking at the difference between the polls on the day before the elections and the actual outcome. Nevertheless as long

[5]Remember MED is the ticker for Mediolanum not Mediaset

[6]For instance, after the negative outcomes of the 1999 EU and 2000 regional elections, D'Alema (then prime minister) was forced to resign because of the consecutive defeats of his party. Similarly Berlusconi was forced by his allies to dramatically reshuffle his cabinet, his policy, and his economics minister after the loss of the 2004 EU and 2005 regional elections.

[7]n.28 of the Feb 23, 2000. The ratio of the law emanated during the last months in office of the center left government was preventing the hectic campaigner Berlusconi from gaining too much support through last minute polls. Nevertheless the law has remained in place also during Berlusconi's rule

as either a significant fraction of investors do not have access to private polls or the polls evolve continuously our surprise metric will be correlated with the true surprise. Moreover, it will be an upper bound for the true surprise, and therefore our estimates will be a lower bound for the distortions due to conflict of interest.

I constructed the surprise metric for the electoral outcome using the average of the polls commissioned by the most widespread and trustworthy newspapers in Italy (Il Corriere della Sera, and La Repubblica) and published just before the silence zone (15 days before the elections). The list of all the survey companies used is: Termometro politico, IPR marketing, SWG espresso, abacus/skytg24, and Ispo. The sample sizes of these polls were always of at least 1000 and the sampling method stratified by the relevant metrics[8].

To compute the surprise metrics I considered the difference in the difference of support between Berlusconi's coalition and the center-left coalition between the poll and the actual outcome.

$$Surprise = (BerlusconiElectoralsupport - CenterLeftElectoralSupport)$$

$$-(BerlusconiPollsupport - CenterLeftpollSupport)$$

A higher value of this variable implies a bigger positive surprise for Berlusconi. Figure 3-1 shows the cumulative abnormal returns through time and the election dates with their corresponding surprise metric. The dotted lines are political elections, the dashed EU elections and the remaining are regional or local elections.

3.4 Empirical Strategy:

We wish to test the hypothesis that conflict of interest during Berlusconi's political era has influenced the profitability of his companies and estimate the effects of these distortions. In the next subsection we review the EMH in its three formulations. We then show what each of these formulations implies for our test of distortions and detail our H_0 and H_1 hypothesis. In 3.4.2, we focus on Semi-strong EMH and detail what its empirical predictions would be. In 3.4.3 we show that if markets are more than Semi-strongly efficient we would not get any effects from a simple regression and that Cumulative Abnormal Returns would be the correct way to estimate the distortions. In the last section we conclude with some caveats before introducing the results.

[8] Sex, residence in rural or urban areas, age etc

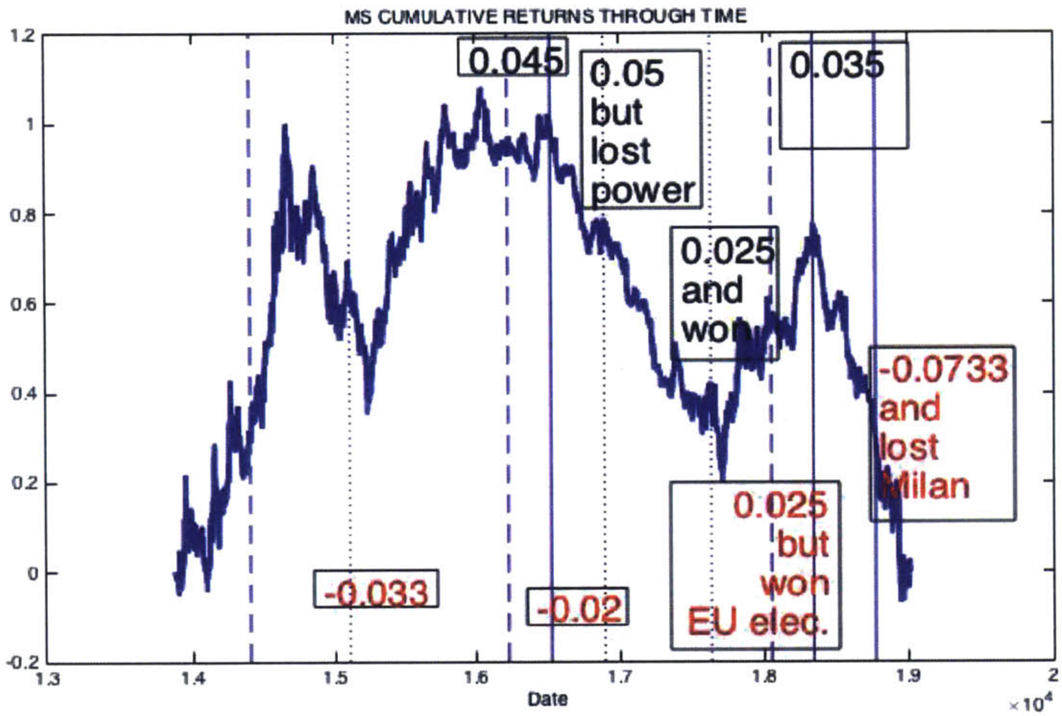


Figure 3-1: Cumulative Residuals for MS and Election dates.

3.4.1 *The finance theory in the estimation strategy:*

The Efficiency Market Hypothesis specifies the degree to which financial markets prices embed information. The definition of the EMH distinguish between Strong, Semi-Strong and Weak EMH.

- Strong EMH (SEMH) assumes that asset markets incorporate in their pricing all available information – both public and private.
- Semi strong EMH (SSEMH) assumes that markets incorporate all public information but not necessarily private information.
- the Weak EMH (WEMH) only assumes that markets incorporate the past history of assets' prices in current prices.[9]

As we reviewed before, there is ample evidence that, while in power, Berlusconi benefited his companies (DellaVigna et al. [2013]); and in particular his media company: MS. If this was true, how would the stock market react under the different EMHs to a surprising outcome in his favor?

[9]That is that there is no money to be made by looking at the past history of assets and their correlations.

	WEMH	SSEMH	SEMH
Political Distortion	No Reaction	MS ↑on results' day	MS ↑before results' day
No Political Distortion	No Reaction	No Reaction	No Reaction

Table 3.2: Summary of MS performance for different Informational and distortion scenarios.

Consider an investor that knows that Berlusconi will benefit his company if in power and that has access to the news that Berlusconi is about to win. He will purchase a lot of stocks from MS. If enough stocks are purchased, the price of the stock will rise. The EMH is concerned with whether that variation will occur. The outcome is summarized in the table 3.2.

If markets are only WEMH then those investors will be too few to affect the price and the MS stock price won't even reflect the public news of a surprisingly good outcome for Berlusconi. Instead if the markets are SSEMH then they will react with price increases on the date of the election - i.e. when the surprising outcome becomes public news - but not before. Lastly if markets are more than SSEMH then people with access to private information (new secret polls during the silence period) will also purchase enough stocks so that the price rise will occur even before the election day.

We want to test empirically whether Berlusconi distorts the markets favoring MS, but notice that given our summary table above we will only be able to test the joint hypothesis that markets are at least SSEMH consistent AND that Berlusconi will favor his firm. Therefore we assume that markets are SSEMH and our null hypothesis becomes: [10]

$\mathbf{H_0}$: Berlusconi will not favor his firm

vs.

$\mathbf{H_1}$: Berlusconi will favor his firms

A benefit of this asset pricing theoretical approach is that one can use the contrary assumption to test informational efficiency in financial markets. Rather than assuming that markets are strongly efficient, we could assume that Berlusconi will benefit his firms[11] and test whether markets are at least semi strongly efficient.

[10] Without assuming anything our hypothesis would be:

$\mathbf{H_0}$: Either Berlusconi will not favor his firm OR financial markets are not at least semi-strongly efficient

$\mathbf{H_1}$: Berlusconi will favor his firms AND financial markets are at least semi strongly efficient.

[11] Based on political economics arguments, history, character analysis, and biography.

3.4.2 *The Twist of the EMH: why we need a surprise metric*

Even before thinking about the required market efficiency, one may want to run a regression of the returns of Berlusconi's stocks (at the electoral date) on his electoral results (percentage points of support). These kind of regressions should find no effect regardless of what the level of informational efficiency is.

This might seem somewhat surprising, and is the reason why we need to have a metric of surprise in our analysis so let us detail why this is.

There are three cases to consider:

- If markets are only Weakly efficient then they would not react to the public news of victory;
- If markets are Semi Strong efficient then we should expect them to embed all public information at any point in time and not just the electoral results on the election day. So, to the extent that the electoral outcome is not a surprise then markets should not react. Of course, sometimes markets will be positively surprised, sometimes negatively, but since we do not observe their expectations we should not find any correlation between the idiosyncratic return on Berlusconi's companies and his electoral performance.
- If markets are Strongly efficient then we should observe no effect because of the previous point and the fact that previous information from secret polls is already embedded in prices before the election date.

This is the twist of the EMH assumption, a twist that can be tested by running precisely the standard regression of returns on electoral outcomes: we should find no effect. In conclusion, to find results we need to have at least SSEMH and a measure, even noisy, of the markets surprise to the electoral outcome. As detailed in section 3.3.3, I use the difference between the last allowed poll and the electoral outcome for such a metric. [12]

Therefore we can run the regression:

$$Tot\_Ret\_Stock_t = \alpha + \beta_{mkt} \cdot Tot\_Ret\_NatioanlINDEX_t + \beta_{Ind} \cdot Tot\_Ret\_IndustryINDEX_t + \varepsilon_t \quad (3.1)$$

Where $Tot\_Ret\_Stock_t$, is the daily return for his stock, $Tot\_Ret\_NatioanlINDEX_t$ represents the return for the national stock index (the equivalent of the SP500 for the US), and $Tot\_Ret\_IndustryINDEX_t$ is the daily return for the Industry index to which the company belongs.

[12] Of course in between the last poll and the election day new events will occur and these will be priced in by the market but because these are probably not correlated and do not completely over-ride the prior given by the last poll, this metric constitutes a valid, although noisy, signal of the underlying market surprise.

We then estimate the residuals, and focus on the part of the daily returns that are not due to general macroeconomic news or to Industry specific news. Notice though that to the extent that Berlusconi's victory is beneficial to the whole Media industry we are underestimating its effect.

Then we run the regression

$$\hat{\varepsilon}_t = \alpha + \beta \cdot Surprise_t \tag{3.2}$$

where

$$\hat{\varepsilon}_t = Tot\_Ret\_Stock_t - \alpha - \hat{\beta}_{mkt} \cdot Tot\_Ret\_NatioanlINDEX_t - \hat{\beta}_{Ind} \cdot Tot\_Ret\_IndustryINDEX_t \tag{3.3}$$

If the financial markets are consistent with the SSEMH, we should observe a significant effect on the coefficient of this regression although it would be biased towards 0 because of the noisy surprise metric, providing a lower bound for the true effect. Our theoretical assumption was that financial markets are at least semi strongly efficient. The next section shows how to construct this test if markets are more than SSEMH.

Regardless of the informational efficiency remember that the twist of EMH is that we should observe no results if we do not use the surprise metric:

$$\hat{\varepsilon}_t = \alpha + \gamma \cdot BerlusconiPoliticalOutcome_t \tag{3.4}$$

So we should find that γ is not significantly different from 0.

3.4.3 *SEMH : Event Study and Cumulative Residuals:*

While finding statistically significant results in the previous section would be encouraging, failing to do so should not automatically lead us to accept the null H_0.

If markets are strongly efficient, rather than semi strongly, traders could place orders in the stock market based on private not-publishable polls. The implication (see table 3.2) is higher daily returns before the election and markets reacting less or not at all when the official results come out - because that information has already been priced in.

Event studies pioneered by Fama et al. [1969] are precisely the way to test this hypothesis. Event studies test if the underlying event was anticipated by the stock market in the weeks before its occurrence.

Consider an event that occurs at time T (e.g. an election). We could then calculate the cumulative returns as

$$CR(t) = \sum_{i=T-15}^{t} R_j(i) \tag{3.5}$$

where $R_j(i)$ is the total daily return for stock j at time i, and therefore $CR(t)$ are the cumulative returns up to time t from the starting point $T - 15$. Notice that the effects of an increase in the returns is carried over time so that this function at time T+1 ($CR(T+1)$), captures the whole gains from the event occurring independently of which of SSMH or the SMH hold.

Choosing the time T as the electoral dates and R as the total returns on Berlusconi's company, we have another strategy to test our hypothesis and estimate a lower bound for the conflict of interest. We can look at the graph of the Cumulative returns function and if, for positive surprises, it is increasing up to the date T+1 and then flattens, this would be strong evidence of financial markets consistent with the EMH and Berlusconi in power being beneficial to his firms. Moreover, the value of the function CR(T+1) would represent the estimate of the total value of conflict of interest.

Since many pieces of private and public news hit the market that affect the whole media sector (e.g. industry specific information) or other Italian companies (e.g. macro news) we can improve our previous specification by looking at the residuals rather than the returns.

$$CRES(t) = \sum_{i=T-15}^{t} \hat{\varepsilon}_j(i) \tag{3.6}$$

If the markets are in between SSEMH and SEMH, we should expect both the regressions(3.2) and 3.6 to provide support for the hypothesis. Notice that, in this case, markets should not react to news of victory of Berlusconi per se (only to the surprise), an implication we can empirically test.

3.4.4 *Caveats and interpretation of results:*

If we reject the null H_0, this does not necessarily imply that Berlusconi is corrupting and bending the law in his favor.[13] It could be that his adversaries would have passed revenge-legislation against his firms and that this is what the stock market is reacting to.

We compare the returns of his companies MS and MON to those of MED. All of these companies are his property but for some of them he has direct regulatory access while for others he has no access to the regulatory body (MED)[14]. According to our theory the results should be different for each of these companies.

Another robustness test we will perform is contrasting the cumulative residuals on his media company in Italy (MS) with his sister company in Spain (TL5). Both are in the same sector, with similar programs and

[13] Although there is anecdotal evidence that Berlusconi is benefitting his companies through favorable regulation, subsidies, and weakening the only competitor in the Italian media duopoly (the State TV company RAI).

[14] See table 3.1.

VARIABLES	(1) Total Return Mediaset	(2) Total Return Mediaset	(3) Total Return Mediaset	(4) Total Return Mediaset
Total Return	0.493***	0.493***	0.493***	0.493***
Media Index	(0.0382)	(0.0383)	(0.0383)	(0.0383)
Total Return	0.542***	0.542***	0.541***	0.542***
Italian Stock Index	(0.0362)	(0.0362)	(0.0362)	(0.0362)
Victory in	0.0205		0.00802	
Percentage	(0.0204)		(0.0375)	
Dummy Victory		0.000892		0.000892
		(0.00101)		(0.00101)
Elections			0.00177	
			(0.00270)	
Constant	0.000104	0.000104	0.000100	0.000104
	(0.000261)	(0.000261)	(0.000261)	(0.000261)
Observations	3,573	3,573	3,573	3,573
R-squared	0.489	0.489	0.489	0.489

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.3: Total Return regressed on Electoral outcome.

business model. Nevertheless, Berlusconi in power can only affect the regulation and subsidies in Italy[15], so that if the conflict of interest channel is the right one we should see the difference between the Cumulative residuals of the two companies trending upwards while if the right channel is the second one we should have their difference being flat.

3.5 Results:

3.5.1 *Regressions:*

Before testing our hypothesis on the whole sample of stocks, we regress the total daily returns of the MS company on the FTSEMIB (Italian index), the European Median index and a metric of political victory at the elections (we try both a dummy for victory and percentage of victory). According to our empirical strategy section - equation (3.4)-we should find no effects . Results are provided in table 3.3, and consistent with our prior, we find no results regardless of the victory specification and the inclusion of fixed effects for election days.

[15]Even though there is anecdotal evidence that he used his close relationship with the spanish PP to benefit his company abroad.

VARIABLES	(1) Residuals Mediaset	(2) Residuals Mediaset	(3) Residuals Mondadori	(4) Residuals Mediolanum	(5) Residuals Mondadori	(6) Residuals Mediolanum
Elect. Surprise	0.106***	0.109**	0.0889	-0.0125		
in Perct. Points	(0.0391)	(0.0362)	(0.155)	(0.0657)		
Victory in					0.0579	-0.0692
Percentg Points					(0.202)	(0.0383)
Constant	3.96e-07	0.00309	-0.000581	-0.00276	-0.00192	-0.00131
	(0.000261)	(0.00171)	(0.00763)	(0.00351)	(0.00562)	(0.00339)
Observations	3,573	8	3,573	8	3,573	8
R-squared	0.000	0.573	0.041	0.004	0.023	0.155

Robust standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 3.4: Residuals regressed on Electoral surprise.

This result is unsurprising and consistent with the theory so we turn to the surprise metric. A possible concern is that we are improperly borrowing strength from the number of observations.

To be conservative, we can calculate the residuals from a market regression of the MS daily total returns on the daily total returns of the Italian index and the European media industry index and consider the regression of these residuals on the surprise metric both at all dates and even just at our electoral dates. See equations (3.2). Consider that this strategy is biased against finding results because the low number of observations makes the necessary t-stat burdensome, and because the surprise metric is a noisy observation of the true (unobservable) market operators' surprise. The results do not change when using all observations or just the electoral dates.

As shown in Table 2, we find a significant effect at the 5% level and an estimated coefficient of 0.1 - large in real world terms. This coefficient implies that a shock of 5p.p. to the expectations metric (the surprise size that occurred in the 2006 elections) would have a 1 standard deviation effect in terms of residuals.

In table 3.4, we also reviewed similar results for the other two companies in which he has broad stakes: Mondadori and Mediolanum. We see that in this case, no result is statistically significant and that the sign for Mediulanum is even a negative rather than positive. Remember that under stronger than SSEMH we may see no effect for the surprise metric on the total return because the result is already expected by the markets through the aggregation of agents' private information.

Now if we assume that "some surprise" is not embedded in the price [16], we would expect the "surprise part" to have a bigger effect on the MS stock.

[16] So that markets are more than Semi strong efficient but not quite strongly efficient.

The reason is that MED is a financial company regulated by the Bank of Italy and outside the control of the Italian government, so it cannot be favored by the government through concessions, subsidies or regulations. Moreover, MED has a very small market share so that general regulation favorable to the whole financial industry would have very little bang for the buck.

Mondadori operates in the publishing industry and has a significant market share so that it is a good candidate company but has significantly less revenue and profits than MS. MS is heavily subject to the regulation of the state, has 40% of the market and when Berlusconi is in power he manages to control also the only competitor of MS; the state owned RAI. It is unsurprising then that MS stock reacts so strongly.

The evidence uncovered so far is consistent with Semi strong efficient markets or (almost) Strongly efficient markets. It suggests that conflict of interest is likely to be a problem and that the market perceives that MS benefits from Berlusconi being in power. Although the estimates of the effects are big, for the reasons reviewed above they are likely to be a lower bound. In the next section we explore the reason why Berlusconi's firms earn extra returns and precise estimate of the effects.

3.5.2 *Event study: Cumulative Residuals*

The best strategy to estimate the full effect of the benefit that Berlusconi's companies obtain from his political power, is to estimate the Cumulative Residual function. A problem with the previous strategy is that if those who know the secret party polls are buying stocks before the election, we will miss those returns in our regression estimates. In this section, I compute and plot the cumulative residuals function for each of Berlusconi's companies as discussed in the empirical strategy section (see equation (3.6)).

We average the residuals across elections correcting for the sign and the magnitude of the surprise effect by dividing the residuals by the corresponding value of the surprise metric. We can interpret the resulting graph as the cumulative residual of a (favorable to Berlusconi) 1 percentage point change in the surprise metric. According to the Strong market Efficiency Hypothesis, if Berlusconi's political power favors a firm, we should see an upward sloping curve before the election day and a quasi flat line thereafter. Even though the number of events for each stock is low, the results are remarkably close to the predictions of the theory pointing to a strong market expectation that Berlusconi political power is beneficial to MS and MON.

Notice how, now that we take into account the cumulative effects of the surprise metric, we find that each 1 percentage point increase in the surprise metric leads to a cumulative 6 percentage points increase (residuals) for the total returns of MS and 4 percentage points for MON. See figures 3-2 and 3-3 respectively. These are huge effects and the fact that both graphs show a flattening of the curve right after the electoral date suggests that the shape is due to an embedding of private and public information in the run up to the election.

Unsurprisingly given the discussion about the different regulatory bodies and the smaller bang for buck, we do not find such an effect for his third company MED.

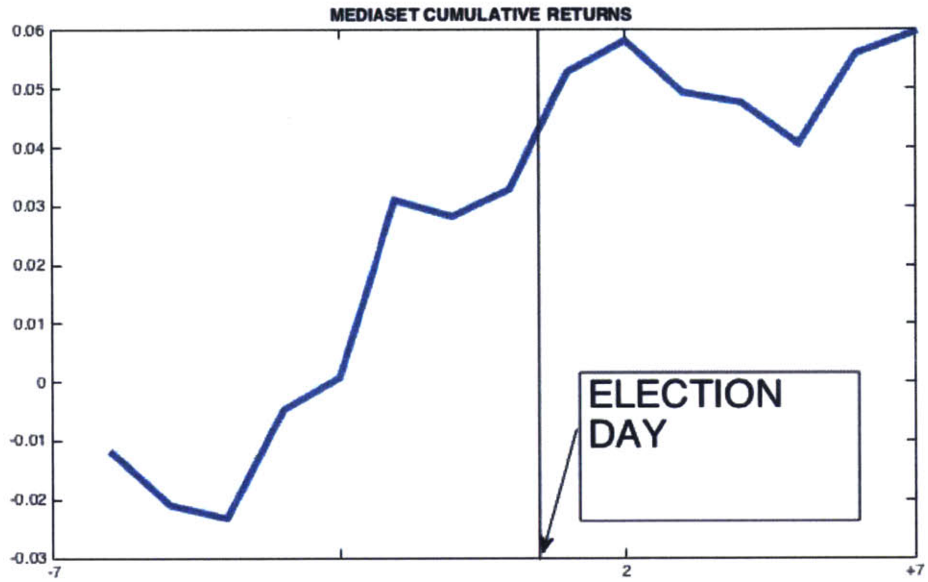


Figure 3-2: Mediaset Cumulative Returns.

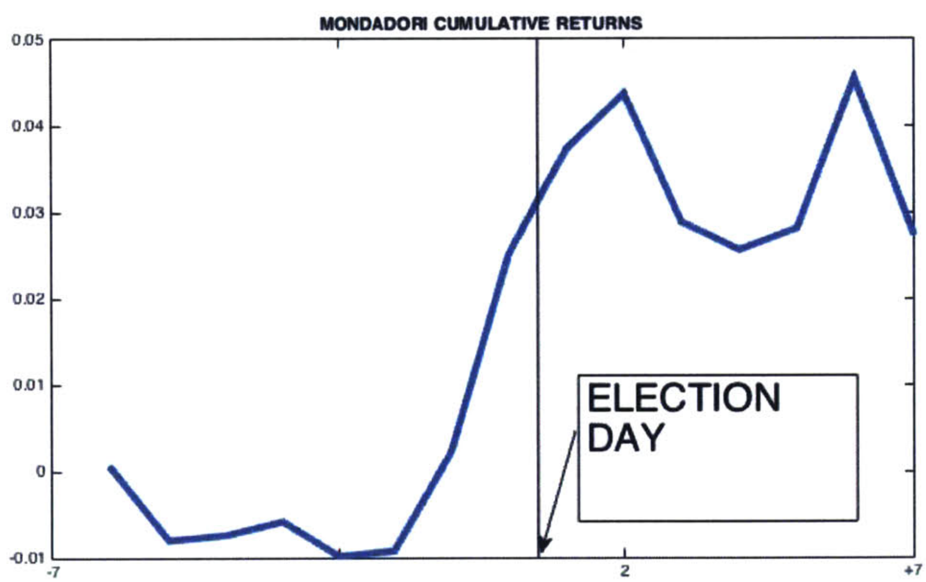


Figure 3-3: Mondadori Cumulative Returns.

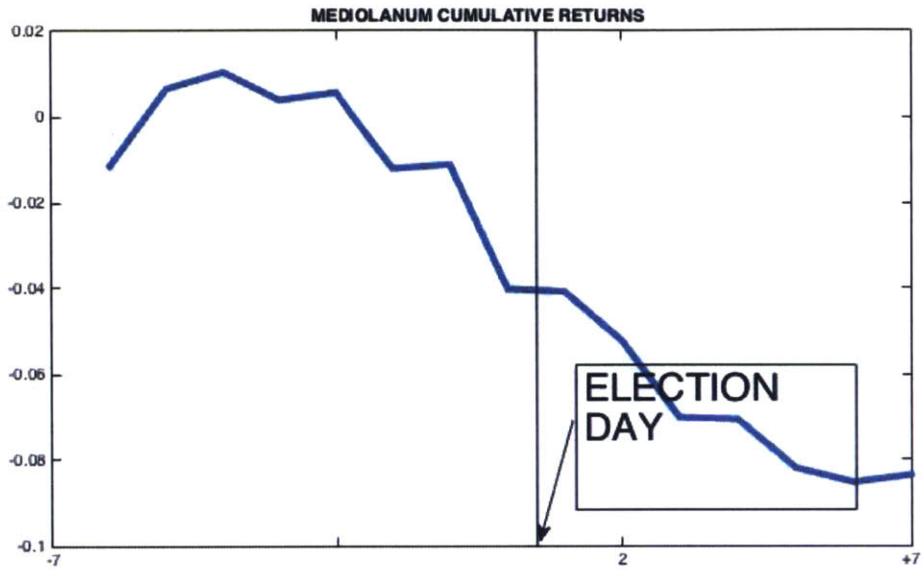


Figure 3-4: Mediolanum Cumulative Returns.

For MED, if anything it seems that Berlusconi's political power has a negative effect. See figure 3-4. Perhaps, a possible explanation for this graph is that MED loses from politically aware customers' boycotts without the benefits of ad hoc laws.

We showed that the companies which are under Berlusconi's regulatory reach react positively to his political power. The only company that is completely outside his regulatory reach (in the banking sector) and which has a low market share reacts negatively. These facts already suggest that one of the proposed explanations in the Caveat subsection is unlikely to hold. If the driver of the extra returns were concerns about his leadership in the companies we should expect similar effects on all of them.

Yet, one may counter that Berlusconi made his fortune precisely in the media sector and that his leadership skills have a disproportionate effect on his media company accounting for the difference between the above results. Luckily, since 2004 also the Spanish media company TL5 (controlled again by Mr. Berlusconi) has been listed. TL5 operates in the exact same market and with the same generalist business model as MS (they even share the same brand symbol), the only difference coming from the regional markets within the EU in which they operate.

If the rationale behind the Italian stock movements is conflict of interest, we should observe no effect for TL5. If the rationale is the loss of a media leader à la Steve Jobs, we should observe similar results on both companies. The graph of the difference is plotted at figure 3-5.

Similarly we can run a regression for the difference of residual returns on the surprise effect. Table 3.5 reports the results.

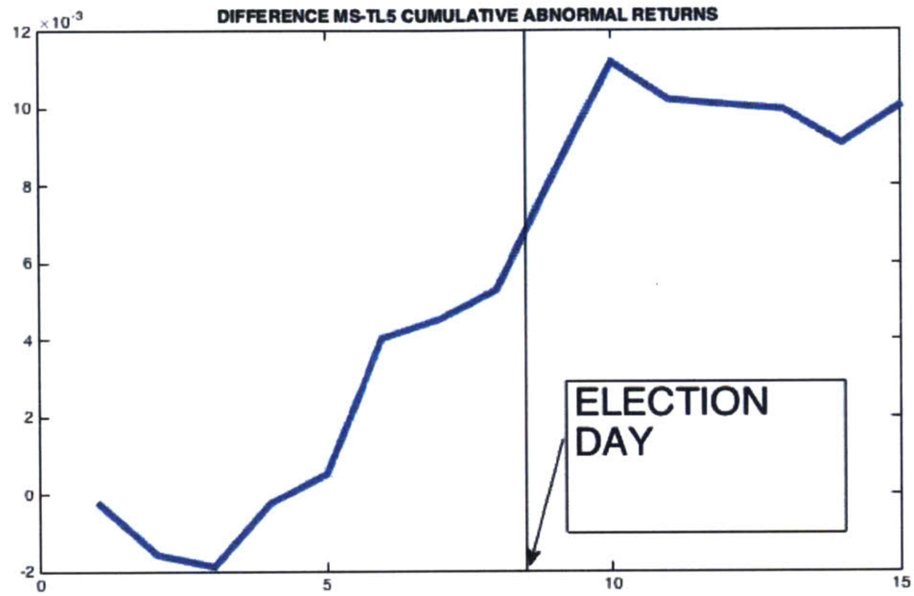


Figure 3-5: Mediaset-Tele5 CR.

VARIABLES	(1) Resid. Mediaset -Resid. Tele5	(2) Resid. Mediaset -Resid. Tele5	(3) Resid. Mediaset -Resid. Tele5
Elect. Surprise	0.143**		0.149**
in Perct. Points	(0.0606)		(0.0485)
Victory		-0.0120	
Percent.		(0.0234)	
Constant	-0.000506	-0.000507	0.00244
	(0.000469)	(0.000469)	(0.00237)
Observations	1,928	1,928	6
R-squared	0.000	0.000	0.655

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.5: Residual Returns on surprise.

As we can see the estimated coefficient is very similar to the one of Mediaset abnormal returns (slightly larger) and statistically significant even when we use only the election dates[17]. The rejection of the hypothesis of a 0 coefficient suggests that the *"CEO-founder negative effect"* is probably not what lies behind the previous results.

Lastly as a further test we look at whether the three days or five days volatility is affected by the electoral date. The tables with the results are presented in the appendix, but we fail to find evidence that volatility is significantly higher for any stock around electoral dates.

3.6 Robustness:

3.6.1 *Simulations:*

In this section we simulate randomly the 8 electoral dates, and keep the surprise metric as before. We worry that the interaction between the weights of the residuals (given by the surprise metric) and the random occurrence of other events that cause serial correlation of returns through time, can be at the origin of the graphs we showed above.

Suppose instead of graphs for the cumulative abnormal returns we were estimating coefficients. We would be interested in knowing the p-value of our estimates. In particular we would like to know, under the null hypothesis that Berlusconi in power is not beneficial to his firms, how likely it would be to observe cumulative returns pictures as the ones we showed above.

Simulating the CAR under the null H_0 that Berlusconi in power is not beneficial to his firms can be done by selecting the election dates randomly. At those dates no real elections are being held so only the "standard public information" and "insider trading" hits the market. It would be worrying if a picture similar to the above could be produced easily with random election dates.

Before detailing how we will compute the "p-value" of our graphs, let us plot a Cumulative Abnormal Returns for a percentage point of surprise with random election dates. The CAR are computed for MS, which will be the focus of the exercise presented here.

As we can see, not only there is no trend before the elections but there are frequent changes in the direction of the CAR function - at least 5 before the simulated election. We can contrast this with the original picture of the CAR for the true election dates where there was a clear trend upwards before the election, and there were just 3 changes of direction before the election date. It would be useful to have a way to simulate these graphs from the true distribution under H_0 without having to analyze each graph.

[17]Here the observations drop to 6 because Tele5 was listed only in 2004; so we lose the elections of 2001 and 2004.

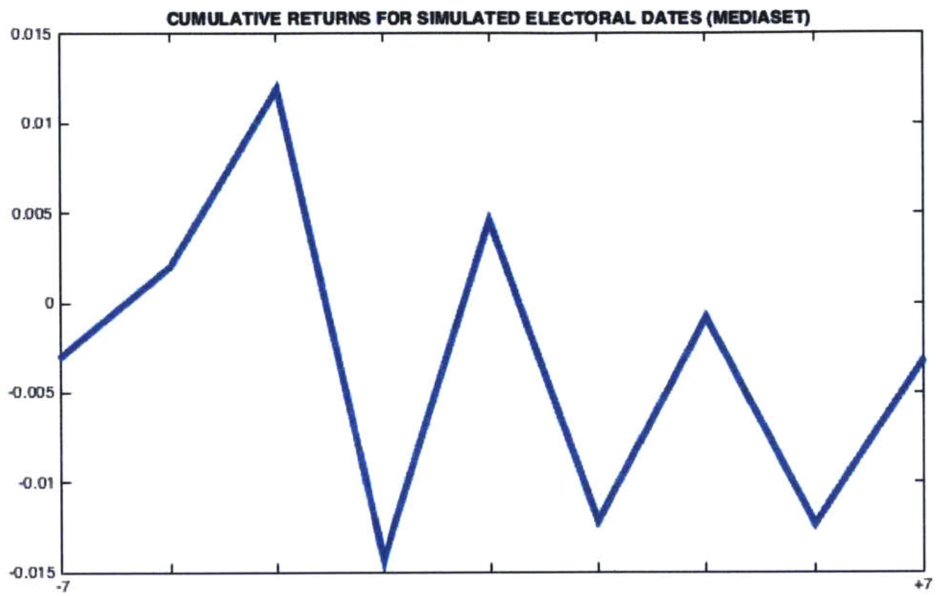


Figure 3-6: Simulated MS Cumulative Returns.

To do so I construct a Monte Carlo simulation program in MATLAB that simulates drawing 10000 groups of 8 random electoral dates. From here I can calculate 10000 random simulated graphs of cumulative abnormal returns per percentage point of Surprise. I consider two parameters of the original MS graph relevant: the fact that the slope is upward sloping and the fact that there are few "changes of direction". Among the 10000 graphs I make Matlab calculate the regression trend and the number of changes of direction. I am interested in a kind of "p-value" exercise so I wonder what is the fraction of graphs that have a slope at least as high as the MS original graph and at least as few "turning points" as the original MS graph. The answer is a reassuring 7%

3.7 Conclusion:

Contrary to the Fisman case study on Dick Cheney's personal ties, we find very big and statistically significant effects for the ruling of Berlusconi being beneficial to his companies. The effects imply swings of up to 20% of market capitalization for surprising electoral outcomes like the 2006 elections. At current market capitalizations this implies by back of the envelope calculations direct costs of over 1 billion \$. The effects are large and very much consistent with independent estimates (using a completely different methodology by DellaVigna et al. [2013]) and suggest that, even in developed countries, the costs of conflict of interest are not trivial. The results call for further research to test the external validity of the estimates in other developed economies and the costs introduced by conflict of interest laws. Unfortunately we are not able to shed any light on whether the distortions are due to revenge legislation by the opposition or bending of the law by Berlusconi himself but we provided suggestive evidence that the estimates are not due to the "exit of founder-CEO" effect.

References

R Michael Alvarez and Jonathan Nagler. A new approach for modelling strategic voting in multiparty elections. *British Journal of Political Science*, 30(1):57–75, 2000.

Kenneth J Arrow. Informational structure of the firm. *The American Economic Review*, 75(2):303–307, 1985.

Sandra E Black. Do better schools matter? parental valuation of elementary education. *Quarterly journal of economics*, pages 577–599, 1999.

David Blackwell et al. Comparison of experiments. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 1, pages 93–102, 1951.

Andrew Caplin and Mark Dean. Rational inattention and state dependent stochastic choice. *unpublished, New York University*, 2013.

Gabriel Carroll. Robustness and linear contracts. *The American Economic Review*, 105(2):536–563, 2015.

Gabriel D Carroll. Robust incentives for information acquisition. *Available at SSRN 2214984*, 2013.

Sylvain Chassang. Calibrated incentive contracts. *Econometrica*, 81(5):1935–1971, 2013.

Stijn Claessens, Erik Feijen, and L Laeven. Political connections and preferential access to finance: The role of campaign contributions. *Journal of Financial Economics*, 2008.

Stephen Coate, Michael Conlin, and Andrea Moro. The performance of pivotal-voter models in small-scale elections: Evidence from Texas liquor referenda. *Journal of Public Economics*, 92(3-4):582–596, April 2008a. ISSN 00472727.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Tri Vi Dang, Gary Gorton, and Bengt Holmström. The information sensitivity of a security. *Unpublished working paper, Yale University*, 2013.

Tri Vi Dang, Gary Gorton, Bengt Holmström, and Guillermo Ordonez. Banks as secret keepers. Technical report, National Bureau of Economic Research, 2014.

Henrique De Oliveira, Tommaso Denti, Maximilian Mihm, and M Kemal Ozbek. Rationally inattentive preferences. *Available at SSRN 2274286*, 2013.

S DellaVigna, Ruben Durante, and BKE La Ferrara. Influence for Sale: Evidence from the Italian Advertising Market. *Unpublished Manuscript*, 2013.

Joel S Demski and David EM Sappington. Delegated expertise. *Journal of Accounting Research*, pages 68–89, 1987.

Peter Diamond. Managerial incentives: on the near linearity of optimal compensation. *Journal of Political Economy*, 106(5):931–957, 1998.

Arindrajit Dube, T William Lester, and Michael Reich. Minimum wage effects across state borders: Estimates using contiguous counties. *The review of economics and statistics*, 92(4):945–964, 2010.

John Duffy and Margit Tavits. Beliefs and Voting Decisions: A Test of the Pivotal Voter Model. *American Journal of Political Science*, 52(3):603–618, 2008a.

Maurice Duverger. *Political parties: Their organization and activity in the modern state.* Methuen, 1959.

M Faccio. Politically connected firms. *The American economic review*, (1999), 2006.

Mara Faccio, RW Masulis, and J McConnell. Political connections and corporate bailouts. *The Journal of Finance*, LXI(6):2597–2635, 2006.

E Fama, L Fisher, M Jensen, and R Roll. The adjustment of stock prices to new information. *International economic review*, 1969.

T Ferguson and HJ Voth. Betting on Hitler—the value of political connections in Nazi Germany. *The Quarterly Journal of Economics*, (February), 2008.

D Fisman, Ray Fisman, Julia Galef, and Rakesh Khurana. Estimating the value of connections to Vice president Cheney. *Unpublished paper*, 2006.

R Fisman. Estimating the value of political connections. *The American Economic Review*, 2001.

Thomas Fujiwara. A Regression Discontinuity Test of Strategic Voting and Duverger's Law. *Quarterly Journal of Political Science*, 6(3-4):197–233, 2011.

Tim Ganser and Stan Veuger. A difficult choice: strategic voting in proportional representation systems. *Unpublished*, 2012.

Matthew Gentzkow and Emir Kamenica. Costly persuasion. *The American Economic Review*, 104(5): 457–462, 2014.

Eitan Goldman, Jörg Rocholl, and J So. Do politically connected boards affect firm value? *Review of Financial Studies*, (December), 2009.

Benjamin Hebert. Moral hazard and the optimality of debt. *Available at SSRN 2185610*, 2015.

Bengt Holmstrom and Paul Milgrom. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica: Journal of the Econometric Society*, pages 303–328, 1987.

S Jayachandran. The Jeffords Effect. *Journal of Law and Economics*, 2006.

Kei Kawai and Yasutora Watanabe. Inferring strategic voting. *The American Economic Review*, pages 1–42, 2013a.

Kei Kawai and Yasutora Watanabe. Inferring strategic voting. *The American Economic Review*, 103(2): 624–662, 2013b.

DS Lee and Thomas Lemieux. Regression discontinuity designs in economics. 48(June):281–355, 2009.

David K. Levine and Thomas R. Palfrey. The Paradox of Voter Participation? A Laboratory Study. *American Political Science Review*, 101(01):143, February 2007a. ISSN 0003-0554.

Hongbin Li, Lingsheng Meng, Qian Wang, and Li-An Zhou. Political connections, financing and firm performance: Evidence from Chinese private firms. *Journal of Development Economics*, 87(2):283–299, October 2008. ISSN 03043878.

Filip Matejka, Alisdair McKay, et al. Rational inattention to discrete choices: A new foundation for the multinomial logit model. 2011.

Filip Matejka, Jakub Steiner, and Colin Stewart. Dp10720 rational inattention dynamics: Inertia and delay in decision-making. 2015.

Lars-Göran Mattsson and Jörgen W Weibull. Probabilistic choice and procedurally bounded rationality. *Games and Economic Behavior*, 41(1):61–78, 2002.

Roger B Myerson and Robert J Weber. A theory of voting equilibria. *American Political Science Review*, pages 102–114, 1993b.

Suresh Naidu. Suffrage, schooling, and sorting in the post-bellum us south. Technical report, National Bureau of Economic Research, 2012.

Francisco Perez-Gonzalez. Inherited Control and Firm Performance. *SSRN Electronic Journal*, 2002. ISSN 1556-5068.

Maxim L Pinkovskiy. Economic discontinuities at borders: Evidence from satellite data on lights at night. Technical report, Working Paper, 2013.

Doron Ravid. Bargaining with rational inattention. Technical report, Working paper, 2014.

MA Satterthwaite. Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 1975.

Marilyn L Shaw and Peter Shaw. Optimal allocation of cognitive resources to spatial locations. *Journal of Experimental Psychology: Human Perception and Performance*, 3(2):201, 1977.

Christopher A Sims. Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690, 2003.

J Spenkuch. On the Extent of Strategic Voting. *Available at SSRN 2170997*, (November), 2012.

Michael Woodford. Imperfect common knowledge and the effects of monetary policy. Technical report, National Bureau of Economic Research, 2001.

Michael Woodford. Inattentive valuation and reference-dependent choice. 2012.

Ming Yang. Optimality of debt under flexible information acquisition. *Available at SSRN 2103971*, 2013.

Ming Yang. Coordination with flexible information acquisition. *Journal of Economic Theory*, 2014.

Luis Zermeno. A principal-expert model and the value of menus. Technical report, working paper, Massachusetts Institute of Technology, 2011.

Appendix for Chapter 1

Proofs

Proposition 3

Proof. We want to show that ∎

$$D\left(\alpha p+(1-\alpha)\, q||\alpha\pi+(1-\alpha)\,\eta\right)\leq\alpha D\left(p||\pi\right)+(1-\alpha)\, D\left(q||\eta\right)$$

$D\left(\alpha p+(1-\alpha)\, q||\alpha\pi+(1-\alpha)\,\eta\right)=\sum_x\left(\alpha p\left(x\right)+(1-\alpha)\, q\left(x\right)\right) log\left(\frac{\alpha p(x)+(1-\alpha)q(x)}{\alpha\pi(x)+(1-\alpha)\eta(x)}\right)$

Using the log-sum inequality:

$\sum_x\left(\alpha p\left(x\right)+(1-\alpha)\, q\left(x\right)\right) log\left(\frac{\alpha p(x)+(1-\alpha)q(x)}{\alpha\pi(x)+(1-\alpha)\eta(x)}\right)\leq$

$\sum_x \alpha p\left(x\right) log\left(\frac{\alpha p(x)}{\alpha\pi(x)}\right)+\sum_x(1-\alpha)\, q\left(x\right) log\left(\frac{(1-\alpha)q(x)}{(1-\alpha)\eta(x)}\right)=\alpha D\left(p||\pi\right)+(1-\alpha)\, D\left(q||\eta\right)$

Lemma 9

Proof. The above Problem is constructed so that all signals in the Experiment space occur with positive probability. Hence $p\left(x\right)=\mathbb{E}_{p(\theta)}\left[p\left(x|\theta\right)\right]>0\;\forall x\in X$. Suppose that the were two signals x_1 and x_2 such that $d\left(x_1\right)=d\left(x_2\right)$. Then we can construct the following experiment $X'=X\backslash x_2$, $p'\left(x|\theta\right)=p\left(x|\theta\right)\forall x\neq x_1,x_2$, $p'\left(x_1|\theta\right)=p\left(x_1|\theta\right)+p\left(x_2|\theta\right)$, and the decision rule is identical restricted to the new signal experiment space $d'=d_{|X\backslash x_2}$. We can recompute the posterior distributions for this new experiment. Obviously, all posteriors for signals different than x_1 will be unchanged while $p'\left(\theta|x_1\right)=\frac{\mathbb{E}_{p(\theta)}[p(x_1|\theta)]}{\mathbb{E}_{p(\theta)}[p(x_1|\theta)]+\mathbb{E}_{p(\theta)}[p(x_2|\theta)]}p\left(\theta|x_1\right)+\frac{\mathbb{E}_{p(\theta)}[p(x_2|\theta)]}{\mathbb{E}_{p(\theta)}[p(x_1|\theta)]+\mathbb{E}_{p(\theta)}[p(x_2|\theta)]}p\left(\theta|x_2\right)$. Call $\alpha=\frac{\mathbb{E}_{p(\theta)}[p(x_1|\theta)]}{\mathbb{E}_{p(\theta)}[p(x_1|\theta)]+\mathbb{E}_{p(\theta)}[p(x_2|\theta)]}$. Then we can rephrase the probabilities in terms of posteriors to observe that the Bernoulli utility over wealth is unchanged:

$$\mathbb{E}_{p(\theta)}\left[\mathbb{E}_{p'(x|\theta)}u\left(W+b\left(d'\left(x\right),\theta\right)\right)\right]=$$

$$\sum_{x\in X'\backslash\{x_1\}}p'\left(x\right)\left[\mathbb{E}_{p'(\theta|x)}\left[u\left(W+b\left(d'\left(x\right),\theta\right)\right)\right]\right]+\left(p'\left(x_1\right)\right)\mathbb{E}_{p'(\theta|x_1)}\left[u\left(W+b\left(d'\left(x_1\right),\theta\right)\right)\right]$$

By construction of m', this is equal to:

$$\sum_{x\in X'\backslash\{x_1\}}p'\left(x\right)\left[\mathbb{E}_{p'(\theta|x)}\left[u\left(W+b\left(d'\left(x\right),\theta\right)\right)\right]\right]+$$

$$\left(p\left(x_1\right)+p\left(x_2\right)\right)\mathbb{E}_{\alpha p(\theta|x_1)+(1-\alpha)p(\theta|x_2)}\left[u\left(W+b\left(d'\left(x_1\right),\theta\right)\right)\right]=$$

$$\sum_{x \in X' \setminus \{x_1\}} p'(x) \left[\mathbb{E}_{p'(\theta|x)} \left[u \left(W + b \left(d'(x), \theta\right)\right)\right]\right] + \mathbb{E}_{p(\theta)} \left[p \left(x_1|\theta\right)\right] p \left(\theta|x_1\right) \left[u \left(W + b \left(d'(x_1), \theta\right)\right)\right]$$

$$+\mathbb{E}_{p(\theta)} \left[p \left(x_2|\theta\right)\right] p \left(\theta|x_2\right) \left[u \left(W + b \left(d'(x_1), \theta\right)\right)\right]$$

By definition of $X'$and construction of m', d', this equals:

$$\sum_{x \in X \setminus \{x_1, x_2\}} p(x) \left[\mathbb{E}_{p(\theta|x)} \left[u \left(W + b \left(d(x), \theta\right)\right)\right]\right] + \mathbb{E}_{p(\theta)} \left[p \left(x_1|\theta\right)\right] p \left(\theta|x_1\right) \left[u \left(W + b \left(d(x_1), \theta\right)\right)\right] +$$

$$\mathbb{E}_{p(\theta)} \left[p \left(x_2|\theta\right)\right] p \left(\theta|x_2\right) \left[u \left(W + b \left(d(x_1), \theta\right)\right)\right] = \mathbb{E}_{p(\theta)} \left[\mathbb{E}_{p(x|\theta)} u \left(W + b \left(d(x), \theta\right)\right)\right]$$

Similarly for the second term, using the convexity of $I(\cdot)$ with respect to posteriors - which comes from the concavity of entropy $H(\cdot)$ we get:

$$I(m') = H(p) - \mathbb{E}_{p'(x)} \left[H \left(p'(\cdot|x)\right)\right] =$$

$$H(p) - \sum_{x \in X' \setminus \{x_1\}} p'(x) H \left(p'(\cdot|x)\right) - p'(x_1) H \left(p'(\cdot|x_1)\right) =$$

$$H(p) - \sum_{x \in X \setminus \{x_1, x_2\}} p(x) H \left(p(\cdot|x)\right) - \left(p(x_1) + p(x_2)\right) H \left(\alpha p(\cdot|x_1) + (1-\alpha) p(\cdot|x_2)\right) <$$

$$H(p) - \sum_{x \in X \setminus \{x_1, x_2\}} p(x) H \left(p(\cdot|x)\right) - \left(p(x_1) + p(x_2)\right) \left(\alpha H \left(p(\cdot|x_1)\right) + (1-\alpha) H \left(p(\cdot|x_2)\right)\right) =$$

$$H(p) - \mathbb{E}_{p(x)} \left[H \left(p(\cdot|x)\right)\right] = I(m)$$

Because we were able to construct a new experiment (X', m') and decision rule d' that does strictly better than the original candidate solution. (X, m) and d were not optimal. □

Lemma 13

Proof. A solution exists by the extreme value theorem given that the function is continuous and we can view the choice of the experiment $m \in E(D)$ as the choice of a joint probability.The space $E(D)$ is a closed and bounded subset of $[0, 1]^{|\Theta| \times D}$. Moreover, the set of experiments $E(D)$ is also convex. By the definition of mutual information, we express entropy as the Kullback-Leibler divergence of the joint over the two marginals, and that D_{KL} is strictly jointly convex. Therefore $I(m) = D_{KL}(p(d|\theta)p(\theta)||p(d)\cdot p(\theta))$ and by the strict join convexity of D_{KL}

$$I(\alpha m + (1-\alpha)m') =$$

$$D_{KL}((\alpha p(d|\theta) + (1-\alpha)p'(d|\theta))p(\theta)||(\alpha p(d) + (1-\alpha)p'(d))\cdot p(\theta)) <$$

$$\alpha D_{KL}(p(d|\theta)p(\theta)||p(d)\cdot p(\theta)) + (1-\alpha)D_{KL}(p'(d|\theta)p(\theta)||p'(d)\cdot p(\theta)) =$$

$$\alpha I(m) + (1-\alpha)I(m')$$

So that the cost is a convex function of m while the Bernoulli utility over wealth is a linear function of m. So overall the utility of the agent is strictly convex □

Proposition 21 This is a direct application of Theorem 18 by letting $g(p,\theta,d) = \alpha(\theta)\log(p)$.

Proposition 22

Proof. For all actions $\theta \in \Theta$, we have that either $p(d|\theta) > 0$ for all $d \in D$ or $p(d|\theta) = 0\ \forall d \in D$ because of proposition 14. Therefore if $(m, b(\cdot,\cdot))$ is a solution to problem 15 augmented with the participation constraint then $(m', b'())$, where $(m', b'())$ are the restriction of $(m, b(\cdot,\cdot))$ to the set D' of decisions that occur with positive probability, is the solution to:

$$\max_{m\in E(D'),b(\cdot,\cdot)} \mathbb{E}_{p(\theta)}[\mathbb{E}_{p(d|\theta)}[y(d,\theta) - b(d,\theta)]]$$

such that for all $d \in D'$ and $\theta \in \Theta$:

$$0 \leq p(d|\theta) \leq 1 \text{ and } \sum_{d'\in D'} p(d'|\theta) = 1$$

$$b(d,\theta) \geq 0\ [\phi(d,\theta)]$$

$$\mathbb{E}_{p(\theta)}\left[\mathbb{E}_{p(d|\theta)}u(W + b(d,\theta))\right] - \mu I(m) \geq K[\lambda]$$

$$u\left(W+b\left(d,\theta\right)\right)p\left(\theta\right)-\mu p\left(\theta\right)log\left(\frac{p\left(d|\theta\right)}{p\left(d\right)}\right)=\mu p\left(\theta\right)log\left(\sum_{d'\in D'}p\left(d'\right)e^{\frac{u\left(W+b\left(d',\theta\right)\right)}{\mu}}\right)\left[\Psi\left(d,\theta\right)\right]$$

We prove that for this problem the optimal contract is $b'\left(d,\theta\right)=\max\{0,A\cdot y\left(d,\theta\right)+B\left(\theta\right)+C\left(d\right)\}\ \forall d\in D'$. It then follows by limited liability and the utility of the principal that $b\left(d,\theta\right)=0$ for all $d\in D\backslash D'$ and consequently for all $\theta\in\Theta$. For the rephrased subproblem among decisions occurring with positive probability, we apply the KKT theorem to express the gradient of the objective function at m, $b\left(\cdot,\cdot\right)$ as a linear combination of the gradients of the binding constraints at such points. The constraints for the probabilities will not be binding for decisions taken with positive probability as showed in the previous section, while the constraint on $\sum_{d'\in D'}p\left(d'|\theta\right)=1$ is already embedded in the third constraint. Therefore letting $\phi\left(d,\theta\right)$ and $\Psi\left(d,\theta\right)$ be the Lagrange multipliers for the potentially binding constraints and expressing the maximization with respect to $u\left(W+b\left(d,\theta\right)\right)$ instead of $b\left(d,\theta\right)$ we get:

$$-\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}p\left(\theta\right)p\left(d|\theta\right)+\phi\left(d,\theta\right)-p\left(\theta\right)\Psi\left(d,\theta\right)+$$

$$\mu p\left(\theta\right)\sum_{d'\in D'}\frac{p\left(d\right)e^{\frac{u\left(W+b\left(d',\theta\right)\right)}{\mu}}}{\sum_{d'\in D'}p\left(d'\right)e^{\frac{u\left(W+b\left(d',\theta\right)\right)}{\mu}}}\frac{1}{\mu}\Psi\left(d,\theta\right)+\lambda=0$$

and after some rearrangement we get:

$$\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}=\frac{\phi\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\frac{\Psi\left(d,\theta\right)}{p\left(d|\theta\right)}+\sum_{l\in D'}\Psi\left(l,\theta\right)+\lambda \tag{7}$$

We get the following condition from $p\left(d|\theta\right)$:

$$\frac{y(d,\theta)-b(d,\theta)}{\mu}=\sum_{\theta'}\Psi\left(d,\theta'\right)\frac{p\left(\theta'\right)}{p(d)}-\frac{\Psi\left(d,\theta\right)}{p\left(d|\theta\right)}-\sum_{\theta'\in\Theta}\sum_{l\in D'}\Psi\left(l,\theta'\right)\frac{p\left(\theta'\right)p\left(d|\theta'\right)}{p\left(d\right)} \tag{8}$$

Notice that in this case the presence of the participation constraint does not affect the first order condition on the augmented objective function. This is because in order for m to be optimal it has to already be a solution to the agent's problem.

Now substituting (7) into (8). We have:

$$\frac{y(d,\theta)-b(d,\theta)}{\mu}=$$

$$\sum_{\theta'}\Psi\left(d,\theta'\right)\frac{p\left(\theta'\right)}{p\left(d\right)}-\sum_{\theta'\in\Theta}\sum_{l\in D'}\Psi\left(l,\theta'\right)\frac{p\left(\theta'\right)p\left(d|\theta'\right)}{p\left(d\right)}+\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}-\frac{\phi\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\sum_{l\in D'}\Psi\left(l,\theta\right)-\lambda$$

We can also rearrange terms by adding multiplying and diving the first term by $p\left(d|\theta'\right)$ and recognizing that we get a posterior $p\left(\theta'|d\right)$, we can rewrite:

$$y(d,\theta)-b(d,\theta)=$$

$$\mu\sum_{\theta'}\frac{\Psi\left(d,\theta'\right)}{p\left(d|\theta'\right)}p\left(\theta'|d\right)-\mu\sum_{\theta'\in\Theta}\sum_{l\in D'}\Psi\left(l,\theta'\right)p\left(\theta'|d\right)+\mu\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}-\mu\frac{\phi\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\mu\sum_{l\in D'}\Psi\left(l,\theta\right)-\mu\lambda$$

Now consider that the first two terms are equal to:

$$\mu\sum_{\theta'}\frac{\Psi\left(d,\theta'\right)}{p\left(d|\theta'\right)}p\left(\theta'|d\right)=\mu\mathbb{E}_{p(\theta|d)}\left[\frac{\phi\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}+\sum_{l\in D'}\Psi\left(l,\theta\right)+\lambda\right]$$

$$-\mu\mathbb{E}_{p(\theta|d)}\left[\sum_{l\in D'}\Psi\left(l,\theta'\right)\right]=\mathbb{E}_{p(\theta|d)}\left[y\left(d,\theta\right)-b\left(d,\theta\right)\right]$$

So that we can rewrite the above as:

$$b\left(d,\theta\right)=y\left(d,\theta\right)+B\left(\theta\right)+C\left(d\right)-\mu\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}$$

Where,

$$B\left(\theta\right)=\mu\sum_{l\in D'}\Psi\left(l,\theta\right)$$

$$C\left(d,\theta\right)=-\mu\mathbb{E}_{p(\theta|d)}\left[\frac{\phi\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}+\sum_{l\in D'}\Psi\left(l,\theta\right)+\lambda\right]+$$

$$\mu\frac{\phi\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\mathbb{E}_{p(\theta|d)}\left[y\left(d,\theta\right)-b\left(d,\theta\right)\right]$$

But for $b\left(d,\theta\right)>0$, the limited liability is not binding and then $\phi\left(d,\theta\right)=0$ so that $C\left(d,\theta\right)$ is only a function of the state d $\left(C\left(d\right)\right)$. So that the optimal contract to the original problem is:

$$b\left(d,\theta\right)=\max\{0,y\left(d,\theta\right)+B\left(\theta\right)+C\left(d\right)-\mu\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}\}$$

For the log case and risk neutrality this becomes respectively:

$$b(d,\theta) = \max\{0, \frac{1}{1+\mu} y(d,\theta) + B(\theta) + C(d)\}$$

$$b(d,\theta) = \max\{0, y(d,\theta) + B(\theta) + C(d)\}$$

□

Proposition 23

Proof. For all actions $\theta \in \Theta$, we have that either $p(d|\theta) > 0$ for all $d \in D$ or $p(d|\theta) = 0$ $\forall d \in D$ because of proposition 14. Therefore if $(m, b(\cdot,\cdot))$ is a solution to problem 15 augmented with the wealth constraints then $(m', b'())$, where $(m', b'())$ are the restriction of $(m, b(\cdot,\cdot))$ to the set D' of decisions that occur with positive probability, is the solution to:

$$\max_{m \in E(D'), b(\cdot,\cdot)} \mathbb{E}_{p(\theta)}[\mathbb{E}_{p(d|\theta)}[y(d,\theta) - b(d,\theta)]]$$

such that for all $d \in D'$ and $\theta \in \Theta$:

$$0 \leq p(d|\theta) \leq 1 \text{ and } \sum_{d' \in D'} p(d'|\theta) = 1$$

Figures

$$b(d,\theta) \geq 0 \ \ [\phi(d,\theta)]$$

$$b(d,\theta) \leq w(d,\theta) \ \ [\gamma(d,\theta)]$$

$$u(W + b(d,\theta))\, p(\theta) - \mu p(\theta) \log\left(\frac{p(d|\theta)}{p(d)}\right) = \mu p(\theta) \log\left(\sum_{d' \in D'} p(d')\, e^{\frac{u(W+b(d',\theta))}{\mu}}\right) [\Psi(d,\theta)]$$

$$-\frac{1}{u'(W + b(d,\theta))} p(\theta)\, p(d|\theta) + \phi(d,\theta) - p(\theta)\, \Psi(d,\theta) +$$

$$\mu p(\theta) \sum_{d' \in D'} \frac{p(d)\, e^{\frac{u(W+b(d',\theta))}{\mu}}}{\sum_{d' \in D'} p(d')\, e^{\frac{u(W+b(d',\theta))}{\mu}}} \frac{1}{\mu} \Psi(d,\theta) - \gamma(d,\theta) = 0$$

and after some rearrangement we get:

$$\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}=\frac{\phi\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\frac{\gamma\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\frac{\Psi\left(d,\theta\right)}{p\left(d|\theta\right)}+\sum_{l\in D'}\Psi\left(l,\theta\right) \tag{9}$$

We get the following condition from $p\left(d|\theta\right)$:

$$\frac{y(d,\theta)-b(d,\theta)}{\mu}=\sum_{\theta'}\Psi\left(d,\theta'\right)\frac{p\left(\theta'\right)}{p\left(d\right)}-\frac{\Psi\left(d,\theta\right)}{p\left(d|\theta\right)}-\sum_{\theta'\in\Theta}\sum_{l\in D'}\Psi\left(l,\theta'\right)\frac{p\left(\theta'\right)p\left(d|\theta'\right)}{p\left(d\right)} \tag{10}$$

Now substituting (9) into (10). We have:

$$\frac{y(d,\theta)-b(d,\theta)}{\mu}=$$

$$\sum_{\theta'}\Psi\left(d,\theta'\right)\frac{p\left(\theta'\right)}{p\left(d\right)}-\sum_{\theta'\in\Theta}\sum_{l\in D'}\Psi\left(l,\theta'\right)\frac{p\left(\theta'\right)p\left(d|\theta'\right)}{p\left(d\right)}+$$

$$\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}+\frac{\gamma\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\frac{\phi\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\sum_{l\in D'}\Psi\left(l,\theta\right)$$

We can also rearrange terms by adding multiplying and diving the first term by $m\left(d|\theta'\right)$ and recognizing that we get a posterior $p\left(\theta'|d\right)$, we can rewrite:

$$y(d,\theta)-b(d,\theta)=$$

$$\mu\sum_{\theta'}\frac{\Psi\left(d,\theta'\right)}{p\left(d|\theta'\right)}p\left(\theta'|d\right)-\mu\sum_{\theta'\in\Theta}\sum_{l\in D'}\Psi\left(l,\theta'\right)p\left(\theta'|d\right)$$

$$+\mu\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}+\mu\frac{\gamma\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\mu\frac{\phi\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\mu\sum_{l\in D'}\Psi\left(l,\theta\right)$$

Now consider that the first terms equals:

$$\mu\sum_{\theta'}\frac{\Psi\left(d,\theta'\right)}{p\left(d|\theta'\right)}p\left(\theta'|d\right)=\mu\mathbb{E}_{p\left(\theta|d\right)}\left[-\frac{\gamma\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}+\frac{\phi\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}+\sum_{l\in D'}\Psi\left(l,\theta\right)+\lambda\right]$$

And the second term only depends on the action d:

$$-\mu \mathbb{E}_{p(\theta|d)}\left[\sum_{l\in D'} \Psi\left(l,\theta'\right)\right]$$

So that we can rewrite the above as:

$$b\left(d,\theta\right)=y\left(d,\theta\right)+B\left(\theta\right)+C\left(d\right)-\mu\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}$$

Where,

$$B\left(\theta\right)=\mu\sum_{l\in D'} \Psi\left(l,\theta\right)$$

$$C\left(d,\theta\right)=-\mu \mathbb{E}_{p(\theta|d)}\left[-\frac{\gamma\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}+\frac{\phi\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}-\frac{1}{u'\left(W+b\left(d,\theta\right)\right)}+\sum_{l\in D'} \Psi\left(l,\theta\right)+\lambda\right]$$

$$+\mu\frac{\phi\left(d,\theta\right)}{p\left(\theta\right)p\left(d|\theta\right)}+\mu \mathbb{E}_{p(\theta|d)}\left[\sum_{l\in D'} \Psi\left(l,\theta'\right)\right]$$

But for $b\left(d,\theta\right)>0$, and $b\left(d,\theta\right)<w\left(d,\theta\right)$ the limited liability and wealth constraints are not binding and then $\phi\left(d,\theta\right)=0$ and $w\left(d,\theta\right)=0$ so that $C\left(d,\theta\right)$ is only a function of the state d. Of course if ϕ binds the value of $b\left(d,\theta\right)$ would be 0 and similarly if $\gamma\left(d,\theta\right)$ binds, the value of $b\left(d,\theta\right)$ will be $w\left(d,\theta\right)$. The optimal contract to the original problem therefore is:

$$b\left(d,\theta\right)=\max\{0,min\{y\left(d,\theta\right)+B\left(\theta\right)+C\left(d\right)-\mu\frac{1}{u'\left(W+b\left(d,\theta\right)\right)},w\left(d,\theta\right)\}\}$$

For the log case and risk neutrality this becomes respectively:

$$b\left(d,\theta\right)=\max\{0,min\{\frac{1}{1+\mu}y\left(d,\theta\right)+B\left(\theta\right)+C\left(d\right),w\left(d,\theta\right)\}\}$$

$$b\left(d,\theta\right)=\max\{0,min\{y\left(d,\theta\right)+B\left(\theta\right)+C\left(d\right),w\left(d,\theta\right)\}\}$$

$\square$

Proposition 25

Proof. The proof is the same as the main theorem (and we use the same notation for multipliers) with the addition of the constraints:

$$b(d,\theta) = b(d,\theta') \text{ for all} d \in D, \theta \in \Theta \text{ and} \theta' \in \theta(d) \ [\zeta(d,\theta,\theta')]$$

to the principal's program. The KKT condition for $b(d,\theta)$ becomes:

$$-p(d)\,p(\theta|d) + \lambda(d,\theta) + \sum_{\theta'\in\theta(d)} \zeta(d,\theta,\theta') - \sum_{d'} \frac{\phi(d,d',\theta)\,u'(W + b(d,\theta))}{\mu} = 0$$

And then using the same manipulations from the main theorem:

$$b(d,\theta) = y(d,\theta) - \frac{\mu}{u'(W + b(d,\theta))} + \frac{\mu\lambda(b,\theta)}{u'(W + b(d,\theta))\,p(d)\,p(\theta|d)} + \frac{\tau(d)}{p(d)} + \frac{\mu\sum_{\theta'\in\theta(d)}\zeta(d,\theta,\theta')}{u'(W + b(d,\theta))\,p(d)\,p(\theta|d)} + \beta(\theta)$$

□

Because the $b()$ and $y()$ are the same over the each partition $\theta(d)$. Summing over all the $\theta^* \in \theta(d)$ and dividing by $|\theta(d)|$, yields:

$$b(d,\theta) = y(d,\theta) - \frac{\mu}{u'(W + b(d,\theta))} + \frac{1}{|\theta(d)|}\sum_{\theta^*\in\theta(d)} \frac{\mu\lambda(d,\theta^*)}{u'(W + b(d,\theta^*))\,p(d)\,p(\theta^*|d)}$$

$$+ \frac{\tau(d)}{p(d)} + \frac{1}{|\theta(d)|}\sum_{\theta^*\in\theta(d)} \frac{\mu\sum_{\theta'\in\theta(d)}\zeta(d,\theta^*,\theta')}{u'(W + b(d,\theta^*))\,p(d)\,p(\theta^*|d)} + \frac{1}{|\theta(d)|}\sum_{\theta^*\in\theta(d)} \beta(\theta^*)$$

And because the $\zeta(d,\theta,\theta')$ will elide with $\zeta(d,\theta',\theta)$ and we can apply the same arguments for $\lambda(d,\theta)$ as in the main theorem, we get the desired result.

Examples and other derivations

The main result can also be derived from the agent's solution without invoking the Caplin and Dean [2013] results. We can substitute the maximization condition for the agent with the first order KKT condition derived in proposition 14, and then we can apply KKT again to the new constrained problem obtaining the shape of the optimal contract as the solution.

Theorem 32. *The optimal contract for problem 15 under the assumption that the agent has log utility or risk neutral preferences is:*

$$b(d,\theta) = \max\{0, K \cdot y(d,\theta) + B(\theta) + C(d)\}$$

with K=1 for the risk neutral case and $K = \frac{1}{1+\mu}$ *for the log utility.*

Proof. For all actions $\theta \in \Theta$, we have that either $p(d|\theta) > 0$ for all $d \in D$ or $p(d|\theta) = 0$ $\forall d \in D$ because of proposition 14. Therefore if $(m, b(\cdot,\cdot))$ is a solution to problem 15 then $(m', b'())$, where $(m', b'())$ are the restriction of $(m, b(\cdot,\cdot))$ to the set D' of decisions that occur with positive probability, is the solution to:

$$\max_{m \in E(D'), b(\cdot,\cdot)} \mathbb{E}_{p(\theta)}[\mathbb{E}_{p(d|\theta)}[y(d,\theta) - b(d,\theta)]]$$

such that for all $d \in D'$ and $\theta \in \Theta$:

$$0 \leq p(d|\theta) \leq 1 \text{ and } \sum_{d' \in D'} p(d'|\theta) = 1$$

$$b(d,\theta) \geq 0 \;\; [\phi(d,\theta)]$$

$$u(W + b(d,\theta)) - \mu log\left(\frac{p(d|\theta)}{p(d)}\right) = \mu log\left(\sum_{d' \in D'} p(d') e^{\frac{u\left(W + b\left(d',\theta\right)\right)}{\mu}}\right) \; [\Psi(d,\theta)]$$

We prove that for this problem the optimal contract is $b'(d,\theta) = \max\{0, A \cdot y(d,\theta) + B(\theta) + C(d)\}$ $\forall d \in D'$. It then follows by limited liability and the utility of the principal that $b(d,\theta) = 0$ for all $d \in D \backslash D'$ and consequently for all $\theta \in \Theta$. For the rephrased subproblem among decisions occurring with positive probability, we apply the KKT theorem to express the gradient of the objective function at m, $b(\cdot,\cdot)$ as a linear combination of the gradients of the binding constraints at such points. The constraints for the probabilities will not be binding for decisions taken with positive probability as showed in the previous section, while the constraint on $\sum_{d' \in D'} p(d'|\theta) = 1$ is already embedded in the third constraint. Therefore letting $\phi(d,\theta)$ and $\Psi(d,\theta)$ be the Lagrange multipliers for the potentially binding constraints and expressing the maximization with respect to $u(W + b(d,\theta))$ instead of $b(d,\theta)$ we get:

$$-\frac{1}{u'(W + b(d,\theta))} p(\theta) p(d|\theta) + \phi(d,\theta) - p(\theta)\Psi(d,\theta) + \mu p(\theta) \sum_{d' \in D'} \frac{p(d) e^{\frac{u\left(W + b\left(d',\theta\right)\right)}{\mu}}}{\sum_{d' \in D'} p(d') e^{\frac{u(W + b(d',\theta))}{\mu}}} \frac{1}{\mu} \Psi(d,\theta) = 0$$

and after some rearrangement we get:

$$\frac{1}{u'(W+b(d,\theta))} = \frac{\phi(d,\theta)}{p(\theta)\,p(d|\theta)} - \frac{\Psi(d,\theta)}{p(d|\theta)} + \sum_{l\in D'} \Psi(l,\theta) \tag{11}$$

We get the following condition from $p(d|\theta)$:[18]

$$\frac{y(d,\theta)-b(d,\theta)}{\mu} = \sum_{\theta'} \Psi(d,\theta')\frac{p(\theta')}{p(d)} - \frac{\Psi(d,\theta)}{p(d|\theta)} - \sum_{\theta'\in\Theta}\sum_{l\in D'} \Psi(l,\theta')\frac{p(\theta')\,p(d|\theta')}{p(d)} \tag{12}$$

Now substituting (11) into (12). We have:

$$\frac{y(d,\theta)-b(d,\theta)}{\mu} =$$

$$\sum_{\theta'} \Psi(d,\theta')\frac{p(\theta')}{p(d)} - \sum_{\theta'\in\Theta}\sum_{l\in D'} \Psi(l,\theta')\frac{p(\theta')\,p(d|\theta')}{p(d)} + \frac{1}{u'(W+b(d,\theta))} - \frac{\phi(d,\theta)}{p(\theta)\,p(d|\theta)} - \sum_{l\in D'} \Psi(l,\theta)$$

We can also rearrange terms by adding multiplying and diving the first term by $p(d|\theta')$ and recognizing that we get a posterior $p(\theta'|d)$, we can rewrite:

$$y(d,\theta)-b(d,\theta) =$$

$$\mu\sum_{\theta'} \frac{\Psi(d,\theta')}{p(d|\theta')}p(\theta'|d) - \mu\sum_{\theta'\in\Theta}\sum_{l\in D'} \Psi(l,\theta')\,p(\theta'|d) + \mu\frac{1}{u'(W+b(d,\theta))} - \mu\frac{\phi(d,\theta)}{p(\theta)\,p(d|\theta)} - \mu\sum_{l\in D'} \Psi(l,\theta)$$

Now consider that the first two terms are equal to:

$$\mu\sum_{\theta'} \frac{\Psi(d,\theta')}{p(d|\theta')}p(\theta'|d) = \mu\mathbb{E}_{p(\theta|d)}\left[\frac{\phi(d,\theta)}{p(\theta)\,p(d|\theta)} - \frac{1}{u'(W+b(d,\theta))} + \sum_{l\in D'} \Psi(l,\theta)\right]$$

$$-\mu\mathbb{E}_{p(\theta|d)}\left[\sum_{l\in D'} \Psi(l,\theta')\right] = \mathbb{E}_{p(\theta|d)}\left[y(d,\theta)-b(d,\theta)\right]$$

So that we can rewrite the above as:

$$b(d,\theta) = y(d,\theta) + B(\theta) + C(d,\theta) - \mu\frac{1}{u'(W+b(d,\theta))}$$

Where,

[18]Taking care to differentiate all the constraints for all d and θ.

$$B(\theta) = \mu \sum_{l \in D'} \Psi(l, \theta)$$

$$C(d, \theta) = -\mu \mathbb{E}_{p(\theta|d)} \left[\frac{\phi(d, \theta)}{p(\theta) p(d|\theta)} - \frac{1}{u'(W + b(d, \theta))} + \sum_{l \in D'} \Psi(l, \theta) \right] + \mu \frac{\phi(d, \theta)}{p(\theta) p(d|\theta)} - \mathbb{E}_{p(\theta|d)} [y(d, \theta) - b(d, \theta)]$$

But for $b(d, \theta) > 0$, the limited liability is not binding and then $\phi(d, \theta) = 0$ so that $C(d, \theta)$ is only a function of the state d. So that the optimal contract to the original problem is:

$$b(d, \theta) = \max\{0, y(d, \theta) + B(\theta) + C(d) - \mu \frac{1}{u'(W + b(d, \theta))}\}$$

For the log case and risk neutrality this becomes respectively:

$$b(d, \theta) = \max\{0, \frac{1}{1 + \mu} y(d, \theta) + B(\theta) + C(d)\}$$

$$b(d, \theta) = \max\{0, y(d, \theta) + B(\theta) + C(d)\}$$

□

Example 33. Consider the random variable X defined over outcome space $X = \{a, b, c, d\}$ with $p(a) = p(b) = \frac{1}{8}$, $p(c) = \frac{1}{4}, p(d) = \frac{1}{2}$. If we were to write a binary code to transmit this information we could associate $a \to 1$, $b \to 01$, $c \to 001$ and $d \to 0001$. Now this code would clearly be binary and self punctuating - as upon receiving the code 1 we know that we need not wait for further signals. Nevertheless it would not be the most efficient way to transmit information. The expected number of transmitted digits for such code would be:

$$\frac{1}{8} 1 + \frac{1}{8} 2 + \frac{1}{4} 3 + \frac{1}{2} 4 = \frac{25}{8}$$

It would be more efficient to associate the shorter binary code to the messages that occurs more frequently, for instance using $a \to 0001, b \to 001, c \to 01, d \to 1$. So that the resulting expected length is:

$$\frac{1}{8} 4 + \frac{1}{8} 3 + \frac{1}{4} 2 + \frac{1}{2} 1 = \frac{15}{8}$$

Notice that the entropy of this random variable is:

$$H(x) = -\left(\frac{1}{8}log\left(\frac{1}{8}\right) + \frac{1}{8}log(\frac{1}{8}) + \frac{1}{4}log(\frac{1}{4}) + \frac{1}{2}log(\frac{1}{2})\right) = \frac{7}{4}bits = \frac{14}{8}bits.$$

The theorem by Shannon, shows that the minimum expected number of binary signals required to determine X is between $H(X)$ and $H(X)+1$. Entropy as a measure of uncertainty captures the expected length of the most efficient self-punctuating binary encoding.

Discussion of reduction in variance as a measure of uncertainty One could think that Variance of a random variable could also good measure of the uncertainty and some authors have proposed so (see Gentzkow and Kamenica [2014]). While this intuition is correct for normal variables, it is not very compelling for discrete variables. Variance is a measure of dispersion - how far apart the values of a random variable are from its mean - not of uncertainty - how difficult it is to transmit/acquire information about the realization. Consider the example of a Bernoulli variable, its variance will be $p(1-p)$. Now if we multiply the Bernoulli variable by a constant, say 10, nothing changes in the underlying uncertainty. We have just relabeled one of the outcomes from 1 to 10. Consistently with this fact, the entropy of such a random variable would not change while the variance would be increased by a factor of a 100. The fact that a dispersion measure such as variance is not ideal to measure "uncertainty", and therefore information, does not imply that the two concepts are unrelated. Mechanically, if the variance is 0 and there is no dispersion, then the variable is degenerate and it has 0 entropy/uncertainty as well. The next example shows why how for normal variables entropy and variance are related.

Example 34. Gaussian state of the world and gaussian signal with mutual information:

Consider a model where both the state θ and the signal X are continuous variables over $\mathbb{R}$, and suppose we restrict to the case of jointly normal variables. Then given that the entropy of a gaussian variable $X \sim N(\mu, \sigma^2)$ is $ln(2\pi e\sigma^2)$, we get that the cost of mutual information is $ln\left(\frac{1}{1-\rho^2_{e\theta}}\right)$. This can be derived from the fact that the conditional distribution of two jointly normal variables is $\theta|X \sim \mathcal{N}\left(\mu_\theta + \frac{\sigma_\theta}{\sigma}\rho(x-\mu), (1-\rho^2)\sigma^2_\theta\right)$.

By the above derivation of the mutual information $I(\theta, X) = H(\theta) - H(\theta|X)$ and therefore $ln(2\pi e\sigma^2_\theta) - ln\left(2\pi e\sigma^2_{\theta|X}\right) = ln\left(\frac{\sigma^2}{\sigma^2_{\theta|X}}\right) = ln\left(\frac{1}{1-\rho^2_{e\theta}}\right)$.

Appendix for Chapter 2

Proof

Remark 27

Proof. Given that by assumption C is never a likely contender, if a voter can be pivotal it has to be that she is pivotal between her favorite candidate and another. It is therefore a dominant strategy (δ_ι is also positive) to choose $r(1)$. □

Proposition 28

Proof. If a C supporter is pivotal this means that the total votes for his $r(2)$, denoted as $n_{r(2)}$ are either: $n_{r(2)} = n_{r(3)}$ or $n_{r(2)} = n_{r(3)} - 1$. In either case, assuming that ties are broken by the flip of a fair coin, the marginal benefit of switching vote from C to $r(2)$ in ballot M, is given by

$$\frac{1}{2}\left(\nu^z_{r(2)} - \nu^z_{r(3)}\right)\pi^M(\alpha_M, \beta_M)$$

and the marginal cost of doing so is δ_ι.

Therefore for voter of type z[19] $\overline{\delta}^{M,j} = \frac{1}{2}\left(\nu^z_{r(2)} - \nu^z_{r(3)}\right)\pi^M(\alpha_j, \beta_j)$. □

Proposition 29

Proof. From Proposition 28, we know that $\overline{\delta}^{N,j} = \frac{1}{2}\left(\nu^z_{r(2)} - \nu^z_{r(3)}\right)\pi^N(\alpha, \beta) = \overline{\delta}^{N,l}$. Then equation (2.1) becomes $\left(F\left(\overline{\delta}^{j,j}\right) - F\left(\overline{\delta}^{l,l}\right)\right)$ and because $\alpha_j > \alpha_l$, and because $\pi\,()$is decreasing in α. We have $\overline{\delta}^{l,l} > \overline{\delta}^{j,j}$. And because F is the same across regions the result follows. □

[19]Remember that the types here are just one per possible preference ranking.

Tables:

Table 6: Poll Results

Region	Center-Left	Center-Right	Region	Center-Left	Center-Right
ABRUZZO	37.4%	27.9%	MARCHE	40.8	21.8%
BASILICATA	38.5%	23.1%	MOLISE	34.6%	26.7%
CALABRIA	35.4%	24.8%	PIEMONTE	31.2%	29%
CAMPANIA	30%	29.6%	PUGLIA	30.4%	27.2%
EMILIA-ROMAGNA	44.1%	22.1%	SARDEGNA	41.9%	28.5%
FRIULI	32.4%	31.8%	SICILIA	24.3%	31.8%
LAZIO	34.6%	27.8%	TOSCANA	48.4%	26.9%
LIGURIA	36.1%	25.6%	UMBRIA	41.6%	24%
LOMBARDIA	31.7%	31.7%	VENETO	27.1%	38.8%

Notes: The Table presents the share of votes that were expected to go to each of the top two Coalitions according to the Regional poll run by Tecne' for Sky. The paper presents more information about the nature of the poll and the source of this information.

Table 8: Controls Lombardy vs. Emilia-Romagna

		Border	10km	15km	20km
Outcomes		(1)	(2)	(3)	(4)
Diff. Voters %	$Toss-up\,Region$	-0.213	-0.279	-0.254	-0.309
		(0.233)	(0.269)	(0.233)	(0.245)
Income P.C.		-0.599	-0.949	-0.661	-0.415
		(0.872)	(1.217)	(1.069)	(1.085)
Pop.		-0.0961	-0.236	-0.129	0.0912
		(0.244)	(0.351)	(0.313)	(0.340)
Retired %		-0.529	-0.912	-0.0477	-0.905
		(0.787)	(1.255)	(1.066)	(1.037)
	Obs.	83	181	264	347

Notes: This Table present the result of a linear local regression model around the border of Lombardy and Emilia-Romagna. Column (1) uses only Municipalities at the border. Between Column (2) and (4) we use a bandwidth of 5, 7.5 and 10 miles respectively. The outcomes are reported in the first Column, and in particular they are difference in votes, Income per Capita, Population, Retired pop. %, as previously described. The main variable $Toss-up\,Region$ is a dummy equal to one for Municipalities in Lombardy, which is the most Toss-up and pivotal Region in Italy. The regression is weighted by the total number of votes in the Congress at the border. All regressions include a constant and standard errors are clustered at level of the border. *** denotes significance at the 1% level, ** at the 5%, and * at the 10%.

Table 7: Border Regressions

			$\Delta^{C-S}_{\%ij}$				
		Border Municipalities			10km	15km	20km
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$Toss-up\,Region$	-1.447***	-1.557***	-1.215***	-1.508***	-1.552***	-1.599***	-1.632***
	(0.147)	(0.126)	(0.314)	(0.364)	(0.190)	(0.141)	(0.128)
$f(Distance)$			Y	Y	Y	Y	Y
Controls		Y		Y	Y	Y	Y
Observations	83	83	83	83	181	264	347

Notes: Each column presents a different city-level regression, where the outcome is the difference in votes between the Congress and the Senate for the two major coalitions (center-right and center-left), scaled by the total votes in the Senate. The first four Columns is using only Municipalities exactly at the border between Emilia-Romagna and Lombardy. Columns (5) to (7) uses Municipalities between 10km, 15km,20km respectively. The outcome variable is constructed from electoral data, provided by the Archive of the Italian Department of State ("Ministero dell'Interno"). The main variable $Toss-up\,Region$ is a dummy equal to one for Municipalities in Lombardy, which is the most Toss-up and pivotal Region in Italy. In Column (1) we do not add any control. In Column (2) we add a linear function of the distance to the relevant border, with different slope per side. In Column (3) and (4), we repeat the previous two adding also controls. In (5) to (7) we add both controls and distance. For controls, we mean a set of standard covariates (Income per Capita, Population, Retired pop. %, and difference in votes). Variable description is in the Appendix. The regression is weighted by the total number of votes in the Congress. All regressions include a constant and standard errors are heteroskedasticity robust. *** denotes significance at the 1% level, ** at the 5%, and * at the 10%.

Table 9: National Regression Discontinuity

	$\Delta^{C-S}_{\%_k}$					
	10km		15km		20km	
	(1)	(2)	(3)	(4)	(5)	(6)
$Toss-up$	6.560***	7.055***	6.376***	6.273***	4.399**	3.880**
	(1.172)	(1.591)	(1.072)	(1.635)	(1.752)	(1.537)
$Controls$		Y		Y		Y
$Border\,F.E.$	Y	Y	Y	Y	Y	Y
Obs.	50	50	54	54	54	54

Notes: This Table present the result of the second-stage of the Regression Discontinuity model, as in Pinkovskiy(2014), discussed in the paper. The outcome is the estimate of the conditional expectation at the border of the difference in votes between the Congress and the Senate for the two major coalitions (center-right and center-left), scaled by the total votes in the Senate. The standard errors are clustered at the regional level and the second stage has a border FE as discussed in the paper. Therefore every observation is measured at border-side level, and it is estimated in a first stage, as described in the paper. Given that there are 27 regional border under analysis, the total number of observation is in general 54. The variable Toss-up is a score equal to the absolute value of the difference between the top two Coalitions, measured at Regional level, according to the latest pool available before the election (Sky). In Columns (1), (3) and (5) we report the simple regression. In Columns (2), (4) and (6) we augment it with the controls. The controls are obtained from a first stage equivalent to the outcome variable, and in particular we consider the usual set of standard covariates (Income per Capita, Population, Retired pop. %, and difference in votes). The outcome variable is constructed from electoral data, provided by the Archive of the Italian Department of State ("Ministero dell'Interno"). The regression is weighted by the total number of votes in the Congress at the border. All regressions include a constant and standard errors are clustered at level of the region. *** denotes significance at the 1% level, ** at the 5%, and * at the 10%.

Table 10: Balancing Controls in National RD

		10km	15km	20km
Outcomes		(1)	(2)	(3)
Diff. Voters %	$Toss-up$	-1.551	-1.941	-0.230
		(3.333)	(2.169)	(1.760)
Income P.C.		1.810	2.472**	2.564***
		(1.524)	(0.974)	(0.770)
Pop.		0.138	0.255***	0.187*
		(0.148)	(0.0617)	(0.0889)
Retired %		6.105	6.389	6.864
		(5.740)	(5.627)	(5.166)
Pop. 18-24%		0.00163	-0.00294	0.0162*
		(0.0140)	(0.0111)	(0.00919)
$Congress_{Top2}\%$		-9.759	-12.06*	-13.94**
		(5.618)	(6.592)	(5.152)
$Senate_{Top2}\%$		-16.32**	-18.44**	-18.34**
		(5.822)	(6.912)	(6.265)
	Obs.	54	54	54

Notes: This Table present the result of the second-stage of the Regression Discontinuity model, as in Pinkovskiy(2014), discussed in the paper, but estimated on other outcomes, as a control, with fixed effect at border. The outcomes are reported in the first Column, and in particular they are difference in votes, Income per Capita, Population, Retired pop. %,Share of population between 18-24 over total population (2012), Votes for top 2 parties in the Congress, votes for top 2 parties in Senate. The two-stage estimation is completely identical to the one used for the main specification. As before every observation is measured at border-side level. Given that there are 27 regional border under analysis, the total number of observation is in general 54. When considering a bandwidth from the border in the first stage of 5 miles only, we had to drop 2 borders because of lack of observations in the first stage (in particular, the border between Emilia-Romagna and Piemonte and the border between Lazio and Marche. The variable Toss-up is a score equal to the absolute value of the difference between the top two Coalitions, measured at Regional level, according to the latest pool available before the election (Sky).The regression is weighted by the total number of votes in the Congress at the region. All regressions include a constant and standard errors are clustered at level of the border. *** denotes significance at the 1% level, ** at the 5%, and * at the 10%.

Table 11: Border Regressions EL Placebo

	$\Delta^{C-S}_{\%ij}$						
	Border Municipalities				10km	15km	20km
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$Toss-up\,Region$	-0.258*	-0.295**	-0.0197	-0.320	-0.314	-0.233*	0.174
	(0.151)	(0.117)	(0.316)	(0.313)	(0.192)	(0.141)	(0.401)
$f(Distance)$			Y	Y	Y	Y	Y
Controls		Y		Y	Y	Y	Y
Observations	83	83	83	83	181	264	347

Notes: Each column presents a different city-level regression, where the outcome is the difference in votes between the Congress and the Senate for the two major coalitions (center-right and center-left), scaled by the total votes in the Chamber. The first four Columns is using only Municipalities exactly at the border between Emilia-Romagna and Lombardy. Columns (5) to (7) uses Municipalities between 10km, 15km,20km respectively. The outcome variable is constructed from electoral data, provided by the Archive of the Italian Department of State ("Ministero dell'Interno") for the 2008 elections. The main variable $Toss-up\,Region$ is a dummy equal to one for Municipalities in Lombardy, which is the most Toss-up and pivotal Region in Italy both for 2013 and 2008 but much less so in 2008. In Column (1) we do not add any control. In Column (2) we add a linear function of the distance to the relevant border, with different slope per side. In Column (3) and (4), we repeat the previous two adding also controls. In (5) to (7) we add both controls and distance. For controls, we mean a set of standard covariates (Income per Capita, Population, Retired pop. %, and difference in votes). Variable description is in the Appendix. The regression is weighted by the total number of votes in the Congress. All regressions include a constant and standard errors are heteroskedasticity robust. *** denotes significance at the 1% level, ** at the 5%, and * at the 10%.

Table 12: National Regression Discontinuity Placebo Test

	$\Delta^{C-S}_{\%_k}$					
	10km		15km		20km	
	(1)	(2)	(3)	(4)	(5)	(6)
$Toss-up$	-0.857	-0.490	-0.588	-0.318	-0.336	-0.521
	(2.192)	(2.193)	(1.480)	(1.368)	(1.439)	(1.244)
$Controls$		Y		Y		Y
Obs.	50	50	54	54	54	54

Notes: This Table present the result of the second-stage of the Regression Discontinuity model, as in Pinkovskiy(2014), discussed in the paper. The outcome is the estimate of the conditional expectation at the border of the difference in votes between the Congress and the Senate for the two major coalitions (center-right and center-left) in the 2008 elections, scaled by the total votes in the Senate. The standard errors are clustered at the regional level and the second stage has a border FE as discussed in the paper. Therefore every observation is measured at border-side level, and it is estimated in a first stage, as described in the paper. Given that there are 27 regional border under analysis, the total number of observation is in general 54. The variable Toss-up is a score equal to the absolute value of the difference between the top two Coalitions in 2013, measured at Regional level, according to the latest pool available before the election (Sky). In Columns (1), (3) and (5) we report the simple regression. In Columns (2), (4) and (6) we augment it with the controls. The controls are obtained from a first stage equivalent to the outcome variable, and in particular we consider the usual set of standard covariates (Income per Capita, Population, Retired pop. %, and difference in votes). The outcome variable is constructed from electoral data, provided by the Archive of the Italian Department of State ("Ministero dell'Interno"). The regression is weighted by the total number of votes in the Congress at the border. All regressions include a constant and standard errors are clustered at level of the region. *** denotes significance at the 1% level, ** at the 5%, and * at the 10%.

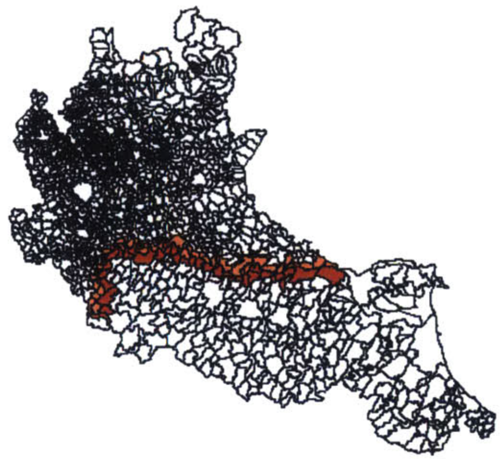
Table 13: Robustness

	$\Delta^{C-S}_{\%_k}$					
	Standard -20km			*Bootstrap* -20km		
	(1)	(2)	(3)	(4)	(5)	(6)
$Toss-up$	-3.36*	-3.89***	-4.71**	-2.56***	-3.02***	-4.54**
	(1.87)	(0.86)	(2.260)	(0.77)	(0.93)	(2.24)
Controls		Y	Y		Y	Y
Border F.E.			Y			Y
Obs.	54	54	54	54	54	54

Notes: This Table presents two Robustness to our main result, using the two step methodology discussed above. a national Regression, where only Municipalities within a certain bandwidth from the border are employed. The outcome is the estimate of the conditional expectation at the border of the difference in votes between the Congress and the Senate for the two major coalitions (center-right and center-left), scaled by the total votes in the Senate. Therefore every observation is measured at border-side level, and it is estimated in a first stage, as described in the paper. Given that there are 27 regional border under analysis, the total number of observation is in general 54. Results are here presented at a 20km bandwidth. In Column (1) and (2), we present the usual results using a 10 mile bandwidth. We report this here for sake of clarity in the comparison with the robustness. In Column (3), we add a border F.E. to the specification in Column (2). We therefore have a total of 27 F.E.. In Column (4)-(6), we repeat exactly the same procedure as in (1)-(3), but the estimates are obtained using the two-stage boostrap procedure, which is described in detail in the paper. All regressions include a constant and standard errors are clustered at level of the border. *** denotes significance at the 1% level, ** at the 5%, and * at the 10%.

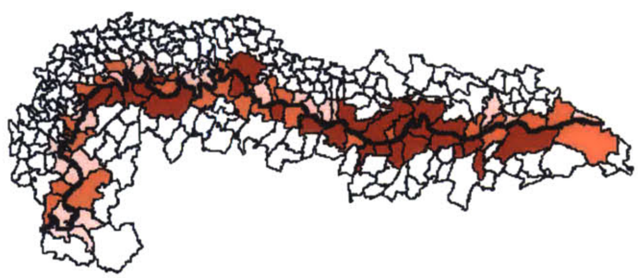
Figures

Figure -7: Municipalities at the border between Lombardy and Emilia-Romagna



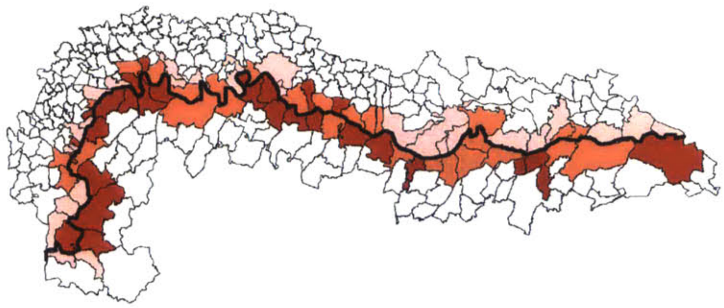
Notes: this map contains all the Municipalities in Lombardy (north) and Emilia-Romagna (south). The border of each Municipality is defined by the black lines. Furthermore, we highlight with red color the sub-set of Municipalities that are at the border between the two.

Figure -8: Distribution of Income by Municipality, Lombardy vs. Emilia-Romagna



Notes: this map contains all the Municipalities in Lombardy or Emilia-Romagna, that are within 10km from the other Region border. Colored are the Municipalities that are on the border. A darker color signal higher level of income per capita in that Municipality. Data on income are provided by ISTAT. The black line is the border between the two Regions.

Figure -9: Distribution of the outcome by Municipality, Lombardy vs. Emilia-Romagna



Notes: this map contains all the Municipalities in Lombardy or Emilia-Romagna, that are within 10km from the other Region border. Colored are the Municipalities that are on the border. A darker color signal higher level of the outcome variable, which is the difference in the share of votes going to one of the top-two Coalitions between Congress and Senate at Municipality level. Data on income are provided by Italian Department of State. The black line is the border between the two Regions.

Appendix for Chapter 3

Tables

	(1)	(3)	(5)
	vol11	vol13	vol15
VARIABLES	vol3D_MS	vol3D_MON	vol3D_MED
vol3D_MXEU0MD	0.393***	0.380***	
	(0.0322)	(0.0330)	
vol3D_FTSEMIB	0.438***	0.285***	0.717***
	(0.0320)	(0.0356)	(0.0527)
elezioni	-0.0306	0.00366	0.112
	(0.0252)	(0.0480)	(0.0842)
vol3D_BEFINC			0.115***
			(0.0431)
Constant	0.0992***	0.126***	0.144***
	(0.00529)	(0.00557)	(0.00638)
Observations	3,571	3,571	3,565
R-squared	0.306	0.200	0.299

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 14: Average 5 Days Volatility against election dates

	(1)	(3)	(5)
	vol11	vol13	vol15
VARIABLES	vol5D_MS	vol5D_MON	vol5D_MED
vol5D_MXEU0MD	0.530***	0.475***	
	(0.0285)	(0.0284)	
vol5D_FTSEMIB	0.446***	0.387***	0.927***
	(0.0282)	(0.0309)	(0.0507)
elezioni	-0.00208	0.00649	0.0657
	(0.0287)	(0.0393)	(0.0560)
vol5D_BEFINC			0.00619
			(0.0393)
Constant	0.0895***	0.111***	0.148***
	(0.00496)	(0.00485)	(0.00561)
Observations	3,569	3,569	3,563
R-squared	0.468	0.367	0.422

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 15: Average 5 Days Volatility against election dates