Università degli Studi di Milano Bicocca

DEFAP - Doctorate School in Public Economics

Ph.D. dissertation

Essays on Networks and Information Dynamics

Candidate
Pietro Battiston

Supervisor
Prof. Luca Stanca

Academic year 2013/2014

Introduction

The theory of networks has been growing at a very fast pace in the last 20 years, in terms of scholars working in the field, of the quality and complexity of theoretical instruments adopted, and of the availability of empirical data.

The first three chapters of this dissertation explore different areas in which networks are being used as fundamental tools to study information dynamics and their economic relevance.

In "*Citations are Forever: Modeling Constrained Network Formation*", I consider a model of constrained network formation, and employ it to identify a lower bound to the relevance of environmental factors on citation flows to scientific articles.

In "*Opinion Dynamics in Social Networks: Theory and Experimental Evidence*" (joint work with Luca Stanca), we consider a new model of opinion formation over a social network, and test it in an experimental setting.

In "*Not that Fundamental: Bubbles and Financial Networks*", I highlight the possibility of rational asset price bubbles in financial markets with finite number of agents and amount of wealth, and sketch some consequences for systemic risk, in light of the results of recently proposed models of financial networks.

In "*The Impact of Peer Pressure on Tax Compliance: a Field Experiment*" (joint work with Simona Gamba), the social network in which agents move is only implicit, but the information flow is clear: the experimental design allows us to manipulate the sellers' perception about social norms on tax compliance, and to identify an effect on their own propensity towards compliance.

Contents

Chapter 1

Citations are Forever: Modeling Constrained Network Formation

"Always go to other people's funerals; otherwise they won't go to yours."
Yogi Berra (confused by a typical "constrained growth" network)

1.1 Introduction

In the last 20 years, the theory of networks has been recognized an important role in explaining the formation and functioning of social and economic settings in which *relationships* among agents under observation are of fundamental importance. In particular, several models of network formation were developed targeting the understanding of how the structure of a network is endogenously determined by the characteristics of the nodes, whose tendency to form and break links depends on given parameters (typically, the cost of creating/keeping alive a link, compared to the utility received from becoming - directly or indirectly - connected to some other nodes). A stream of literature, starting from the seminal work of Bala and Goyal (2000), has developed focusing on a *noncooperative* approach, in which the choice of adding a link is made independently by one of the two nodes concerned by it, and only that node bears the cost, although both nodes can potentially benefit from such link. Based on this framework, a definition of *stability* can be given, typically based on the concept of *pairwise Nash* equilibrium (such as in Galeotti, 2006 and Haller et al., 2007), or some refinement of it (for instance Dutta and Mutuswami, 1997 consider *coalition* choices, while the concept of *"farsightedly stable networks"* formulated by Herings et al., 2009 assumes that nodes have a longer horizon of reasoning).

Those studies share the implicit assumption that links can be added and destroyed freely (though at some cost). Even experimental works on endogenous network formation have most often been based on the assumption that participants can *at any point in time* - or at least *repeatedly* - decide to create/break a link (see for instance Goeree et al., 2009 and Kirchsteiger et al., 2011). This is a very natural setup for several reasons. Firstly, many real world networks are indeed characterized by links which are at least potentially volatile (i.e. computer networks, social relationships...). Secondly, even in cases in which the network under study is characterized by exogenous restrictions, such as geographic ones, the interest of researchers, as well as the available data, is often focused on an inherently volatile *flow* of some good (such as influence, or information) over it. Thirdly, even considering networks which are typically characterized by a *stratification* of links over time (such as connections in Internet social networks, or the network of roads between cities), most databases in use are *snapshots* of the network at given points in time, hence discarding its temporal evolution, and allowing the researchers

to focus only on *static* analysis.

Only very recently a form of *temporal hierarchy* of nodes/links has been considered in the context of such theories, by Haller (2012). His study provides interesting conclusions applying to networks which are endogenously formed around an exogenously given subset of links. That subset is shown to potentially change drastically the existence, numerosity, stability and efficiency of pairwise Nash equilibria. An interesting insight is that *backbone infrastructures*, that is, sets of links which are guaranteed to exist independently from individual incentives, are *restrictions* to the set of possible actions available to nodes, which however can bring important *welfare improvements*. The present study aims at generalizing the analysis to the *repeated* addition of nodes and/or links, under positive *and negative* constraints. Differently from the work of Haller, the set of guaranteed/forbidden links will not necessarily be exogenously given, but can come instead from the previous iteration of the network formation process.

In Section 1.3, I will then adopt the new setting to model the network of citations among scientific publications, and test some of its stylized predictions. The scientometric literature, although relatively young, has developed extremely in the last decades.[1] The first bibliometric studies, such as the work by Gross and Gross (1927), were motivated by the practical motivation of determining which scientific works ought to be present in a scientific library in order to satisfy the needs of scholars and of faculty members. The idea that the encoding of citations between scientific papers could be of use to researchers themselves, and in particular that a *citation index* could prove of tremendous utility to them, is due to Garfield (1955), who went as forward as to propose the term "impact factor". Most importantly, he created the first implementation of such an index, the evolution of which is today the *ISI Web of Knowledge*. The stream of literature qualifying and quantifying the different roles that a citations can have in enhancing the quality of exposition, or trustworthiness, of a scientific paper can also be traced back to the work of Garfield et al. (1965), who proposed a classification, admittedly incomplete, of fifteen different motivations which can explain why a scientific paper cites another one. Such classification includes categories for *negative* citations, aimed at *"disclaiming work or ideas of others"* or *"disputing priority claims of others"*, which can be seen today as an implicit criticism to the *normative* view, according to which instead citations are a mere way to attribute credit to previous works, or even a reward for the research work done by their authors. It is however distinct from the alternative *constructive* view, according to which citations are only the result of *strategic* decisions, tar-

[1] This selection of works is largely based on the review by Baccini (2010).

geted at gaining *"a dominant position in their scientific community"* (Moed and Garfield, 2004). The approach taken in the present study is also, in principle, agnostic with respect to those main views, in the sense that it does not recognize citations as the mere consequence of a norm, but neither assumes they are necessarily taken strategically. For instance, the results which will be exposed are consistent with the hypothesis that citations are mainly a way to attribute credit, but that the choice an author makes of papers to cite is influenced by environmental factors. On the other hand, the *Matthew effect* which is identified can be partly explained by the fact that most cited papers are also the most authoritative, and hence the ones which can better *persuade* the reader. In any case, the consequences go in the direction of reconsidering the value which can be attributed to bibliometric indicators as measures of scientific productivity and/or impact.

From the empirical point of view, the analysis of the *scientific network* - a broad term which can refer to the networks defined by several different *relationships* characterizing the functioning of the academia - has been the subject of many studies. Virtually all of them employ data related to scientific journals and publications on them: in particular, many focus on the *co-authorship* relation between researchers (see Goyal et al., 2006, Cainelli et al., 2010 and De Stefano et al., 2013 for recent examples), or on the links between journals (for instance Baccini and Barabesi, 2010 considers the relation "having a non-empty intersection between members of editorial boards", while West et al., 2010 discuss the *Eigenfactor*, a new method for ranking journals by importance, based on the network of citations between them).

The analysis of the network of citations *between paper themselves* also has a long history, dating at least to de Solla Price (1965), who estimated some of its relevant statistics, focusing in particular on the very skewed distribution of indegree (when compared to the outdegree).[2] Some attempts have also been targeted at extracting from the characteristics of the network some insights on the behavioral aspects of the act of citing, such as in Baldi (1998). This approach is however hampered by huge endogeneity issues. The problem is worsened by the tendency to attribute citations an important role as a measure of research *impact* and/or *quality*: to the "natural" behavioral aspects of the citation choice, one must add the incentives to act *strategically* in the choice of papers to cite (i.e. citing papers from a given journal in order to increase the chances of getting published, citing papers from a given author in order to be cited/positively referred in return).

Probably because of the historical difficulties in obtaining and analyzing

[2]The *indegree* of a given paper is, here and in the following, the number of papers citing it, while the outdegree is the number of papers it cites.

entire citational databases, most works on the network of citations have also been limited to the observation of *strictly local* properties - typically, correlations among different characteristics on a *per node* (paper) basis, or at most among characteristics of the citing paper and the cited paper. Section 1.3 and subsequent ones of the present study, although still based on local properties, go beyond this limit, adopting an empirical strategy based on particularly defined sub-networks of diameter 3, and exploiting information about authors and journals. The aim is to provide a *causal identification* of the importance of environmental factors on the citing process (more precisely, establishing a lower bound to the importance of such effects). This approach is complementary to the literature on statistical properties of bibliometric indicators, such as the analysis of the h-index by Pratelli et al. (2012). Their model accounts for the strong and intrinsic non-independence of citation flows across time, but does not consider the possibility of effects such as the *fame* of the author, or even strategic citing behavior, influencing the *citation flows*. Having a lower bound on the determinants of citations which are entirely unrelated to the *content* of articles provides an additional error term to be considered *on top* of what they find, and can have important policy implications concerning the use of bibliometric indicators for the distribution of resources among publicly funded research facilities (and, in general, research institutions targeting scientific goals not limited to the mere *impact* on the academia).

The fame effect was already considered in several studies. As suggested by de Solla Price (1976), "success breeds success": the more a paper is known (/cited), the more its fame (/flow of citations) will grow in the future, and the same can hence be said for an *author*. However, while this is undoubtedly true from the *predictive* point of view (a clear correlation between past and future flow of citations has been found in the empirical literature), it is particularly hard to identify and quantify the *causal effect*, which often goes under the name of *Matthew effect* (Merton, 1968). The scope of my work is hence precisely to extract from bibliometric data some evidence of *herding*: may papers gain popularity (citations) not (just) because of their content but because of some *environmental* characteristics which focus on them the attention of the scientific community? This can well be a circular process (the more they are cited, the more they will be read, the more they will be cited). Notice the term "herding" does not necessarily imply any irrationality on the behalf of agents (in the same way in which it does not when used in the context of the financial market): it can be the consequence of bounded rationality, but also of limited information (in particular *costly* information - the duty of keeping at pace with the existing literature typically takes up a relevant share of the work time of a researcher), and, as already mentioned,

strategic behavior.

The literature cited so far, in particular with respect to the debate between the normative and the constructive views, leaves no doubt that investigating the factors which affect the citation behavior, and discussing its possible interpretations, is not a novel idea. The contribution I make to the empirical literature consists however in providing a comprehensive *measure*, or at least a lower bound, to the importance of *environmental variables*, which is not based on merely anecdotal evidence, and which can be attributed a causal meaning. To the best of my knowledge, the only other successful attempt at identifying a causal effect of fame on citation patterns is due to Azoulay et al. (2014), who find a clear, although short-lived, increase in the inflow of citations after a scientist is appointed the title of Howard Hughes Medical Investigator. While their approach, compared to the one developed in the present paper, has the clear advantage that such appointment has a simple and unambiguous interpretation in terms of publicity and fame, it is also intrinsically limited to a small population of scientists of a specific field. Instead, the approach described in sections from 1.3 to 1.6 of the present study can in principle be applied to any scientific field, or even to subpopulations of researchers defined according to various criteria (nationality, other proxies of fame...)

1.2 The model

The fundamental building block adopted for the formalization of the constrained network formation is the model of Galeotti et al. (2006). A network is composed by $N = \{1, \ldots, n\}$ nodes: for each pair of nodes i, j, a cost parameter $c_{ij} > 0$ and a value parameter $v_{ij} > 0$ are given. A *directed* network g is formally a set of pairs of nodes: if a pair (i, j) is in g, we say that i *sponsors* a link to j, and we write $g_{ij} = 1$. $\bar{g}$ represents the corresponding *undirected* network, that is, the smallest network containing g and (j, i) for each (i, j) contained in g. Each node extracts from the network a benefit which depends on the values of the nodes which are *connected* to it. That is, denoting as $N_i(g)$ the set of nodes j such that the network g contains a path from i to j, the benefit extracted by i is defined as:

$$B_i(g) = B_i(\bar{g}) = \sum_{j \in N_i(\bar{g})} v_{ij};$$

i also pays a cost which is the sum of costs of sponsored (outgoing) links:

$$C_i(g) = \sum_{j:g_{ij}=1} c_{ij},$$

and the resulting payoff deriving from the network is simply the difference between the benefit and the cost:

$$\Pi_i(g) = B_i(g) - C_i(g).$$

Some other standard graph-theoretic concepts and notations will be used. $\mathcal{G}$ denotes the space of the $2^{n(n-1)}$ possible networks, and e the empty network. A set of nodes $S \subset N$ is said to be *connected* if for any $i, j \in S$ there exists a path from i to j in $\bar{g}$, and is said to be a *component* if moreover for any $i \in S$, $j \notin S$ there is no path from i to j in $\bar{g}$; a link is said to be a *bridge* if the number of components of the network changes (increasing by 1) when it is removed; a network is said to be *minimal* if all links are bridges, and *minimally connected* if it is connected and minimal. Moreover, the notation

$$g_i = (g_{i,1}, \ldots, g_{i,n}) \in \{0,1\}^n$$

summarizes the outgoing links from a given node i in the network g (in the present work, it is always assumed that $g_{ii} = 0$). A *strategy* for a node is also an element of the set $\{0,1\}^n$, and I will say that a strategy is *included* in another one (and write $g_i \subset g_i'$) if it involves sponsoring only links which are sponsored according to the other one (that is, if $g_{ij} = 0$ whenever $g_{ij}' = 0$).

1.2.1 Internal constraints

Haller (2012) enriches this basic model with the presence of *constraints*: in his work, a pre-existing and exogenously given network $\mathfrak{g} \in G$. The payoff function is modified in order to set the cost of links in $\mathfrak{g}$ to 0, and this implies that such links are always incentive compatible. The aim of the present section is to generalize the seminal idea of Haller with the concept of *negative* constraints: a model of network formation will be characterized not only by $\mathfrak{g}$, which will be denoted henceforth as $\mathfrak{g}^+$, but also by another network $\mathfrak{g}^-$, containing links which will be *absent* in any possible network (by assumption, $\mathfrak{g}^+$ and $\mathfrak{g}^-$ will be disjoint). Although it is possible to introduce this generalization by setting the cost of links in $\mathfrak{g}^-$ high enough, a more tractable approach is to neglect their benefits in the payoff function,[3] which is hence defined as

[3] As in the approach of Haller (2012), the original cost of links in $\mathfrak{g}^+$ and $\mathfrak{g}^-$ should be taken again into consideration when doing comparative statics and welfare analysis.

$$\Pi_i\left(\mathfrak{g}^+, \mathfrak{g}^-, g\right) = B_i\left(g \oplus \mathfrak{g}^+ \ominus \mathfrak{g}^-\right) - C_i\left(g\right) \qquad \text{for } g \in \mathcal{G}.$$

where $\oplus$ and $\ominus$ denote respectively the operations of union and difference between networks.[4] It can be easily verified that when $\mathfrak{g}^- = e$, this coincides with the payoff function defined by Haller (2012). With all the components of the model exposed, we can proceed to the generalization of some of his results concerning Nash networks - that is, networks which are stable with respect to individual deviations.

Proposition 1. *Consider a strategic model of network formation with payoff functions $\Pi_i(\mathfrak{g}^+, \mathfrak{g}^-, g)$, $g \in \mathcal{G}$, $i \in N$. Suppose that costs are owner-homogeneous. Then there exists a Nash network g^*.*

This proposition is an immediate generalization of Proposition 1 by Haller. The assumption that costs are owner-homogeneous is one of the reasons why it is impractical to define negative constraints just as prohibitively costly links: if this was the case, in order for a owner-homogeneous model of network formation to remain such after the imposition of negative constraints, such constraints could not consist in arbitrary sets of links, but rather include all outgoing links from a given set of nodes.[5] In order to prove Proposition 1, let us first introduce the following definition.

Definition 1. *$g \in \mathcal{G}$ is a* Nash extension *for $\Pi(\mathfrak{g}^+, \mathfrak{g}^-, \cdot)$ if it is disjoint from $\mathfrak{g}^+$ and $g \oplus \mathfrak{g}^+$ is a Nash network.*

That is, a Nash extension is the set of non-exogenously given links contained in a given Nash network. We can then state the following simple result.

Lemma 1. *If $\vec{g}$ is a Nash extension for $\Pi(\mathfrak{g}^+, \mathfrak{g}^-, \cdot)$, all of its links are bridges for $\vec{g} \oplus \mathfrak{g}^+$.*

Proof. Let $i, j \in N$ be such that (i, j) is in $\vec{g}$ and it is not a bridge for $\vec{g} \oplus \mathfrak{g}^+$. By definition, $(i, j) \notin \mathfrak{g}^+$. Then, by removing this link, the connected components (of $\vec{g}_{-i,j} \oplus \mathfrak{g}^+$) remain the same as those of $\vec{g} \oplus \mathfrak{g}^+$. Hence, the benefit of node i, which is only determined by the extent of its component, is unchanged:

$$B_i(\vec{g} \oplus \mathfrak{g}^+) = B_i(\vec{g}_{-i,j} \oplus \mathfrak{g}^+),$$

[4] With a slight abuse of notation, when the network to be added/removed is composed of a single link, I will write $g \oplus (i, j)$ or $g \ominus (i, j)$, rather than $g \oplus \{(i, j)\}$ or $g \ominus \{(i, j)\}$.

[5] Another reasons is that this would make the definition of *endogenous* negative constraints, as described in Section 1.2.2, much more complicated.

while the total costs for the node have decreased:

$$C_i(\vec{g} \oplus \mathfrak{g}^+) > C_i(\vec{g} \oplus \mathfrak{g}^+) - c_{i,j} = C_i(\vec{g}_{-i,j} \oplus \mathfrak{g}^+).$$

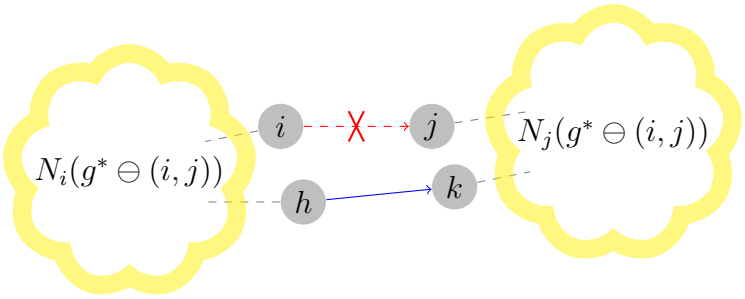
So finally the payoff for i has increased: $\mathfrak{g}^+ \oplus \vec{g}_{-i,j} \succ_i \mathfrak{g}^+ \oplus \vec{g}$. This implies that $\mathfrak{g}^+ \oplus \vec{g}$ was not a Nash equilibrium, and hence $\vec{g}$ is not a Nash extension for $\Pi(\mathfrak{g}^+, \mathfrak{g}^-, \cdot)$. □

Proof of Proposition 1. Consider any $\mathfrak{g}^+$. I will show that, given any g^{m-} composed by m links, if there exists a Nash network g^* for $\Pi(\mathfrak{g}^+, g^{m-1-}, \cdot)$ with $g^{m-1-} \subset g^{m-}$ of numerosity $m-1$, then one also exists for $\Pi(\mathfrak{g}^+, g^{m-}, \cdot)$. g^{m-} is $g^{m-1-} \oplus (i,j)$, for some $(i,j) \notin g^{m-1-} \oplus \mathfrak{g}^+$. If $(i,j) \notin g^*$, then g^* itself is a Nash network for g^{m-} (i's strategies set having being restricted, and all of the others nodes' ones staying unchanged, the equilibrium is still such), and hence this step is trivial. So let us assume that $(i,j) \in g^*$. The link (i,j) is contained in the Nash extension $g^* \ominus \mathfrak{g}^+$ (since it is by assumption not in $\mathfrak{g}^+$), so Lemma 1 guarantees that it is a bridge in g^*. Two things may happen:

A) there exist h, k in the union of $N_i(g^* \ominus (i,j))$ and $N_j(g^* \ominus (i,j))$ such that $h \notin N_k(g^* \ominus (i,j))$ (see Figure 1.1), $(h,k) \notin g^{m-}$, and

$$c_{h,k} < \sum_{k' \in N_k(g^* \ominus (i,j))} v_{h,k'}. \tag{1.1}$$

Figure 1.1: Link $[h,k]$ replaces link $[i,j]$.



B) There are no such h, k.

Let us consider case A, and denote as g^{*A} the network $g^* \ominus (i,j) \oplus (h,k)$. For any node $l \in N \setminus \{i,h\}$, the strategies set is unchanged from g^* to g^{*A}, as well as the payoffs. For i, all *available* strategies now deliver a payoff increased by $c_{i,j}$ (since the cost of connecting the two components is now borne by h),

so their preference ordering does not change. For what concerns h, since costs are owner-homogeneous, she does not strictly prefer to replace the link (h,k) with a link to any other $k' \in N_k(g^* \ominus (i,j))$. Apart from that, all her available strategies now deliver a payoff decreased by $c_{h,k}$, except the ones where $g_{h,k} = 0$, which however are strictly dominated because of (1.1).[6] So g^{*A} is a Nash equilibrium.

Consider case B instead, and denote as g^{*B} the network $g^* \ominus (i,j)$. For any node $l \neq i$, the strategies set is unchanged from g^* to g^{*B}, while the payoffs are decreased for all nodes in $N_i(g^* \ominus (i,j))$ and $N_j(g^* \ominus (i,j))$, but the preference ordering over available strategies does not change (except for those connecting the two components, which are however dominated). For what concerns i, in virtue of (1.1) and of the assumption of owner-homogeneity of costs, we know that (i,j') is in g^{m-} for each $j' \in N_j(g^* \ominus (i,j))$. The preference ordering of available strategies then clearly corresponds to the preference ordering of the same strategies with the addition of link (i,j), and in particular $g^{*B} = g_i^* \ominus (i,j)$ is optimal. So g^{*B} is a Nash equilibrium.

Whatever the case, the existence of a Nash network for $\Pi(\mathfrak{g}^+, g^{m-}, \cdot)$ was proved, by assuming that one exists for $\Pi(\mathfrak{g}^+, g^{m-1-}, \cdot)$. The case $\Pi(\mathfrak{g}^+, g^{0-}, \cdot)$, that is $\Pi(\mathfrak{g}^+, e, \cdot)$, is Proposition 1 by Haller (2012). The result is hence proved for any possible $\mathfrak{g}^-$ by induction.

□

What follows is instead the natural generalization to negative restrictions of Haller's Proposition 2.

Proposition 2. *Consider a strategic model of network formation with payoff functions $\Pi_i(\mathfrak{g}^+, \mathfrak{g}^-, g), g \in \mathcal{G}, i \in N$. Suppose that the pre-existing network or infrastructure $\mathfrak{g}^+ \in \mathcal{G}$ is Pareto optimal. Then the empty network is a strict Nash network and the only Nash network.*

Proof. Let $\mathfrak{g}^+$ be Pareto optimal. The case $\mathfrak{g}^- = e$ is proved by Haller (2012). When considering $\mathfrak{g}^- \neq e$, the actions set of some nodes is restricted, but the links in $\mathfrak{g}^+$ are left untouched (recall that $\mathfrak{g}^+$ and $\mathfrak{g}^-$ are disjoint). Hence, the empty network is still a strict Nash network, because the preference ordering on available strategies does not change.

Suppose next that some $g^* \neq e$ is a Nash network. The proof develops as in the original result: given some player i with $g_i^* \neq 0$, it must be that g_i^* is a best response against g_{-i}^*. But then g^* is strictly preferred to $\mathfrak{g}^+$ by at least i, while it is at least equally preferred by all other agents. This contradicts the Pareto optimality of $\mathfrak{g}^+$. □

[6] This line of reasoning holds with minimal changes also in the case in which $i = h$.

Propositions 1 and 2 are pure generalizations of the theory of network extension to the presence of negative constraints. The effort in this direction can be motivated with two main arguments:

1. considering negative constraints is important to understand the growth of some real world network settings,

2. from a social planner point of view, imposing negative constraints can improve the beneficial effects of an endogenously formed network, possibly at a lower cost than through positive constraints.

The first motivation has already been discussed in Section 1.1. I will now devote some attention to the second. Haller (2012) shows several ways in which positive exogenous constraints can impact on the equilibria of a network: examples include a *stabilizing* effect (some models of network formation exhibit non-existence of Nash network, which can instead exist for given $\mathfrak{g}^+$), a *welfare improvement* effect (constraints can raise the overall sum of payoffs in Nash equilibrium), and others. Those exogenous constraints can hence be imagined as publicly provided infrastructures which are undertaken by the social planner. Can some of the described effects be attained as well through *negative* constraints - i.e. with a social planner acting through *prohibition* of a set of given links? For what concerns the attainment of *efficient* Nash equilibria, negative constraints cannot simply replace positive ones. See for instance Example 3 by Haller (2012):[7] it is clear that in the absence of *positive* constraints, no link will ever connect the two sets of nodes $\{1,2,3,4\}$ and $\{5,6,7,8\}$ (and hence the welfare improvement represented by such a link will be lost), since for any i and k in the two different sets, we have

$$\sum_{j \neq i} v_{ij} = 8 < 16 = c_{ik}.$$

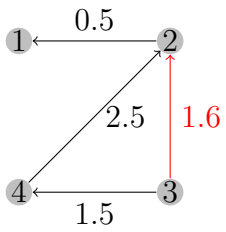
Something more can be said instead about the stabilizing effect. In fact, an example of it is easily found: by simply setting $\mathfrak{g}^-$ as the complementary of

[7] For ease of reading, the example is reported here:
Let $N = \{1,2,3,4\}$, $K = \{1,2,3,4\}$, $L = \{5,6,7,8\}$. *Further set* V_{ij} *for all* $i \neq j$, *and*

$$c_{ij} = \begin{cases} 0.8 & \text{if } i \neq j,\ i,j \in K, \\ 0.8 & \text{if } i \neq j,\ i,j \in L, \\ 16 & \text{if } i \in K, j \in L, \\ 16 & \text{if } i \in L, j \in K. \end{cases}$$

It is easy to see that no link between components K and L will be privately provided, although it would be welfare improving.

Figure 1.2: An example of the stabilizing effect of negative constraints.



$\mathfrak{g}^+$ (hence e if we are assuming $g^+ = 0$), we transform g^+ itself into a trivial Nash equilibrium - analogously, the trivial but uninteresting way to stabilize any model of network formation with only positive constraints is clearly to set $\mathfrak{g}^+$ to the complete network. A more interesting case comes from Example 2 of Haller et al. (2007)[8], represented in Figure 1.2, where $v_{i,j} = 1$ for any i, j (all edges absent from the picture have a cost higher than 3 and are hence irrelevant). Haller (2012) shows that by setting $g^+ = \{[4, 2]\}$, we obtain the Nash equilibrium $g^* = \{[3, 4], [4, 2], [2, 1]\}$; the same result can be obtained by setting $g^- = \{[3, 2]\}$.

1.2.2 Repeated internal constraints, and non-decreasing network models

A crucial ingredient of any real world example of endogenous network formation is *time*. A study of the consequences of repeated internal constraints, going beyond the analysis of static Nash equilibria relative to exogenous constraints, is hence a natural development of the theory exposed so far. This will be exemplified in the empirical application, concerning bibliometric networks, proposed in sections from 1.3 to 1.6 of the present paper. In what follows, I will assume that the formation of the network happens in a discrete time setting. For each $t = 1, 2 \ldots$, I will define as $\mathfrak{g}^{t^+}$ and $\mathfrak{g}^{t^-}$ respectively the positive and negative constraints at that period. At each time, each node's best reply is the one maximizing $\Pi_i(\mathfrak{g}^{t^+}; \mathfrak{g}^{t^-}; \cdot)$.[9] The outcome, if any, of the step t, denoted as g^t, will hence be a Nash equilibrium for these payoffs functions. Clearly, such outcome needs not be unique, neither to exist: if it does not, the network formation process *terminates* at time t.

The result which follows considers the specific class of *non-decreasing network models*, defined as those for which $\mathfrak{g}^{t^+} = g^{t-1}$ (the positive restriction

[8] Example 1 by Haller (2012).

[9] Clearly, the framework is also an ideal context for the study of a less myopic type of rationality, such as the farsightedly stable networks (Herings et al., 2009).

coincides with the outcome of the previous step of the process): such class naturally maps to several real world contexts, including the case of bibliometric networks. A peculiarity of non-decreasing network models is that, since the number of links present at time t is (weakly) increasing in t itself, and since it can never exceed $n^2 - n$, it must, for some t, terminate *or* stabilize in some configuration, which I will call a *limit network*. A limit network will then be defined as *strict* if there is no other limit network composed by a subset *or* a superset of its links.

Proposition 3. *If* $\mathfrak{g}^{t^-} = \mathfrak{g}^{t-1^-}$ *for every* t*, then the set of (strict) limits of the non-decreasing network model corresponds to the set of (strict) Nash equilibria of the static model associated to the payoffs function* $\Pi_i(\mathfrak{g}^{1^+}; \mathfrak{g}^{1^-}; \cdot)$*.*

Proof. Consider a (strict) Nash equilibrium g^* of the model associated to payoffs functions $\Pi_i(\mathfrak{g}^{1^+}; \mathfrak{g}^{1^-}; \cdot)$. By, definition, it is also a (strict) Nash equilibrium for the first step of the non-decreasing network model. In order to prove that it is a limit network, it is hence sufficient to show that it is still a (strict) Nash equilibrium for $\Pi_i(g^*; \mathfrak{g}^{1^-}; \cdot)$. Assume it is not: that is, there is some i who (weakly) prefers some $g_i' \supset g_i^*$. But then, g^* was not a (strict) Nash network in the first place. The same applies hence for $t = 2, 3 \ldots$

Now assume g^* is a (strict) limit for the non-decreasing network model, reached at some time t^*. By construction, g^* is a union of t^* subsequent Nash extensions (some of them possibly empty), so because of Lemma 1 we know that all of its elements are bridges. Given any time t and any link (i, j) in $g^t \ominus \mathfrak{g}^{t^+}$, let $\Delta_{i,j}^t$ be the profit which the link (i, j) yields to i in g^t, that is,

$$\Delta_{i,j}^t = \Pi(\mathfrak{g}^{1^+}, \mathfrak{g}^{1^-}, g^t) - \Pi(\mathfrak{g}^{1^+}, \mathfrak{g}^{1^-}, g^t \ominus (i, j)).$$

This profit is necessarily (strictly) positive, since the node is part of the best reply of i. Any $\Delta_{i,j}^{t'}$ with $t' > t$ will also be positive - all new links are bridges, and so the connected component of j can only grow, while no paths from i to j alternative to (i, j) can appear. So no node has a (strictly) positive individual incentive to simply break one or more existing links in g^t. If g^* is not a Nash equilibrium of $\Pi_i(\mathfrak{g}^{1^+}; \mathfrak{g}^{1^-}; \cdot)$, this means that some node has individual incentives to *add* some link, or to *replace* some link with some other. The first case is impossible: since $\mathfrak{g}^{t^-}$ is constant, this would make g^* unstable also at time t^*. But the second is also impossible: since $\Pi_i(\mathfrak{g}^{1^+}; \mathfrak{g}^{1^-}; \cdot)$ is positive, the new link should still connect i to $N_j(g^* \ominus (i, j))$. So to be incentive compatible, it should cost less than (i, j). But then, it would have been chosen at time t in its place. □

Proposition 3 in particular implies that when no Nash equilibrium exists, no limit network exists, and the iterative process necessarily terminates at

time 1. Another interesting implication is that network models satisfying only the more general condition $\mathfrak{g}^{t+} \supseteq g^{t-1}$ do not exhibit richer limit structures than non-decreasing network models: imposing $(i,j) \in \mathfrak{g}^{t+}$ would not make a change, in terms of limit networks, compared to imposing $(i,j) \in \mathfrak{g}^{0+}$. Richer dynamics could instead be expected when

1. considering *partially* non-decreasing network - networks in which only *some* previously provided links can be destroyed, at *some* instants in time, or

2. introducing time-dependent *negative* constraints.

The empirical application described in the remaining of the paper falls in this second case.

1.3 The growth of the citations network

The network of citations between scientific papers is an eminent example of an endogenously formed network in which the time component is not just crucial for the endogenous growth mechanism, but also easily observable in the available data. Indeed, papers have a well defined publication date, which imposes a clear temporal hierarchy among them and hence strong restrictions to the set of "actions" - that is, of citations - they can make.

Before diving in the details of the time-related properties of the network of citations among scientific publications, it is worth reviewing why, from a static point of view, the non-cooperative approach *à la* Bala and Goyal (2000), introduced in Section 1.2, is appropriate for the setting under analysis.

- A citation is a purely one-sided sponsored kind of relation: an author can very well find out *ex-post*, if ever, that she's been cited.

- A citation represents a two-ways channel for benefits. The fact that being cited can, at least in some cases, represent a gain for a researcher is unanimously recognized, and is part of the motivation for the present study. Apparently less intuitive would be the utility obtained from *making* a citation. However, the hypothesis that *there is* some benefit is supported by the evidence that the overwhelming majority of scientific articles do have a list of bibliographic references.[10] See Section 1.1 for

[10]The kinds of benefit flowing in each way may still seem very different, and for instance it would be very hard to compare them in terms of importance. This is however not needed, since a paper *cannot* in any way create ingoing links, and hence its "rational choice" is based only on the values added and costs of *outgoing* links.

a brief survey of the literature on citation behavior.

- Though apparently there is no cost involved in "sponsoring" a citation, it is evident that the number of bibliographic references contained into a single scientific work is limited: many authors, starting with de Solla Price (1965), have analyzed different aspects of its distribution, evidencing a strong concentration for small values (as will be confirmed in Section 1.5). While this evidence does not help in disaggregating the implicit costs born by authors in making citations, which may be due partly to editorial/formatting choices and partly to the work involved in processing the literature to be cited, it does provide striking evidence that some costs are indeed hidden in the process.

- As best exemplified by the phenomenon of *literature reviews*, it is very natural to assume that the benefit of a citation to a given paper depends in turn also on the citations *from* that paper. The hypothesis of *perfectly reliable links* - meaning that being connected to another paper through an arbitrarily long path is equivalent to being directly connected - is instead a non-harmful approximation of reality for the present analysis. As will be motivated in Section 1.4.1, it does not affect its qualitative results, and on the other hand an alternative specification would make the analysis much more complex and require some arbitrary choices.

Having clarified those points, the growth of the citations network is clearly a particular case of the model described in the previous section: citations cannot be removed, so the network model is clearly non-decreasing. At the same time, there are obvious restrictions to the set of available strategies.

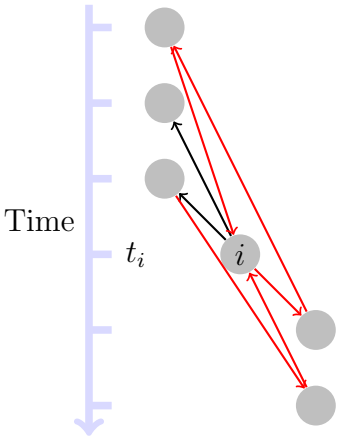
It is clear that the choice of adopting a pure value-based model of citation flows implies the exclusion of a wide range of factors that can possibly affect them, and some of which have been already mentioned in Section 1.1. The aim of this empirical application is not to convince the reader that such factors are irrelevant: well on the contrary, as will be clear in Section 1.6, the model is instrumental in showing that they have a significant effect on the citation flows. In particular, the model completely negates the constructive view on the citation behavior, since the strategic behavior of agents (papers) is only directed towards an increase in the quality of the publication: this is clearly an unrealistic assumption. Moreover, many empirical works have focused on the fundamental role of individual researchers and scientific fields as "attractors" of citations: the choice of neglecting such aspects in the model is, again, consciously made in order to *reject* a model according to which bibliometric indicators would be perfect measures of research quality.

For simplicity of exposition, I will assume that the set of nodes of the network of citations is predetermined (i.e. contains all scientific papers which are going to be published in a given time span). The number of instants in time is equal to this number of nodes, so that there is a one-to-one relation between each node i and its instant t_i (and vice-versa, between an instant t and its node n_t). The negative restrictions are then defined as follows:

$$\mathfrak{g}^{t_i^-} = \{(j,k) : j \neq i \text{ or } t_k > t_i\}$$

which means that at each instant in time, only the scientific publication being published can establish links, and it cannot cite works which are yet to be published (Figure 1.3).[11]

Figure 1.3: Examples of allowed and forbidden links at time t_i.



Allowed links in black, forbidden links in red. Links arrows point *from* the citing *to* the cited paper.

In Section 1.2.2, the fundamental building block of the development of a network with repeated internal constraints was assumed to be the Nash equilibrium of a given step t. Under the specification given for the network of citations, in which at each step only one node is able to choose among different strategies, such a Nash equilibrium degenerates to the simple best response of such node. It is important to notice that the hypothesis of all

[11] In principle, given the typical publication process, which goes through a period of open discussion in seminars/workshop, an often lengthy referral process, and finally a delay from the definitive acceptance to the publication, it can happen that two papers i and j cite some version of each other. This very special case, which is not admissible under the simplified settings just described, is relatively non influential in the global picture, but would possibly deserve a specific analysis.

nodes existing since time 0 does not distort in any way the strategic choice, which is determined simply as a best response *among allowed links*, which means that *later* links are irrelevant.

The values $v_{i,j}$ and costs $c_{i,j}$ will be assumed to be coming from two probability distributions $\mathcal{V}(v,i,j)$ and $\mathcal{C}(v,i,j)$. In order to formulate formally the hypothesis that citation flows are a valid proxy for the *value* of a paper, it is crucial to define what we mean by "*value*": even letting aside measurement issues, while for a public research institution the value of a paper may consist, for instance, in the efficacy of a new pharmaceutical it presents, for a journal it may more be its mere *impact* on the scientific community. Taking this difference at the extreme, if a paper is able to gather a high attention, but is based on fabricated evidence, the overall gain for the journal having published it, in terms of additional publicity and possibly subscribers, may still be large, while the overall gains for a public institution financing the research will almost certainly be negative.[12]

In the present analysis, the only restrictions imposed on the concept of *value* and on the citing behavior of scientists are summarized in two assumptions on $v_{i,j}$ and $c_{i,j}$.

> **Assumption 1**: each paper i is characterized by an *intrinsic value* $\bar{v}_i$ such that, for any $t_j > t_i$,
>
> $$\mathbb{E}[v_{j,i}] = f_v(\bar{v}_i, t_j - t_i).$$
>
> where f_v is an age effect.

> **Assumption 2**: the aggregate dynamics of $\mathcal{C}(v,i,j)$ are such that
>
> $$\mathbb{E}[c_{j,i}] = f_c(v_{j,i}, t_j - t_i)$$
>
> where f_c is an age effect.

Notice that no assumption needs to be made on the shape of f_v and f_c: in particular, they can be non-monotonic (the typical citation flow accruing to a paper *is* indeed non-monotonic, peaking after a few months - see Figure 1.7). More importantly, notice that the model does not enforce any interpretation of such "value": it could be *originality*, *adherence to scientific standards*, or

[12] Although the example is indeed extreme, a recent work by Fang and Casadevall (2011) finds a strong correlation between the "*retractation index*" of journals, and their Impact Factor. Many explanations are possible, but this reminds at the very least that journals may have scarce incentives to ascertain the methodological correctness of studies they accept.

a combination of those. The only fundamental issue is that it's a property related to the *node* (content of the paper) only.

Although no further assumption is made on $\mathcal{C}$ and $\mathcal{V}$, it can be worthwhile to consider some of their *plausible* aggregate features, which are compatible with assumptions 1 and 2. It is reasonable to expect

$$\mathrm{Cov}[v_{i,j}, |t_i - t_j|] < 0$$

(the value of recent papers for new publications is higher) and

$$\mathrm{Cov}[|v_{i,j} - v_{i',j}|, |t_i - t_{i'}|] < 0$$

(the value of papers which have a higher value for new publications is higher).

1.4 Hypotheses

With those concepts clarified, the null hypothesis which we want to test with the present econometric exercise is the following.

> **Hypothesis H0**: the growth of the citation network can be modeled as a non-decreasing model based on the value/cost based approach *à la* Bala and Goyal (2000).

Notice that this assumption *does not resort* to stating that the probability of any link from j to i, with $t_j > t_i$, is only a function of $\bar{v}_i$: it can instead also depend on the links already present in the network, and namely on the paths *starting* from i. In other terms, the value of a citation can also depend on the bibliographic references of the cited paper.

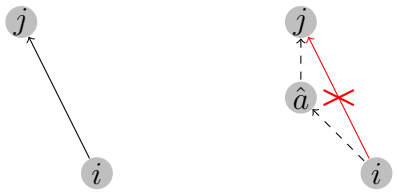
Proposition 4. *Let $\mathcal{M}$ and $\mathcal{M}'$ be two network models, with identical values $v_{i,j} = v'_{i,j}$ and costs $c_{i,j} = c'_{i,j}$, except for the presence of an additional node $\hat{a}$ in $\mathcal{M}'$. Given i, j with $t_i > t_{\hat{a}} > t_j$ in $\mathcal{M}'$, if assumptions 1 and 2 hold then*

$$\mathbb{P}_{\mathcal{M}}\{g^t_{i,j}\} \leq \mathbb{P}_{\mathcal{M}'}\{g^t_{i,j}\}. \tag{1.2}$$

that is, paper i will cite j with a lower *probability in presence of $\hat{a}$.*

Proof. Assume that $v_{i,j}$ is high enough as to make a direct link to j optimal for i in $\mathcal{M}$. In $\mathcal{M}'$, it may be that $\hat{a}$ is already connected (possibly indirectly) to j, and that it is convenient for i to connect (possibly indirectly) to $\hat{a}$ rather than directly to j (see Figure 1.4). On the other hand, if in $\mathcal{M}'$ it is convenient for i to directly link to j, then the same is necessarily true in $\mathcal{M}$. □

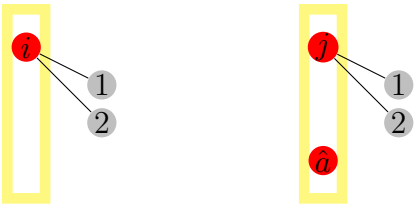
Figure 1.4: Comparison of $\mathcal{M}$ (left) and $\mathcal{M}'$ (right).



The rest of the paper is devoted to testing Hypothesis H0 on real data. In particular, I try to ascertain if the citation decision can be seen as "context-free", in the sense that it only depends on the (subjective) *value* $v_{j,i}$ of some node i, and possibly on i's "reference network" (paths starting from i). The particular *environmental characteristic* I will put on trial is the scientific activity of the author of i: does the subsequent publication of a new paper increase the flow of ingoing links for i?

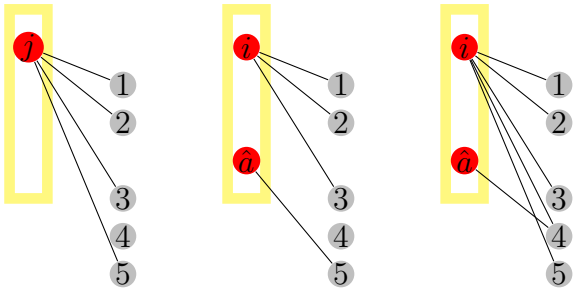
Ideally, to test such an issue one would compare the real network of citations (corresponding to the $\mathcal{M}'$ of Proposition 4) and the same network without a given publication $\hat{a}$ ($\mathcal{M}$). More specifically, the comparison would regard the citation flows of works of the same author of i. A feasible alternative is possible: notice that the networks emerging from $\mathcal{M}$ and $\mathcal{M}'$ are perfectly identical *until* $\hat{a}$. Hence, we can compare the citations inflow of *different* papers, as follows. Consider two scientific publications i, j of two different authors, with comparable observable characteristics, including year of publication and citations flow in the first years after publication, but such that i's author (and *not* j's author) publishes another work after x years, as shown in Figure 1.5. According to Proposition 4, we would expect the number of citations for i *after* those x years to raise less than those for j, as shown in the center of Figure 1.6.

Figure 1.5: Ex-ante comparable papers (same year of publication, same initial flow of citations)



The yellow rectangle regroups works by a same author.

Figure 1.6: Effect of an additional paper $\hat{a}$ by a same author



The new publication can cause citations to previous papers to decrease (center) rather than increase (right).

If instead the right pattern is found to be significantly more common than the central one, we have a clear sign that the choice of linking to a node is affected by factors which are not considered by the model. The above reasoning can be translated into the following regression:

$$cit_{ij}^{T} = \beta_0 + \beta_1 pub_j^{T-1} + \beta_2 cit_{ij}^{\rightarrow T-1} + u_{ij} \tag{1.3}$$

where $cit_{ij}^{\rightarrow T-1}$ is the indegree (cumulated number of citations) of the ith paper of author j at time $T-1$, cit_{ij}^{T} is one if and only if such paper is cited by the paper published at time T, and pub_j^{T-1} is a dummy variable taking value 1 if and only if author j published another paper at time $T-1$. While in the model presented so far there is a one-to-one match between the periods of time and the published papers, so that cit_{ij}^{T} is indeed a dummy variable, the empirical analysis will have to be adapted to the temporal resolution of available data, and so cit_{ij}^{T} will be a discrete variable *counting* the number of citations to the ith paper of author j at time T.

If the aforementioned "success breeds success" phenomenon only had a descriptive interpretation in terms of correlation, we would see $\beta_2 > 0$ with still $\beta_1 \leq 0$: the existing citations are the best possible predictor for the future citational success of a paper, while the publication of additional papers has a negative effect, if any. Hence Proposition 4, in the context of Equation (1.3), can be translated into the following testable hypothesis:

Hypothesis H1: $\mathcal{H}_0 : \beta_1 \leq 0$ vs $\mathcal{H}_1 : \beta_1 > 0$.

Moreover, if recent publications of a same author do eclipse, at least in part, previous ones, we should expect the effect to be even stronger if the new paper benefits from a larger circulation. To test for this additional assumption, the following alternative formulation of (1.3) could be considered:

$$cit_{ij}^{T} = \beta_0 + \beta_1 pub_j^{T-1} + \beta_2 c_{ij}^{\rightarrow T} + \beta_3 infl_j^{T-1} + u_{ij} \tag{1.4}$$

where $infl_j^{T-1}$ is a proxy for the influence of the journals on which author j published her papers (if any) at time $T-1$, such as a function of their Impact Factor at the time of publication. The prediction would then be the following:

Hypothesis H2: $\mathcal{H}_0 : \beta_3 \leq 0$ vs $\mathcal{H}_1 : \beta_3 > 0$.

1.4.1 On perfectly reliable links

The hypothesis of perfectly reliable links, intrinsic in the basic model of Galeotti et al. (2006) and hence in the approach to network extensions by Haller (2012), which the present study builds upon, is at odds with both the intuition and the evidence coming from the citations network (for instance because, under such hypothesis, networks formed by rational nodes would never contain cycles). However, this assumption, which is widespread in the literature on social networks, is the only alternative to more complicated and arbitrary formalizations of the *dispersion* of value along the paths, such as imposing a maximum path length through which value is able to flow. The aim of this paragraph is not to convince the reader that perfectly reliable links are a realistic approximation of the flow of information over bibliometric networks, but rather to show that hypothesis H1 is valid even if an imperfect flow of information through the links is assumed.

As already mentioned, the fundamental assumption of the present approach is that citations toward some node i only depend on the private values it has for the subsequent nodes, and on its *reference network* - the subnetwork containing paths starting i, and composed only of papers published *before* i. Now, recall the proof of Proposition 4. Under imperfect information flow, the value of being directly linked to j, both because of its private value and because of the value of its reference network, is unchanged between $\mathcal{M}$ and $\mathcal{M}'$, while the value of being *indirectly* connected to j may decrease. Hence, inequality (1.2) could only be *less strict*: pushing the imperfection to its limit, if all paths of length higher than 1 contributed *no value* at all to nodes, it would become an equality.

1.5 Data

Citation databases have existed at least since the 1961 *Science Citation Index* produced by Eugene Garfield. Today, the most prominent ones for biblio-

metric studies are ISI Web of Knowledge (which is the evolution of Garfield's database) and Elsevier Scopus.[13] However, for the present exercise the possibility to analyze the *whole* network - or some large subset of it - is crucial, and hence it will be developed on a specialized database, provided by the American Physics Society to researchers in an aggregate form. The APS database (available at `http://publish.aps.org/datasets`) contains metadata for all articles published over the years on any of the 11 highly popular journals of the society, as well as a record of the citations among them, making it possible to reconstruct a relevant bibliographic network: overall, it covers 464.817 articles and 4.710.547 of citations among them. From the metadata, provided in XML format, it is possible to extract, for each article, the names of the authors and the date and journal on which it was published. Citations are instead reported in a single file containing two *doi* identifiers per line: the first corresponds to the paper which contains the citation, the second to the cited paper. Given its quality and availability, the APS database has been the subject of many studies on the social aspects of scientific research (see for instance Radicchi et al., 2009 and Adamic et al., 2010; the website at `http://www.physauthorsrank.org` provides a tool entirely based on such dataset). Although the network it represents is truncated, in the sense that it misses citations to and from papers published on other journals, the journals of the APS are not simply among the most important (although clearly a small minority) in the realm of physics: they also represent a sort of *ecosystem* (each paper has an average of 10 citations *to other papers of the sample*) inside which it is possible, as will be shown in Section 1.6, to detect social effects.

In light of the availability of a large amount of high quality data, and of the possibility on the other hand that the characteristics of interest may be changing with time, it is possible, and convenient, to decompose the analysis in intervals of time. While the definition of *a period* will simply be "a month", the analysis itself will focus on data from a single year, in order to have an optimal tradeoff between temporal localization and samples numerosity. For all papers published in 1990, Equation (1.3) will be estimated on the 120 months following the publication - that is, approximately until 2000. Sensitivity tests will then be ran for the following years, until 1999 - since the data ends in 2010, it would not be possible to analyze the ten years *following* the publication for works published in the new millennium.

[13] Google Scholar is getting much momentum, but compares substantially worse than the others two for what concerns the quality of filter to what is defined as "scientific publishing": this is due mainly to the automatized process which is behind it, which has been shown to be easily manipulable (an extreme case which received some publicity is described by Labbé, 2010).

Table 1.1: Descriptive statistics

	Observations	Total	Min	Median	Mean	Max
Citations	9121	89559	0	4	9.82	326
Citations (non-self)	9121	73255	0	3	8.03	308
Publication	9121	841207	0	21	92.23	30751
Cit./month	1094520	89559	0	0	0.08	11
Cit./month (non-self)	1094520	73255	0	0	0.07	11
Pub./month	1094520	841207	0	0	0.77	1173

Data from publications in the year 1990, observed over the 10 years after publication. The unit of observation is the individual work, "*Publications*" are subsequent works by same author(s). A citation is "non-self" if the set of authors of the citing paper and of the cited paper are disjoint.

Table 1.1 reports some descriptive statistics for the sample of interest. Notice that the number reported as "Total" of publications (in the year 1990) is almost double than the number of *all* scientific articles in the database: this is due to the fact that multiple authored papers are counted once for each author (this holds true also in the rest of the analysis). As can be observed, the distribution of citations per month is very uneven, with an average of 0.8 but a median of 0 and a maximum of 11: even more so the number of subsequent publications by coauthors. This can be better observed in figures 1.8 and 1.9, which show that not only citations (as already suggested by Redner, 1998), but also publications tend to be distributed roughly according to a *power law* distribution.[14]

The large amount of zeros in the data could raise the suspect of zero inflated data. This is however not plausible for two reasons. From the theoretical point of view, there is no obvious reason why some categories of papers should be *excluded* from the realm of "citable" papers: well on the contrary, there is widespread evidence that even inside a given journal there is typically a huge heterogeneity in the number of citations received by each article (see for instance Campbell, 2005). From the empirical point of view, Figure 1.8 shows that the curve describing the frequency of citations per month is remarkably smooth at least in the range from 0 (included) to 5, while a discontinuity would be expected instead in the case of zero inflated data.

In order to test Hypothesis H2, a proxy for the impact, or prestige, of a given publication is needed. Although it is possible, for recent years, to use

[14]The clear non-monotonicity which can be spotted between 100 and 1000 publications per month is presumably a sign of the very peculiar publication habits of large teams of experimental physicists, which will be considered again in Section 1.6.3.

Figure 1.7: Average flow of citations during the first 10 years after publication.

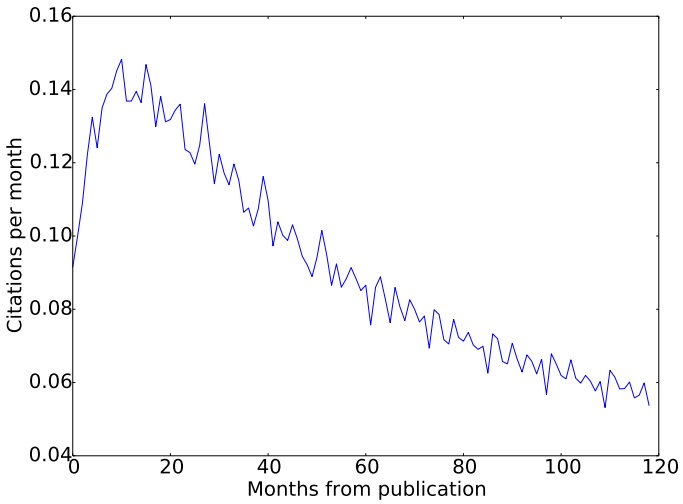


Figure 1.8: Frequency distribution of citations per month per paper.

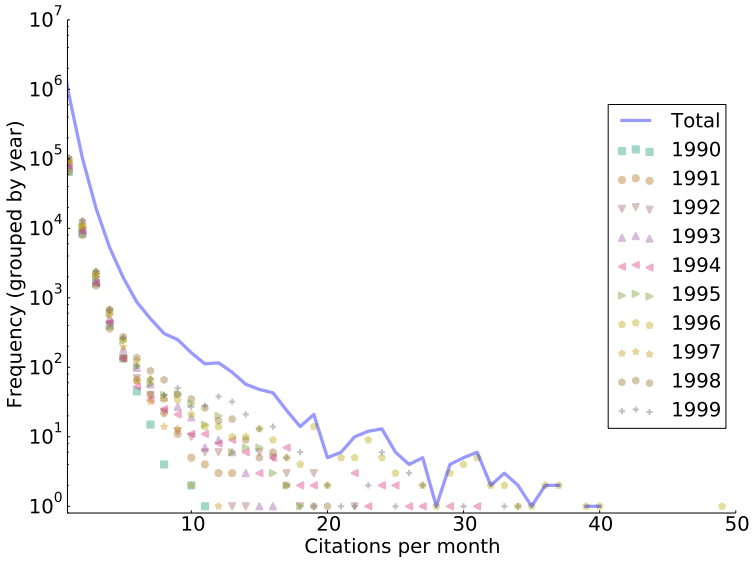
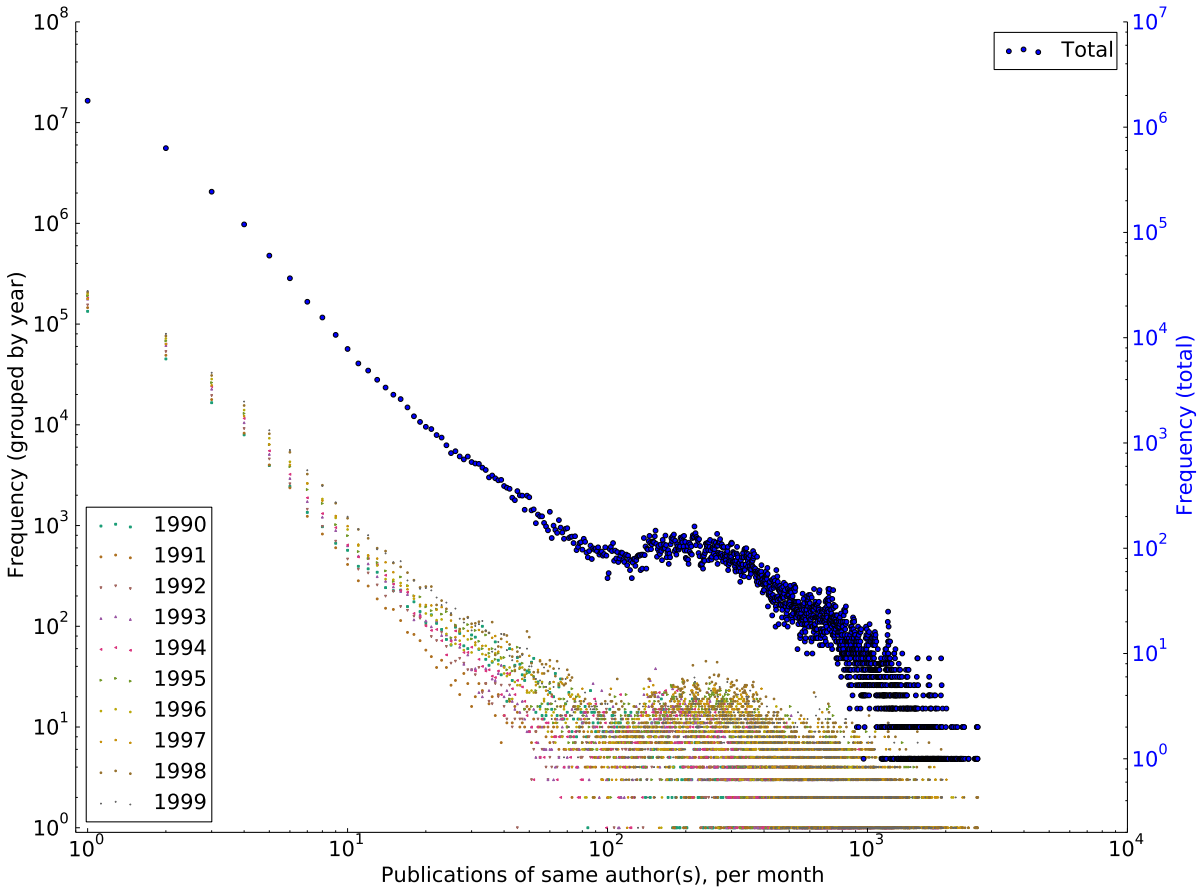


Figure 1.9: Frequency distribution of publications of an author of a given paper, in a given subsequent month.



the well known Impact Factor as a rough measure of the gain in visibility from publishing on a given journal, the characteristics of the database make it preferable to develop the analysis using only data coming from it, in order to avoid confounding errors coming from the truncatedness of the network.

In particular, the choice of relying only on APS data neutralizes the risk of the analysis being distorted by important citation flows coming from other disciplines (i.e. engineering, mathematics), which could have a different response to the publication of a new work in the realm of physics (i.e. because different theoretical works of a same author can have drastically different practical applications). From a more pragmatic point of view, the ISI Web of Knowledge only makes Impact Factor measurements available for years from 2002 onwards.

For this reason, an "*Internal Impact Factor*" is reconstructed; for each of the 11 journals appearing in the database, and for each year considered in the analysis, it is calculated analogously to the well known Impact Factor:

$$IIF_J^y = \frac{\#\text{ of citations to articles on journal } J \text{ in year } y, \text{ in the 2 y.s after}}{\#\text{ of articles on journal } J \text{ in year } y}$$

The influence of a publication will hence be proxied as the IIF for the journal on which it is published, in the year it is published. Notice that this does not introduce endogeneity issues, since the citations to a given publication do not enter in the calculations for the IIF at the date of publication.

1.6 Results

Table 1.2 shows the result of a two-way adaptation of Equation 1.3 ran on different subsets of the data, for papers published in the year 1990. For each paper p considered, and each month T in the 10 years after the paper's publication, the number of citations *per month* received by p *at time* T is regressed on the number of citations per month received until then, and on a dummy variable which takes value 1 if and only if one of the authors of the papers published a paper in the six months *before* T. Time effects are defined based on the number of months elapsed since the publication of a paper (rather than on the effective month of the year), in order to account for the possibility of intrinsic "age effects" affecting the life cycle of a paper. On top of that, the two-way approach involves paper-specific fixed effects, which are strongly recommended given the high level of heterogeneity among papers.

As expected, the *past* citations flow seems to be the most important predictor, since it proxies the impact of the publication. But what is most

Table 1.2: Main results

	Dependent variable: cit				
	all	avg(cit)<1	max(cit)<2	avg(pub)<1	max(pub)<2
Cit/month (pre)	1.328***	1.346***	1.489***	1.330***	1.291***
	(0.005)	(0.005)	(0.007)	(0.005)	(0.010)
Pub. in past 6 m.	0.080***	0.078***	0.051***	0.083***	0.103***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
R^2	0.078	0.073	0.069	0.076	0.063
Observations	1085399	1081710	778141	1000195	272153

Results of the two-way model estimation. While "all" refers to all papers published in 1990, the other samples are defined as containing those papers with average (respectively: maximum) number of citations per month lower than 1 (respectively: 2), and analogously in terms of publications per month of same author(s).

interesting for the current analysis is the coefficient for "Pub in the last 6 months": its interpretation is that, on average, if a new paper published, each previous paper of a same author gains between 0.05 and 0.1 citations per month in the following six months. While the statistical significance is evident, the economic relevance of its estimated value must be judged in light of Table 1.1, showing that the average number of citations per month is 0.08. In other words, on a random article published in 1990, this effect would consist at least temporarily in an *increase* between 63% and 125%. This allows us to state the following.

> **Result 1:** The null hypothesis of H1, $\beta_1 \leq 0$, is rejected: when new articles are published, citations flows to previous ones *increase*.

1.6.1 Interpreting the effect

The work of researchers is characterized by high fixed costs in terms of exploring new fields, and their literature: this often causes strong forms of specialization and an imperfect knowledge of the existing work done by other scholars. It is hence only natural that self-citations - that is, citations from a paper A to previous work of one or more authors of paper A - are a frequent phenomenon in any scientific discipline, even leaving aside the widespread forms of strategic behavior. Their quantitative presence is already evident in Table 1.1, where the "non-self" citational variables are defined net of self-citations: the difference is around 18%. For comparison, Aksnes (2003) reports a rate 26% for a sample of physics articles published between 1981

Table 1.3: Results of different specifications of Equation (1.3).

	Dependent variable:				
	cit (n.s.)	cit (n.s.)	cit	cit (n.s.)	cit
	all	max(pub)<2	all	all	ISI dataset
Cit/period (pre)			1.327***		0.623***
			(0.005)		(0.023)
Cit/per. (pre, non-self)	1.358***	1.309***		1.358***	
	(0.005)	(0.011)		(0.005)	
Pub. in past 6 periods	0.003***	0.004*	0.089***	−0.002	0.209***
	(0.001)	(0.002)	(0.002)	(0.002)	(0.032)
Max. IIF in past 6 m.			−0.004***	0.002***	
			(0.001)	(0.001)	
R^2	0.064	0.053	0.078	0.064	0.257
Observations	1085399	272153	1085399	1085399	2596

Dependent variable "cit": all citations; "cit (n.s.)": only non-self citations. First four columns: APS data, the temporal unit of observation is the month. Last column: ISI data, the temporal unit of observation is the year.

and 1996.[15] Self-citations can be expected to explain some of the estimate β_1: if an author publishes in a given month, this is, all else equal, a sign of high productivity, and more productive authors will likely have more occasions to cite previous papers of them. While a cumulated advantage process driven uniquely by self-citations would still be at odds with a purely value-based interpretation of citations and bibliometric indicators (and specifically, go against the null of Hypothesis H1), it would largely alter the policy implications, since it is relatively easy to consider indicators which discard them (although the benefits of doing so are debated). This justifies a detailed analysis of "non-self" citations, which can be of minor interest for what concerns the accuracy of bibliometric measures, but can provide a better grasp of the mechanisms at work: hence, the main model is also reproduced on such "third" citations. The first model of Table 1.3 shows that indeed the estimate of β_1 previously given is largely driven by self-citations, but that even discarding them, the effect is highly significant (the coefficient of 0.003 corresponds now to a more reasonable percentage increase of 3.75% in the number of citations per month).

Most importantly, the modeling of the network of citations in Section 1.3 was grounded on two assumptions: 1) that the idiosyncratic values (and

[15]Given the numerous differences in the sample selection, it is unclear to what extent the difference can be attributed to different citing behaviors.

hence the citations) reflect some *objective* value, and 2) that citation behaviors reflect such idiosyncratic values. The observation that $\beta_1 > 0$ has been so far implicitly considered as counter-evidence for Assumption 2: that is, new papers getting published alter the citation flows of previous ones. This is not, however, the only possible interpretation: it could be that papers by highly skilled researchers, who write on average high quality papers, tend to get initially get relatively low amounts of citations. Still, since their works are indeed of high quality, they sooner or later get noticed, and cited; and since the authors are good, they manage to get other papers published. In other words, β_1 could be detecting a downward bias of citations in the first periods after the publication of a scientific work (and an infraction of Assumption 1), rather than an upward bias after the publication of a new work of the same author(s).

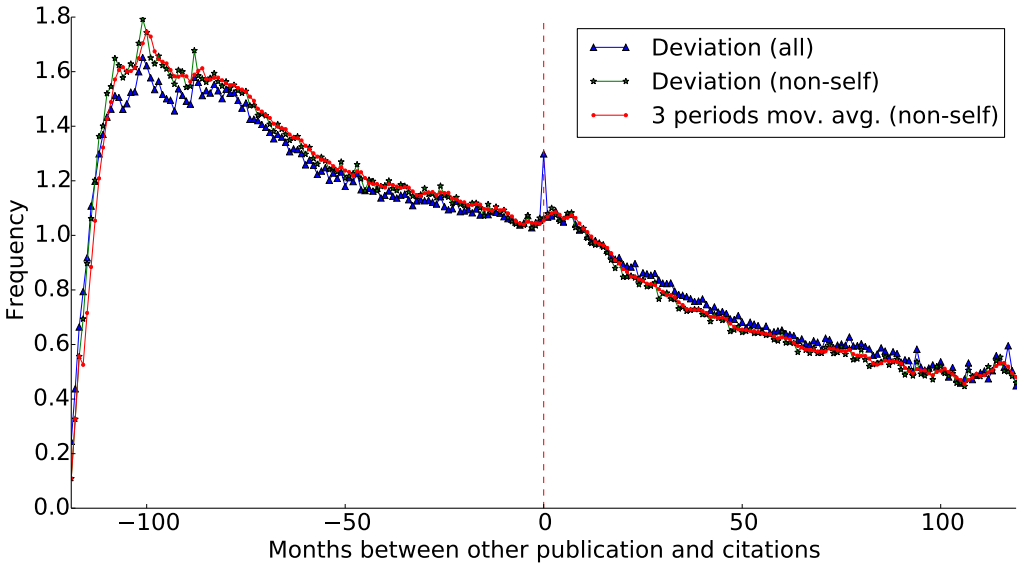
Although such effect is plausible, we can verify that it is not the only one at work through a regression discontinuity design. Figure 1.10 shows the frequency distribution of the distance between a publication and citations to previous publications of the same author, as deviations from a null model with citations distributed homogeneously in time. After time 0 - that is, the month of the publication - there is an evident and unanticipated increase in the flow of citations. Although in the case of all citations there is a clear direct effect (the spike corresponding to 0 is due to the large number of cases in which a new paper cites a previous one by a same author), the increase is evident also when discarding self-citations. The figure can also be interpreted as a difference-in-differences model, comparing the citation increases to "ordinary" papers with those to papers the authors of which just published a new work.

1.6.2 Journals prestige

Since the null $\mathcal{H}_0$ of Hypothesis H1 was unambiguously rejected, the enriched model 1.4 potentially introduces two countervailing effects. On one hand, assuming the "getting noticed" effect is at work, a paper with a high impact will presumably have a stronger influence in getting the author to be cited. On the other hand, if a new paper gets more attention than the previous, it may in part *eclipse* it, effectively reducing its flow of citations.

When the measure of journals influence IIF (see Section 1.5) is introduced as explanatory variable, the results do in fact depend on the exact specification, as summarized in the third and fourth columns of Table 1.3. If all citations are taken into account, having published on a better journal in the last 6 months tends to *decrease* the flow of citations to previous papers, as postulated in Hypothesis H2. But the results change radically if

Figure 1.10: Distribution of citations in time



Frequency of citations, as a function of the distance from the publication of a new work by a same author (time 0). The lines represent the normalized difference from the prediction of a null model with citations distributed homogeneously in time.

self-citations are discarded: in this case, the estimate of β_2 is positive, while the estimate of β_1 is no more significant.

> **Result 2:** the null hypothesis of H2 is valid (and $\beta_2 < 0$) only if self-citations are taken into account; otherwise, the prestige of a journal is correlated with an *increase* of citations to previous works.

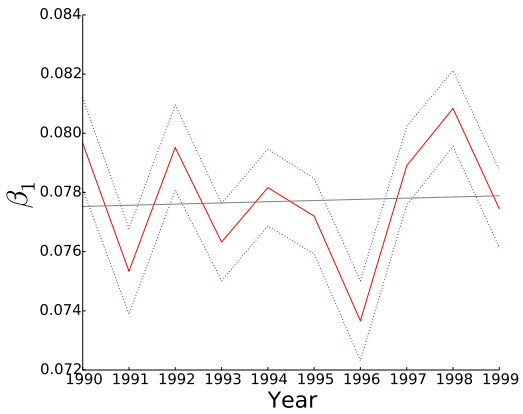
1.6.3 Sensitivity tests

The high heterogeneity characterizing the data was already mentioned in Section 1.5. In particular, the small set of papers with extremely high citations and/or publications of the same author could be to due very specific kinds of research organizations (in the field of high energy physics, for instance, publications with hundreds of authors are not uncommon), or even to homonyms (the APS database does not allow to distinguish authors with same name and surname) which could be influencing the results. Columns 2 to 5 of Table 1.2 report the result for the main model ran only on sets of "low volume" papers, defined in terms of citations or publications of a same author. The results are in line with the full estimation, in some cases actually increasing the estimate of β_1. The lowest value, 0.05, is obtained when restricting to a maximum of 1 citation per month - a criteria which however can be easily falsified even by "unexceptional" publications, and which could be a cause of noise. In general, those sensitivity tests show that the results are not driven by exceptional papers or scientific practices.

All results presented in Table 1.2 come from data about papers published in 1990 only. In order to enhance the external validity of such results, the analysis was repeated for each year between 1990 and 1999. Figure 1.11 shows that the estimation for the relevant coefficient of the base model is always significantly different from 0 - and fairly stable across time. The available data, which ends in 2010, does not allow to run the same analysis for years after 1999.

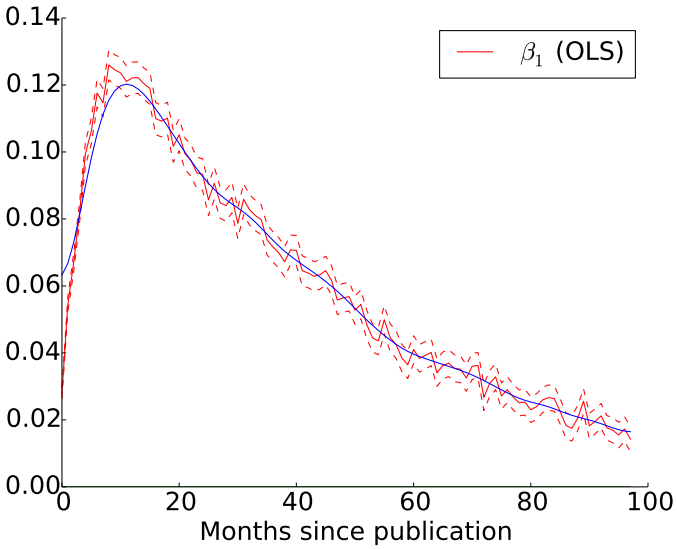
Figure 1.12 investigates the impact of new publications on citation flows as a function of months elapsed since the original publication. Notice that the model is ran on cross sections, so it is a simple OLS rather than the two-way model presented in Table 1.6: this implies that the magnitude of the coefficient is not directly comparable, and that the result could be deeply affected by other confounding factors. All that said, the fact that the coefficient is clearly decreasing in time suggests that the "getting noticed" (or self-citing) effect is particularly strong in the few months after a paper has been published.

Figure 1.11: Evolution of β_1 over the years.



The dashed lines represent the 95% confidence intervals, the solid straight line is the best linear fit.

Figure 1.12: Variation of β_1 over the life of a publication.



Result of cross-sectional OLS versions of Equation (1.3), as a function of the number of months elapsed since publication (in dashed lines, the 95% confidence intervals).

Finally, in order to ascertain the sensitivity of the methodology to the discipline under analysis, the same analysis was applied to a smaller sample of papers in the field of Economics, extracted from ISI Web of Science. The criterion adopted in order to have a meaningful sample for the present analysis was the following: all papers published in 1995 on the Scandinavian Journal of Economics (chosen because of its middle range status in terms of Impact Factor) were considered (those are the units of observation), and for each of them the flow of citations was retrieved, as well as the flow of publications of the same authors. The resulting network, in the scope of the present analysis, suffers less severely from truncation problems. Because of the much lower numerosity of the data, however, the whole year was adopted as fundamental unit of time, rather than the month, and the estimation was made with cross-sectional fixed-effects only, without time effects. The result is reported in the last column of Table 1.3: the estimated effect (0.21) is actually much stronger than what measured so far. Although its entity is not directly comparable, due to the different design, (in particular, the APS journals are presumably more homogeneous, in terms of impact, than the Economics journals appearing in the ISI data set), it provides a strong signal that the cumulative advantage is a phenomenon not restricted to the field of physics.

1.7 Conclusions

A framework for the study of network formation under constraints is being presented. Constraints can be *positive* (sets of links appearing in any possible network) and *negative* (sets of forbidden links), and can change in time, depending on the choices of agents. Some theoretical results of Haller (2012) concerning the existence of Nash networks were generalized to the presence of negative constraints. A model of repeated constraints in discrete time was then proposed, allowing to capture the intrinsic temporal structure of several types of endogenously formed social networks.

The model is then specialized in order to study the formation of the network of citations among scientific papers, and the resulting predictions are tested against the data. The hypothesis by which citation flows truthfully reflect some kind of intrinsic value of cited papers is rejected, showing that instead environmental factors play an important role in shaping them. In particular, it is shown that *after* the publication of a scientific paper, there is an immediate raise in the flow of citations to previous publications by the same author(s), even when discarding self-citations. The status of the journal in which the new work is published, measured as a specialized form

of the Impact Factor, has different consequences depending on whether self-citations are considered: if they are discarded, it implies an increase in the flows of citations to previous papers. Although the intuition behind such results is far from being new, the novelty of the approach lies in its ability to distinguish a *causal* effect of environmental factors which goes beyond anecdotal evidence, or mere correlations.

The methodology adopted in the empirical exercise, even if capturing possibly a very limited share of the environmental factors which shape citation flows (for instance, the effect on the popularity of *subsequent* papers may be much larger, but impossible to measure due to endogeneity issues), can be applied on any given network of citations (for instance, on networks regrouping works in the same disciplines, or in the same period) and give an indication of how strong such factors are.

Finally, the theoretical model can be specialized to study many different kinds of social networks, and provide empirical researchers with tools to go beyond what the mere *static* analysis of networks allows to identify.

Chapter 2

Boundedly Rational Opinion Dynamics in Directed Social Networks: Theory and Experimental Evidence

2.1 Introduction

Research on opinion dynamics and learning in social networks has recently received increasing attention in the economic literature (see Jackson and Yariv, 2010, and Acemoglu and Ozdaglar, 2011, for comprehensive reviews). This growing interest reflected two main factors. At the theoretical level, the development of powerful new tools of analysis. At the empirical level, a significant increase in the availability of data sets to test the theoretical predictions. Several frameworks have therefore been proposed to model opinion dynamics in social networks, based on different assumptions regarding the information transmission mechanisms and the sophistication of individuals. Two main streams can be identified within this literature.

A first group of works includes those that assume perfect rationality, usually together with perfect information about the network structure and the probability distributions of states of nature. The aim of these studies is to establish the optimal strategies and the feasibility conditions in order to reach an estimate of some state of the world which is correct, or asymptotically correct, depending on the setting. The basic framework was laid down by Gale and Kariv (2003). Compared to previous works in the field of social learning (such as Bikhchandani et al., 1992, Banerjee, 1992 and Smith and Sørensen, 2000), their main contribution was to analyze the *repeated* interaction of Bayesian individuals over non-trivial (exogenously given) network structures. Thereafter, this approach was extended in several directions, as in the work of Acemoglu et al. (2011), who consider the asymptotic properties of opinion aggregation over a growing random network. Acemoglu et al. (2010) assume that information flows are also "tagged", i.e., each agent knows the origin of each element of information she receives, and uses such "meta-information" optimally. In their framework, the only obstacle to recovering the hidden state of nature is the fact that communication is costly, and hence individuals may communicate only for a limited amount of time.

The second group of works follows a more pragmatic approach: since reaching a correct consensus is, even when feasible, generally characterized by a high computational complexity and degree of coordination, it is claimed that studies assuming perfect rationality cannot credibly model the way in which humans – and not only humans – process information when communicating in social networks. As a consequence, rather than starting from individual or social *objectives* and deriving optimal strategies, these works start from *reasonably simple* protocols for belief updating, and examine to what extent, under what conditions and with what dynamics, models of

opinion dynamics lead to plausible beliefs.[1]

The most common framework in this second group of studies is based on the model of opinion aggregation by DeGroot (1974). Although this work does not explicitly refer to a network environment (as do instead French, 1956, and Harary, 1959, in their studies on *social power*, later generalized by Friedkin and Johnsen, 1990), since all agents can communicate with all other agents, each individual can attribute a given weight to others' opinion, so that the weights implicitly define a network. These weights, which are constant over time, then determine the evolution and possible convergence of opinions. DeMarzo et al. (2003) have taken over this model, with some variations: in their work, an existing network of connections between agents is explicitly assumed, while opinions are defined as point estimates rather than probability distributions. After considering a general model in which weights can, to some extent, change from period to period, they focus on the case in which individuals attribute the same weight to all neighbors. The assumption that agents do not take into account the topology of the network, but rather update their opinions by simply taking an average of their neighbors' opinions with equal weights, leads to opinions that, even in strongly connected networks, are biased towards the initial beliefs of the most influential (i.e., better connected) subjects. In this setting, agents' *social influence* depends on their positions in the network and, in particular, on their (and their neighbors') *outdegree*.

Some works in the related field of *distributed sensors* (e.g. Olfati-Saber and Murray, 2004) characterize the class of network topologies in which unbiased estimation is obtained even with very simple communication protocols. Other authors (Jadbabaie et al., 2012, Jadbabaie et al., 2013) extend this framework by assuming that agents receive, instead of an initial signal, a continuous flow of (private) information, which they process in a Bayesian fashion; Golub and Jackson (2010) consider families of increasingly large networks, and provide asymptotic properties that guarantee convergence of opinions to the true state of nature. Bala and Goyal (1998) combine the social learning mechanism with a "learning by doing" approach, in which individuals obtain information also by observing their own previous outcomes. Buechel et al. (2012), in their extension of the basic DeGroot model, allow for opinions reported by individuals to differ from actual beliefs, where the difference reflects a preference for *conformity* or *counter-conformity*. Hegselmann and Krause (2002) add the feature that the *structure* of the network

[1]This perspective is taken not only in studies focusing on opinion dynamics in social networks, but more generally in the literature on social learning (see e.g. Ellison and Fudenberg, 1993).

itself may be altered endogenously depending on the similarity of opinions. Their work is part of a literature focusing on the effects of the *"bounded confidence"* phenomenon in terms of non-convergence (see e.g. Hegselmann and Krause, 2005, Dittmer, 2001, Fortunato, 2004), but its novelty lies in the fact that they consider the interplay of such phenomenon with the structure of an underlying social network.

In motivating the search for a model of opinion aggregation, the first requirement put forward by DeGroot (1974) is that *"The process that it describes is intuitively appealing"*. Given the explicit quest of this stream of literature for *credible* models, such models were often preferred over Bayesian ones for interpreting observational and experimental data. Banerjee et al. (2013) exploit a natural experiment focusing on network data from small municipalities in rural India. They conclude that the best predictor for the influence of "injection points" is their eigenvector centrality (as predicted by DeMarzo et al., 2003). Möbius et al. (2010) present a field experiment investigating the diffusion of information over the network defined by friendship relations on Facebook, and seed such network artificially with noisy signals on some hidden state of the world. This allows the authors to ascertain the presence of strong information decay, test the main predictions of the DeGroot (1974) model, and compare them with the possibility of *tagged* information (as proposed by Acemoglu et al., 2010).

In a study closely related to the present one, Corazzini et al. (2012) present a laboratory experiment aimed at testing the presence of persuasion bias in opinion formation within communication networks. Consistent with the predictions of DeMarzo et al. (2003), their results indicate that the social influence of individuals depends on their eigenvector centrality, which in turn depends on the number of other individuals who, directly or indirectly, listen to them (i.e., their *outdegree*). The findings, however, are also consistent with the presence of an *indegree effect*, as agents with higher indegree have higher social influence. In order to explain this phenomenon, Corazzini et al. (2012) suggest a framework based on the assumption that the weights each individual places on her neighbors' opinions are positively related to their neighbors' indegree. Intuitively, more informed individuals receive higher weights when opinions are updated and, as a consequence, also have higher social influence.

Against this background, the objective of this paper is twofold. First, we run an in depth analysis of the model by Corazzini et al. (2012), deriving some results concerning *linear* updating models in general, and characterizing the way in which the *efficiency* of the specific process of opinion aggregation depends on the topology of the underlying network and on the choice of parameters. Second, we present a laboratory experiment explicitly designed

to test the causal effect of indegree on social influence. The structure of the directed network used in the experiment allows us to manipulate indegree without affecting the outdegree and eigenvector centrality of different nodes, thus providing a clean test of the effects of indegree on social influence, which cannot be explained with the model by DeMarzo et al. (2003).

We show that, in balanced networks, placing higher weight on neighbors with higher indegree is less efficient than placing equal weights on all neighbors. On the other hand, in unbalanced networks it is generally more efficient to place higher weight on neighbors with higher indegree, and there exist networks in which it is optimal to place weight only on agents with highest indegree. Empirically, we find strong evidence of an indegree effect on opinion formation. The social influence of an agent is positively and significantly affected by the number of individuals she listens to.

The remainder of the paper is structured as follows. Section 2.2 provides the theoretical framework (technical details are in Appendix A). Section 2.3 describes the experimental design (experimental instructions are in Appendix B). Section 2.4 presents the results. Section 2.5 concludes with a discussion of the key findings.

2.2 Theoretical Framework

Following DeMarzo et al. (2003), consider a setting where a set $\mathcal{N} = \{1, ..., n\}$ of agents, communicating within a social network, want to estimate some unknown state of the world represented by the parameter $\theta \in \mathbb{R}$. Each agent starts with some initial information x_i (henceforth referred to as a *signal*) about θ. For simplicity, we assume that $x_i = \theta + \varepsilon_i$, with $\varepsilon_i \sim N(0, \sigma^2)$ independent across agents. The structure of the network is represented as a directed graph with adjacency matrix q, where $q_{ij} = 1$ if agent i listens to agent j, and 0 otherwise (we assume $q_{ii} = 1$ for every i).[2] We denote as $S_i \subset \mathcal{N}$ the *listening set* of an individual i, that is, $j \in S_i \iff q_{ij} = 1$. Communication takes place in discrete time: at each $t \geq 0$, agents report their current belief to their neighbors. Defining the vector of initial beliefs as $y^0 = x$, we assume that, for each $t \geq 0$, agent i updates her belief according to an updating rule

$$y_i^{t+1} = f_i(y_{i_1}^t, \ldots, y_{i_K}^t)$$

[2]Notice that in the graphical representations of networks presented below, an arrow from i to j means that agent i *talks to* (rather than listens to) agent j, that is, $q_{ji} = 1$. This is different from standard convention, but consistent with the instructions of the experiment presented in Section 2.3 and with the direction of information flows.

where $i_1, \ldots, i_K$ are the agents in i's listening set S_i (notice that $i \in S_i$). Once the network structure q, the updating rules $f_1, \ldots, f_n$, and the initial signals x are determined, the evolution of opinions is obtained. Given an agent i, if $y_i^\infty = \lim_{t\to\infty} y_i^t$ exists, it will be referred to as her *convergence* or *asymptotic* belief. In what follows, we will be particularly interested in the case in which convergence beliefs exist and coincide for all i: if this is the case, we will refer to such limit as the *consensus* belief, and denote it simply as y^∞.

The foundations for this framework were laid by DeGroot (1974), who considers the case in which "opinions" are probability distributions rather than real numbers, and f is *linear*:[3]

$$y_i^{t+1} = \sum_{j\in S(i)} \pi_{ij} y_j^t. \tag{2.1}$$

where π_{ij} is the weight placed by agent i on agent j. DeMarzo et al. (2003), following the studies on *social power* of French (1956) and Harary (1959), introduce an exogenously given network structure q, and analyze in detail the specific updating rule

$$y_i^{t+1} = \sum_{j\in S(i)} \frac{y_j^t}{|S(i)|}. \tag{2.2}$$

One key feature of the boundedly rational updating rule in (2.2) is that agents do not use in any way the information contained in q: they simply form their opinion by averaging all opinions they get to know, irrespective of the network structure.

Consider now a *generalized boundedly rational* updating rule (henceforth GBR) that is still linear, but with weights defined as follows:

$$\pi_{ij} = \frac{q_{ij} d_j^\rho}{\sum_h q_{ih} d_h^\rho} \tag{2.3}$$

where d_j is agent j's indegree ($d_j = |S(j)| - 1$) and $\rho \in [0, \infty)$ is a fixed parameter. Such rule, which was first introduced by Corazzini et al. (2012), provides a simple generalization of the updating rule in (2.2), while incorporating plausible and interesting features. Intuitively, when weighting the opinions of neighbors, agents attribute relatively more importance to those neighbors who have more direct sources of information, i.e., neighbors with higher indegree. To illustrate, let us consider some examples.

[3] In the context considered by DeGroot (1974), $S(i)$ corresponds to $\{1, \ldots, n\}$.

When $\rho = 0$,

$$\pi_{ij} = \frac{q_{ij}}{\sum_h q_{ih}} = \frac{q_{ij}}{|S(i)|}$$

we obtain the updating rule in (2.2), as in DeMarzo et al. (2003): agents update their opinions by averaging the opinions they get to know, while placing *equal weights* on all neighbors.

When $\rho = 1$,

$$y_i^{t+1} = \frac{\sum_{j=1}^n q_{ij} d_j y_j^t}{\sum_{j=1}^n q_{ij} d_j}, \tag{2.4}$$

i.e., the opinion of each neighbor is weighted *proportionally* to her indegree.

In the limit case $\rho \to \infty$,

$$y_i^{t+1} = \sum_{j \in \arg\max_h q_{hi} d_h} \frac{y_j^t}{|\arg\max_h q_{hi} d_h|}$$

so that each agent only listens to the individual(s) with maximum indegree in her listening set. This limit case is useful to provide an intuition of what happens more generally for *high values* of ρ: individuals with higher indegree tend to be the most influential.[4]

DeMarzo et al. (2003) show that in a *strongly connected* network, i.e. where every agent can influence every other agent, any linear updating rule such that $\pi_{ij} > 0$ whenever $q_{ij} > 0$ guarantees convergence to a consensus belief.[5] This implies that, in our setting, convergence is ensured for any $\rho \in [0, \infty)$.[6] Moreover, if we rewrite Equation (2.1) in matrix form as $y^{t+1} = \Pi y^t$, the vector of consensus beliefs $\bar{y}^\infty$ must satisfy the condition

$$\bar{y}^\infty = \Pi \bar{y}^\infty. \tag{2.5}$$

i.e. y^∞ is a *right eigenvector* of Π, with eigenvalue 1. DeMarzo et al. (2003) also show that y^∞ can be written as a weighted sum of the initial

[4] It should be noted, however, that networks can be designed in which, even for arbitrary large ρ, agents with maximum indegree *and* maximum outdegree do *not* have highest social influence.

[5] This is a particular case of their Theorem 1, exploiting the fact that the listening matrix is row-stochastic, and hence describes a Markov chain which is irreducible and aperiodic.

[6] The result does not apply to the limit case $\rho \to \infty$: in fact, with the rule described in Equation (2.2), convergence of beliefs is *not* guaranteed, as can be verified from the simple counterexample defined by $S(A) = \{B, D\}$, $S(B) = \{A\}$, $S(C) = \{B, D\}$, $S(D) = \{C\}$. In this case, agents A and B, who have maximum indegree and are not directly connected, will never change their own beliefs, and hence, assuming their initial signals differ, agreement will not be reached.

signals:

$$y^{\infty} = \sum_{i=1}^{n} w_i x_i, \tag{2.6}$$

with w being the unique (normalized) solution to

$$w\Pi = w. \tag{2.7}$$

If we consider Π as the adjacency matrix of a *weighted* network, Equation (2.7) can be interpreted as stating that the *social influence* w_i of an individual i corresponds to her *eigenvector centrality* (Bonacich, 1972, Jackson, 2010). It is important to observe that Equation (2.7) implies that, in the updating rule proposed by DeMarzo et al. (2003), social influence is *increasing* in outdegree (see Appendix A, Theorem 6).

Let us now consider whether consensus beliefs are correct, in the sense of being optimal aggregates of agents' initial information. Given that all signals are equally informative, the consensus belief is correct if $w_i = \frac{1}{n}\ \forall i$, i.e.

$$y^{\infty} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}. \tag{2.8}$$

Different updating rules can therefore be compared by using the following measure of efficiency:

$$E = -\sum_{i=1}^{n} \left(w_i - \frac{1}{n} \right)^2 \tag{2.9}$$

where $E = 0$ if consensus beliefs are correct, while $E < 0$ otherwise (notice $E \in [-1, 0]$).

DeMarzo et al. (2003) have analyzed the GBR rule under the restriction $\rho = 0$, showing that it will not, in general, lead to a correct consensus in most network structures, as more connected agents have excessive social influence. When $\rho > 0$, the GBR rule does not always neutralize such inefficiency: while it may lead to a consensus which is closer to the correct one than for $\rho = 0$, this is not a general rule. We thus start by asking more generally whether there exist rules of thumb that lead to correct beliefs for any given network structure. The answer is provided by the following theorems.

Theorem 1. *Given any strongly connected network $\bar{\mathcal{G}}$, there exists a linear updating rule $F_{\bar{\mathcal{G}}}$ which guarantees convergence to the correct consensus.*

Proof. See Appendix A. □

Theorem 2. *Given any linear updating rule $\bar{F}$ with weights $\bar{\pi}_{ij}$ which depend on* local *properties of the network around i, there exists a strongly connected network $\mathcal{G}_{\bar{F}}$ on which $\bar{F}$ does* not *guarantee convergence to the correct consensus.*

Proof. See Appendix A. □

While the first result is positive, stating that for any given network it is possible to find an *optimal* linear rule, the second ends the quest for the "perfect" rule of thumb: linear rules cannot be both correct and *simple*, in the sense that weights are determined only by the *local* properties of a network.[7]

For a given value of ρ, the efficiency of the resulting updating rule will clearly depend on the topology of the network. We will denote as $E_\rho(\mathcal{G})$ the efficiency of the GBR rule when implemented over a network $\mathcal{G}$ with the given value of ρ, and as $\rho^*(\mathcal{G})$ the value of ρ that maximizes $E_\rho(\mathcal{G})$. In the following, some relevant classes of networks will be defined on which the GBR rule displays particular features in terms of efficiency. In all cases, it will be assumed that networks are strongly connected.

Let us define two nodes (j, k) as *equivalent* if, after switching their labels, it is possible to arrange the other labels of the network in order to obtain an exact copy of the original one. Then, we can provide the following definitions:

Definition 2. *A network structure is* anonymous *if all nodes are equivalent.*[8]

Definition 3. *A belief updating rule F is* anonymous *if all f_i are symmetric in arguments y_j^t and y_k^t, for any pair of equivalent nodes $j, k \neq i$.*

In other words, a rule F is anonymous if the *labels* of agents do not play any role. Those two definitions allow us to state the following basic result, which generalizes Theorems 1 and 2 in French (1956),[9] as well as Theorem 9 in Harary (1959).

[7] It is interesting to consider a *constructive* proof of Theorem 1: an algorithm which for any given network $\bar{\mathcal{G}}$ finds a linear rule which is efficient over it. Such algorithm is described in Example 1 of Appendix A, and is based on the existence of a closed path p_n passing through all nodes of the network (possibly more than once). Indeed, it clearly does not qualify as a "rule of thumb": in particular, the choice of p_n requires significant ex-ante coordination among agents.

[8] This is stronger than the concept of *regular graph*, which only considers the indegree of vertices (rather than their position in the network), and corresponds instead to the concept of "*automorphic group*" employed by Harary (1959).

[9] While Theorem 2 by French (1956) is indeed a special case of Lemma 2, his Theorem 1 additionally states that on complete networks, the consensus belief is reached *in one step*.

Lemma 2. *On any given anonymous, strongly connected network $\mathcal{G}$ on which all agents play a same linear anonymous updating rule,*

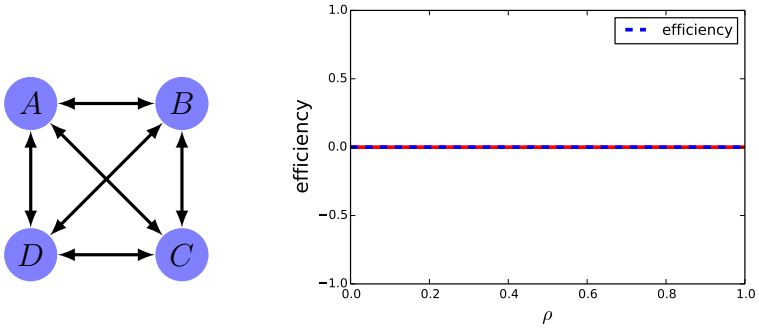
1. *The beliefs in each period are uniquely determined by $\pi_{11}, \ldots, \pi_{nn}$, the self attributed weights,*
2. *The asymptotic consensus beliefs are correct: $w = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$.*

Proof. See Appendix A. □

One may wonder if anonymity *or* linearity alone is a sufficient requirement for Lemma 2 to hold. The answer is negative: two counter-examples are given, respectively, by the rules "weight the opinion coming from the highest labeled neighbor as much as the average of all others" (which is linear but not anonymous) and "weight the highest opinion coming from the neighbors as much as the average of all others" (which is anonymous but not linear).

Since the GBR updating rule is both linear and anonymous, a direct consequence of Lemma 2 is that, given an anonymous, strongly connected network, it will (perform equally, and) converge to the true average of signals for any choice of ρ. For an example, see Figure 2.1, displaying a *complete network* with four nodes: since the value of ρ is irrelevant, the efficiency of the GBR rule is constant; more precisely, since the consensus belief is correct, the loss function is always equal to 0.

Figure 2.1: An anonymous *complete* network



The requirement of anonymity of a network is a very strong one. In fact, the class of networks on which the GBR rule will perform equally for any given $\rho > 0$ is significantly larger than that of anonymous ones. To see this, notice that on any *regular* network,[10] independently of ρ, each agent will

[10] A *directed* network is regular if all nodes have the same indegree and outdegree (the two must necessarily coincide).

attribute the same importance to the opinion of each neighbor, since all of them will have the same indegree. Interestingly, the regularity of the network is also a *necessary* condition for ρ to be irrelevant, as stated in the following theorem.

Theorem 3. *On a strongly connected network $\mathcal{G}$, the convergence belief and social weights obtained under a GBR rule with given $\rho > 0$ coincide with the ones obtained under $\rho = 0$* if and only if *$\mathcal{G}$ is regular.*

Proof. See Appendix A. □

The next result concerns a wider class of networks, which includes all anonymous ones, but also many others, such as undirected networks and regular networks.

Definition 4. *A network is said to be* balanced *if each node has indegree equal to outdegree.*[11]

Theorem 4. *On any given balanced, strongly connected network $\mathcal{G}$, the GBR with $\rho = 0$ is more efficient than with any $\rho > 0$.*

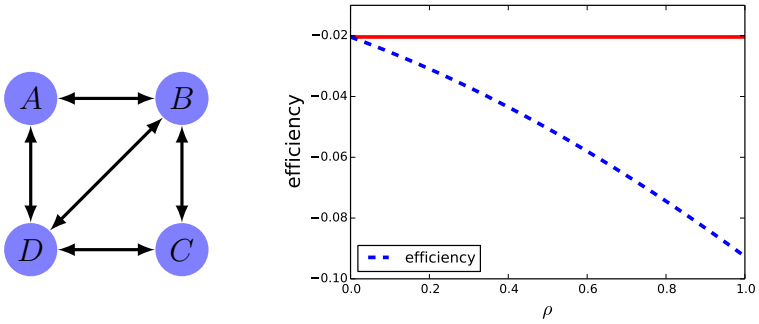
The proof for this theorem is in Appendix A. Still, it can be interesting to consider an intuitive explanation. On this class of networks, the problem of persuasion bias is attenuated. Although it is present in the first period, in the long run agents with higher outdegree exploit their influence to convey *richer information*, since they also have higher indegree. Intuitively, their belief has higher weight, but their own *initial* opinion gets *diluted* in their belief. This does not occur when $\rho > 0$, which causes agents with higher outdegree to place *even higher* weight on their own signal.

For an application of Theorem 4, consider the network structure in Figure 2.2. While it is not anonymous (nodes B and D are identical, but they differ from nodes A and C), it shares with the complete network the feature of being *undirected*. Figure 2.2 shows that the optimal value of ρ is 0. Notice that Theorem 4 *only* holds asymptotically: it is easy to provide counterexamples in finite time, by replacing in the definition of efficiency (Equation 2.9) the weights w_i with those calculated after a finite number of periods.

More importantly for our purposes, outside of those specific classes of networks, things change radically. While there can be cases of *unbalanced*

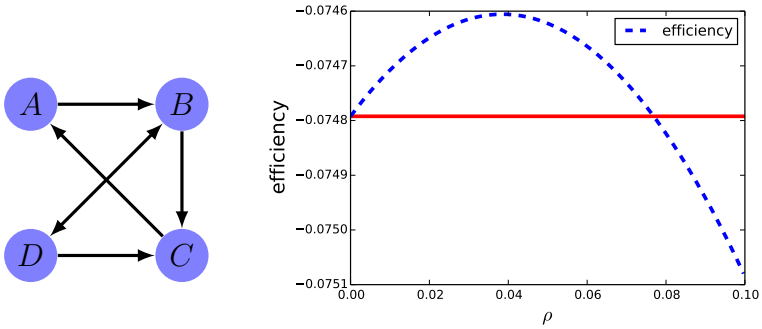
[11] This definition is adopted from Olfati-Saber and Murray (2004). It is more general than the ones used by Corazzini et al. (2012) and DeMarzo et al. (2003), and corresponds to the concept of *isograph* used in graph theory; in particular, all weakly connected isographs are strongly connected and are also referred to as *Eulerian digraphs* (see Sridharan and Parthasarathy, 1972).

Figure 2.2: An *undirected*, and hence balanced, network



networks on which a lower ρ still means higher efficiency, this is not the rule. The network presented in Figure 2.3, which is the one used in the experiment presented in Section 2.3, is characterized by an optimal value of ρ that is positive (0.04). For higher values of ρ, the social influence of B, who has a higher indegree, increases, while the one of D decreases.

Figure 2.3: The network used in the experimental setting



The following theorem states that no value of ρ is in principle "too high".

Theorem 5. *Given any $\bar{\rho} \in [0, \infty)$, there exists a strongly connected network $\bar{\mathcal{G}}$ such that $\rho^*(\bar{\mathcal{G}}) > \bar{\rho}$.*

Proof. See Appendix A. □

In short, our results can be summarized as follows. In anonymous networks the weights placed on neighbors are irrelevant, since all linear rules are efficient. In balanced networks, placing higher weight on neighbors with higher indegree ($\rho > 0$) is generally less efficient than simply placing equal

weights on all neighbors ($\rho = 0$). Finally, in unbalanced networks, it may be optimal to place higher weight on neighbors with higher indegree, and there exist networks in which the optimal value of ρ is arbitrarily high. In the next Section we present an experimental test of the effects of indegree on opinion dynamics and social influence.

2.3 Experimental Design

The experiment is designed to test the effects of agents' position in a communication network on their social influence. More specifically, our experimental design allows us to manipulate agents' indegree without affecting their outdegree and the corresponding eigenvector centrality. Therefore, it enables us to provide a clean test of the effect of indegree on social influence, that would be absent under either Bayesian updating or boundedly rational updating à la DeMarzo et al. (2003).

2.3.1 Task

At the beginning of the experimental task, individuals are anonymously matched in groups of four. In each group, subjects are connected through a communication network, and each subject is assigned a label (A, B, C, D) that defines her position in the network. The task is based on a discrete time setting played over 8 rounds, during which the network structure and positions do not change. Before the first round, each subject is assigned an integer number randomly drawn from a normal distribution, henceforth referred to as a *signal*, denoted with x_A, x_B, x_C, x_D, respectively: concerning the origin of such numbers, subjects are only informed that they are randomly drawn by the system. Then, in each round, the subjects are asked to guess the *average* $\bar{x} = \frac{x_A+x_B+x_C+x_D}{4}$ of the four signals in their group. Notice this is a slight deviation from the theoretical setting, in which individuals want to estimate an unknown state of the world θ. This choice allows subjects to concentrate on the aggregation of information rather than on statistical inference; moreover, it reinforces the link between actions and payoffs, by limiting the role of chance.

In order to be able to update their beliefs, at the beginning of each round subjects receive information from the other group members they are connected to. More specifically, at time t each individual is informed about the guesses at time $t-1$ of the other group members connected to her (the network structure, that defines who receives information from whom, is described in the next subsection). Therefore, while in the first round subjects

only directly know their own signal, over successive rounds they directly or indirectly receive information about the signals received by the other group members. If all four group members optimally process the information they receive, over successive rounds each of them can correctly guess $\bar{x}$.

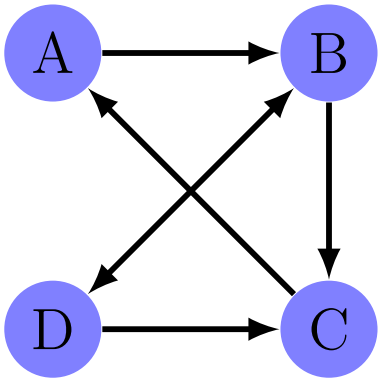
The mechanism for eliciting beliefs is as follows: each individual is informed that at the end of the session, one round will be randomly extracted to determine earnings.[12] Given the guess y^* of the individual in that round, and the average of signals in the group ($\bar{x}$), the individual's payoff is 15 euro minus the absolute difference between y^* and $\bar{x}$, in addition to a show-up fee of 5 euros. This implies that individuals have an incentive to report in each round their best guess for the group average. We adopt a linear scoring rule for three reasons. First, a quadratic scoring rule, commonly used for belief elicitation, would substantially complicate the calculation of payoffs, hence increasing the likelihood of mistakes due to mis-comprehensions. Second, for a given average gain, a quadratic scoring rule would increase the likelihood of earning very small payoffs, thus adversely affecting the incentive to exert effort in the task. Third, given normality of signals, our setting is perfectly symmetric, so that the median of the posterior's distribution coincides with the mean. In addition, experimental subjects were explicitly advised that, in case they got to know a subset of the signals, the optimal strategy was to simply declare their average. The choice of presenting such information, rather than a description of how the signals were generated, was taken in order to relieve subjects from part of the statistical reasoning, still without suggesting any interpretation of the network topology.

2.3.2 Treatments

Figure 2.4 describes the strongly connected directed network structure we use in the experiment. The number of nodes is small in order to provide a simple setting for the experimental subjects, but at the same time sufficiently large to imply interesting opinion dynamics. The network structure implies that A is informed about past beliefs of C, B is informed about past beliefs of A and D, C is informed about past beliefs of B and D, while D is informed about past beliefs of B. The indegree and outdegree of the four nodes are $A = (1, 1)$, $B = (2, 2)$, $C = (2, 1)$, $D = (1, 2)$, respectively. The reason for choosing this specific network structure is that, as explained below, it allows us to cleanly test the key hypotheses of the experiment.

[12] Paying subjects based on a random round is a standard design choice which allows to avoid income effects while constantly providing monetary incentives. In practice, rounds 1 and 2 were never extracted, so the probability of gaining zero euros only because of bad luck was low.

Figure 2.4: Structure of the communication network



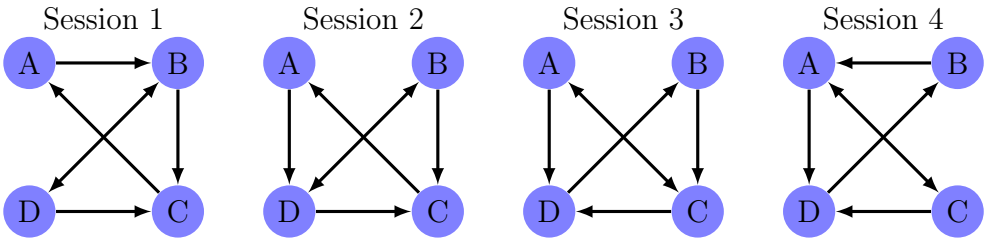
The treatment variable is the node that the subject is assigned to within the network. The four treatments (nodes A, B, C and D) are implemented in a within-subjects design. This means that, in an experimental session, each subject performs the task four times, thus taking part in four subsequent *phases* of 8 rounds (32 rounds overall). In each of the four phases, each subject is randomly assigned to a different node (position) in the network, in such a way that, within a session, each individual is assigned to each node in exactly one of the four phases. Subjects receive a different set of signals at the beginning of each phase, while the composition of the groups is unchanged throughout the four phases.

Since we aim at assessing if and how agents' social influence is affected by their position in the network, it is important to control for the possible confounding effects of the labels attached to each node (A, B, C, D) and of subjects' visual location in the network (upper left, upper right, bottom left, bottom right). It is possible, for instance, that subjects tend to give more importance (higher weight) to nodes denoted by letters that come first in the alphabet (e.g. A vs D). Similarly, subjects might tend to give more importance to nodes located in the top-left of the network visual display, as opposed to the bottom-right. In order to control for such spurious effects, we implemented the four treatments in each of four sessions keeping constant the networks structure, while changing in each session the spatial disposition of the nodes, as detailed in Figure 2.5.[13] The same four sets (one per phase) of signals were used in each of the four sessions, and matched to labels/spatial dispositions. Hence, the only difference between sessions is the *topological* location of each node in the network: this allowed us to cleanly identify the

[13]This means that, for instance, the node that has label B and upper-right position in session 1, has then label D and bottom-left position in session 2, label C and bottom-right position in session 3, label A and upper-left position in session 4, respectively.

causal effects of network structure.

Figure 2.5: Network structure, by session



2.3.3 Hypotheses

Consider the network structure in Figure 2.4. As detailed in Table 2.1, there exists a set of strategies that allows each of the four network members to find out $\bar{x}$ in just four rounds. Indeed, there exist *several* possible combinations of strategies that result in correct beliefs. With such optimal strategies, the four agents have equal social influence weights in consensus beliefs, i.e. $w^* = [0.25, 0.25, 0.25, 0.25]$. It is worth remarking that such strategies do not simply quickly *converge* towards the true $\bar{x}$: in equilibrium, each subject is choosing *in each round* the expected value of $\bar{x}$ conditional on known information. That is, in principle the experimental setup features no tension between accuracy in the final round and in the previous ones.[14]

Table 2.1: Optimal strategies for each network position, by round

Round	A	B	C	D
1	x_A	x_B	x_C	x_D
2	$\frac{x_A+y_C^1}{2}$	$\frac{x_B+y_A^1+y_D^1}{3}$	$\frac{x_C+y_B^1+y_D^1}{3}$	$\frac{x_D+y_B^1}{2}$
3	$\frac{x_A+3y_C^2}{4}$	$\frac{x_B+2y_A^2+y_D^1}{4}$	$\frac{x_C+3y_B^2}{4}$	$\frac{x_D+3y_B^2-y_D^1}{6}$
4	y_A^{t-1}	y_B^{t-1}	y_C^{t-1}	$\frac{x_D+4y_B^3-y_D^1}{4}$
≥ 5	y_A^{t-1}	y_B^{t-1}	y_C^{t-1}	y_D^{t-1}

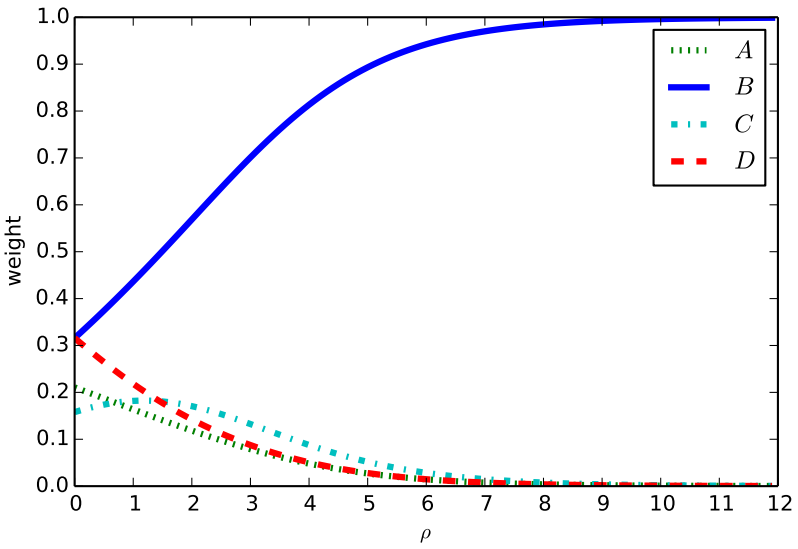
Note: agents' positions in column headings refer to the network structure displayed in Figure 2.4.

Let us now consider the predictions for the GBR updating rule described in equations (2.1) and (2.3). Figure 2.6 shows how the social influence weights

[14]Under alternative (i.e. truncated) distributions of signals, the specific strategies in Table 2.1 would not be optimal *per round* estimates of $\bar{x}$; but on the other hand, whenever the network members did play Bayesian estimates of $\bar{x}$, then the same convergence in 4 rounds would occur.

for each of the four network nodes change as a function of ρ. When $\rho = 0$, agents B and D are the most influential. This, loosely speaking, reflects the fact that both B and D have an outdegree of 2, while A and C have an outdegree of 1: the agents who are listened to by more other agents are the most influential.[15] Also note that A is relatively more influential than C. This reflects indirect social influence, as A communicates to B, who is one of the two most influential subjects, while C communicates to A. For $\rho > 0$, the pattern of social influence weights also reflects agents' indegree. In particular, as ρ increases, B becomes progressively more influential, while the weights of the other three agents tend to zero.

Figure 2.6: Social influence weights as a function of ρ



To summarize, Table 2.2 compares point predictions of social influence weights for $\rho = 0$, $\rho = 1$ and $\rho \to \infty$. When $\rho = 0$, as in DeMarzo et al. (2003), B and D have equal social influence weights, similarly to the case of Bayesian updating. When $\rho = 1$, so that agents update their beliefs using weights that are proportional to indegree, B is the most influential agent. Finally, when $\rho \to \infty$, consensus beliefs tend to agent B's initial opinion.

The pattern described in Figure 2.6 and Table 2.2 provides the predictions to be tested in the experiment. The first hypothesis we test is that agents

[15]Note that B and D have exactly the same weight since they communicate to the same individuals (C and each other) and, while the link from A to B implies that the latter is placing a *lower* weight ($\frac{1}{3}$ rather than $\frac{1}{2}$) on her own belief, she is hence placing lower weight also on the information coming from D.

Table 2.2: Predictions for social influence weights for different values of ρ

	w_A	w_B	w_C	w_D
$\rho = 0$	0.21	0.32	0.16	0.32
$\rho = 1$	0.18	0.39	0.18	0.26
$\rho \to \infty$	0.00	1.00	0.00	0.00

Note: the predictions refer to the updating rule described in equations (2.1) and (2.3) for the network structure displayed in Figure 2.4.

optimally update their beliefs. Empirically, the null hypothesis is that all nodes have equal weights in consensus beliefs:

$$H_0 : w_A = w_B = w_C = w_D = 0.25 \tag{H1}$$

Note that this is a general test of Bayesian updating versus an unspecified alternative. In order to test against the specific alternative of the generalized boundedly rational updating rule, we focus on pair-wise differences between individual weights. More specifically, as shown in Figure 2.6, the updating rule predicts that, *for any* $\rho \geq 0$, an agent in node B is more influential than in either A or C. Conversely, for all other pair-wise comparisons between nodes, the sign of the difference between weights is not independent of ρ. Therefore, the relevant one-sided hypotheses can be stated as follows:

$$H_0 : w_B \leq w_A \quad \text{vs} \quad H_1 \colon w_B > w_A \tag{H2}$$

$$H_0 : w_B \leq w_C \quad \text{vs} \quad H_1 \colon w_B > w_C \tag{H3}$$

Next, we focus on the value of ρ. For $\rho = 0$, analogously to the case of Bayesian updating, the boundedly rational updating rule predicts $w_B = w_D$. On the other hand, for $\rho > 0$, the rule predicts $w_B > w_D$ (see Figure 2.6). We can thus test the effect of indegree on social influence ($\rho > 0$), versus the alternative of no effect ($\rho = 0$), by comparing the social influence weights of agents B and D:

$$H_0 : w_B \leq w_D \quad \text{vs} \quad H_1 \colon w_B > w_D \tag{H4}$$

2.3.4 Procedures

The experiment was conducted in the Experimental Economics Lab of the University of Milan Bicocca between January and March 2013, with 24 subjects participating in each of the four sessions (96 in total). Subjects were

undergraduate students, recruited by e-mail through an online system. The experiment was ran using z-Tree (Fischbacher, 2007). Subjects on average earned 13.8 euro for sessions lasting approximately 80 minutes, including time for instructions, control questions and payments.

Each session consisted of four eight-round phases. Subjects were informed that signals and network positions would be randomly determined at the beginning of each phase, while the composition of the groups would remain the same throughout the session. The four signals for each group/phase were extracted as follows. An integer number θ was extracted from a uniform distribution in a range between 200 and 800. Four positive integers were then randomly drawn from a normal distribution with mean θ and variance 100.

All the experimental instructions, reported in Appendix B, were provided to the participants in written form, and also read aloud at the beginning of the session. Individuals were then asked to answer some control questions. Each participant had the possibility to take notes and make calculations on paper, and also to use an on-screen calculator. Moreover, in each round, the screen reported all the information available (own past guesses and past guesses of neighbors since the beginning of the phase), in order to guarantee perfect recall.

The instructions explicitly explained that, had an individual known with certainty a subset of the signals for her group, her optimal strategy was to report their average. This, together with the fact that individuals had to target the average of four specific numbers (rather than the mean of an underlying distribution of signals) helped us to minimize mistakes caused by inappropriate statistical inference, ensuring that individuals could focus on the process of information aggregation. Control questions guaranteed that such instruction was clearly understood.

2.4 Results

In each of the four sessions, the experimental task was implemented by 24 subjects over 8 rounds in four different phases (32 rounds in total), resulting in 384 observations for each round (24 subjects $\times$ 4 sessions $\times$ 4 phases) and 3072 observations in total. Overall, although there was substantial heterogeneity at individual and group level, subjects generally showed to have clearly understood the experimental task. In the first round of each phase, 94.2 per cent of the subjects truthfully reported their own signal, while 96 per cent of the subjects reported a number within 10 units from their own signal. In the final round of each phase, 24 per cent of the subjects correctly

guessed the average of the four signals within their group. Accounting for rounding errors, 55.5 per cent of the subjects reported beliefs within 10 units from the average of the four signals. Gains (not including the show up fee of 5 euros) averaged to 7.67 euros, and were strictly positive for 63 out of 96 subjects.

2.4.1 Tests of Hypotheses

In order to test hypotheses about the social influence weights of agents in different network positions,[16] we specify each agent's final belief as a linear combination of the initial signals of the four agents in her group:

$$y_i^T = \mu + w_A x_{i,A} + w_B x_{i,B} + w_C x_{i,C} + w_D x_{i,D} + \varepsilon_i \tag{2.10}$$

where y_i^T is agent i's belief in the last round of each phase, $x_{i,j}$ is the signal observed by agent j in i's group, μ is a constant, w_j is the social influence weight of agent j, and ε_i is an idiosyncratic error term. Equation (2.10) is estimated by OLS, under the constraint $\sum_j w_j = 1$ - the presence of the constant μ guarantees that this assumption does not artificially inflate the estimates of interest. The set of regressors also includes full sets of dummy variables for sessions and phases. In order to take into account the dependence of observations belonging to the same group within each session, standard errors are clustered by 24 independent groups (there are 6 independent groups in each of the four sessions).

Table 2.3 presents the main results (the complete results of the estimation can be found in Appendix 2.C) . Since we are focusing on the final observation from each of the four phases, the overall sample includes 384 individual observations. Column (1) reports estimates of social influence weights in absolute terms, as in Equation (2.10). The weights generally differ from 0.25, with a pattern that is qualitatively consistent with the predictions of the generalized updating rule: social influence is highest for node B (0.294) and lowest for node C (0.214). The hypothesis that all nodes have equal weights ($w_A = w_B = w_C = w_D = 0.25$), as predicted under Bayesian updating, is strongly rejected by the data ($p < 0.01$).

Result 1: Bayesian updating is rejected by the data.

Focusing on pair-wise differences between weights (hypothesis H2-H3), w_B is higher than w_C, consistent with the predictions of the generalized

[16]Throughout the discussion of the results, unless otherwise stated, we will refer to the four network nodes using the labels of session 1 (see Figure 2.5). This means that nodes from sessions 2 to 4 are implicitly relabeled so that they are the same as in session 1.

Table 2.3: Estimated social influence weights, overall

	(1) Absolute weights	(2) Relative weights
Signal A	0.268***	0.018
	(0.031)	(0.031)
Signal B	0.294***	0.044***
	(0.013)	(0.013)
Signal C	0.214***	-0.036**
	(0.017)	(0.017)
Signal D	0.224***	-0.026
	(0.023)	(0.023)
Number of observations	384	384

Note: figures reported are OLS estimates of social influence weights associated to the node indicated by the row heading. The weights are expressed in absolute terms (column 1) and as a difference from 0.25 (column 2), respectively. Dependent variable: individual beliefs in final round. All specifications include full sets of session and phase dummies. Standard errors clustered at group level reported in brackets. ***p< 0.01, **p< 0.05, *p< 0.10.

boundedly rational rule, and the difference (+0.080) is strongly significant ($p < 0.01$). Similarly, w_B is higher than w_A, although the difference (+0.026) is not statistically significant ($p < 0.25$).

> **Result 2**: Pairwise differences between social influence weights are consistent with the GBR updating rule.

Next, turning to H4, we find that w_B is higher than w_D and the difference is strongly significant ($p < 0.01$). This leads us to reject the null hypothesis that $\rho = 0$.

> **Result 3**: The social influence of an individual is positively affected by the number of individuals she listens to.

This finding is important, as it indicates that subjects do not place equal weights on all their neighbors, but take into account their neighbors' indegree when aggregating the information they receive from them. As a result, ceteris paribus, subjects with higher indegree ultimately have higher social influence.

In order to shed light on these findings, column (2) reports differences of social influence weights with respect to 0.25, obtained by expressing individual final-round beliefs as deviations from the average of the four group signals. The results indicate that w_B is significantly higher than 0.25 ($p < 0.01$), while

w_C is significantly lower than 0.25 ($p < 0.02$). On the other hand, w_A and w_D are not significantly different from 0.25 ($p < 0.29$ and $p < 0.13$, respectively, for the corresponding one-sided hypothesis). The different test results for nodes B and D provide further evidence against a simple updating rule à la DeMarzo et al. (2003).

2.4.2 Robustness

In order to assess the robustness of the results to the possible effects of outliers, Table 2.4 reports estimates of (relative) social influence weights obtained by eliminating from the sample the groups containing the 1%, 5%, or 10% most extreme observations, where potential outliers are identified by considering, for each group member, the difference between the reported beliefs and the ones predicted by Bayesian updating. This results in a restricted sample size of 380, 364 and 344 observations, respectively. In all cases, the estimates are virtually unchanged relative to the overall sample. The hypothesis that all nodes have equal weights ($w_A = w_B = w_C = w_D = 0.25$) is strongly rejected by the data ($p < 0.01$). The hypothesis that $w_B = w_D$ is also strongly rejected in all cases. Indeed, by eliminating potential outliers, the estimated weights are even more closely consistent with the theoretical predictions of the generalized boundedly rational updating rule. In column (3), for example, where the 10 per cent of the groups reporting the most extreme deviations from optimal predictions are excluded, the estimated relative weights are 0.005, 0.053, -0.045 and -0.012. In all cases, w_B (w_C) is significantly higher (lower) than 0.25.

It should be observed that, although the variance of beliefs held by the four group members falls steadily over successive rounds in all groups, disagreement persists in many cases, so that beliefs do not converge to a consensus in all cases. In order to assess the potential effects of non-convergence, Table 2.5 presents estimates of (relative) social influence weights by individual network position. Focusing on nodes B and D, in columns (2) and (4), respectively, estimated social influence weights are qualitatively unchanged with respect to the overall results in Table 2.3: the relative weight of B is positive and significant, while it is negative and significant for C. Agent D has a negative relative weight in the final beliefs of agent A. Finally, relative social influence weights are not different from zero for C.

Overall, these results indicate that the effects of network structure on social influence reported in Section 2.4.1 are both qualitatively and quantitatively robust to the potential effects of outliers. In addition, they are qualitatively unaffected by the possible non-convergence of beliefs within individual groups.

Table 2.4: Social influence (relative weights), robustness

	(1) 1 %	(2) 5 %	(3) 10 %
Signal A	0.013	0.019	0.005
	(0.031)	(0.032)	(0.027)
Signal B	0.049***	0.051***	0.053***
	(0.014)	(0.015)	(0.014)
Signal C	-0.037**	-0.045**	-0.045***
	(0.017)	(0.018)	(0.017)
Signal D	-0.026	-0.024	-0.012
	(0.023)	(0.023)	(0.017)
Number of observations	380	364	344

Note: the figures reported are estimates of social influence weights, as a difference from 0.25, associated to the subject in the position indicated by the row heading. Dependent variable: individual beliefs in final round. All specifications include full sets of session and phase dummies. Standard errors clustered at group level reported in brackets. ***p< 0.01, **p< 0.05, *p< 0.10. Columns (1) to (3): sample restricted by eliminating groups with most extreme deviations from optimal beliefs (1%, 5%, 10%, respectively).

Table 2.5: Social influence (relative weights), by node

	(1) Node A	(2) Node B	(3) Node C	(4) Node D
Signal A	0.057	0.015	0.007	-0.009
	(0.047)	(0.037)	(0.040)	(0.066)
Signal B	0.038	0.057**	0.015	0.065**
	(0.038)	(0.024)	(0.023)	(0.026)
Signal C	0.013	-0.070**	0.024	-0.110***
	(0.033)	(0.027)	(0.022)	(0.035)
Signal D	-0.109**	-0.003	-0.045	0.053
	(0.041)	(0.025)	(0.031)	(0.048)
Number of observations	96	96	96	96

Note: the figures reported are estimates of the social influence weights, as a difference from 0.25, associated to the subject in the position indicated by the row heading. Dependent variable: individual beliefs in final round. All specifications include full sets of session and phase dummies. Standard errors clustered at group level reported in brackets. ***p< 0.01, **p< 0.05, *p< 0.10.

2.4.3 Further Evidence

The experimental data also allow us to investigate what explains the treatment effects on social influence weights, by looking at how agents at specific nodes aggregate the information they receive in each round. At individual level, there can be two possible, not mutually exclusive, mechanisms explaining differences in social influence between B and D. The first, and most intuitive, mechanism is that may C place a higher weight on the opinion of B than on the one of D, because of B's higher indegree. The second is that, for the same reason, D may be influenced by B relatively more than B is influenced by D.

Table 2.6 sheds light on this issue by presenting estimates of node-specific (absolute) weights based on all updating rounds. Looking at the estimates for node C (column 3), the weight of B (0.435) is substantially higher than the one of D (0.163), and the difference is strongly statistically significant ($p < 0.01$). This provides support to the first of the two mechanisms described above. The comparison of the weights given to each other by B and D is non-trivial, since their respective indegrees are different, as they form their beliefs on the basis of different numbers of neighbors. However, D appears to substantially under-weigh the information coming from B (0.104), whereas B does not substantially under-weigh the information received from D (0.325). The overall effect of indegree on social influence is therefore mainly explained by the way in which information is processed by C: node B, whose indegree is twice the indegree of D, receives a weight that is more than twice as large as the weight for node D.

Finally, since the hypothesis that $\rho = 0$ is strongly rejected, it is interesting to ask what value of ρ provides the best fit for the experimental data. We simulated the generalized updating rule with a wide range of values for ρ, searching for the value that minimizes the sum of squared deviations, over all individuals, between observed (experimental) and simulated final-round beliefs:

$$\hat{\rho} = \arg\min_{\rho} \sum_{g=1}^{96} \sum_{k=1}^{4} (y_{g,k,T} - \bar{y}^{\rho}_{k,T})^2, \tag{2.11}$$

where $y_{g,k,T}$ is the belief of an individual with role k in group g in the final round, and $\bar{y}^{\rho}_{k,T}$ is the corresponding theoretical prediction. This produces an estimate of $\hat{\rho} = 0.30$. Interestingly, this is higher than $\rho^* = 0.04$, the value of ρ that provides the best approximation to the results of the optimal strategy (see Figure 2.3). This indicates that agents, as they should, place higher weight on those neighbors who themselves listen to more peers, but

Table 2.6: Neighbors' absolute weights in current beliefs, by node

	(1) Node A	(2) Node B	(3) Node C	(4) Node D
Lagged belief, node A	0.530**	0.000		
	(0.011)	(0.000)		
Lagged belief, node B		0.675**	0.435**	0.104**
		(0.070)	(0.087)	(0.018)
Lagged belief, node C	0.470**		0.402**	
	(0.011)		(0.087)	
Lagged belief, node D		0.325**	0.163**	0.896**
		(0.070)	(0.049)	(0.018)
Number of observations	672	672	672	672

Note: the figures reported are estimates of neighbors' weights, based on all updating rounds. Dependent variable: current belief of agent at the node reported in column heading, rounds 2-8. All specifications also include full sets of session and phase dummies. Standard errors clustered at independent group level reported in brackets. ***p< 0.01, **p< 0.05, *p< 0.10.

they do so to a greater extent than would be optimal.

2.5 Conclusions

Although the mathematical concept of *digraph*, i.e., a network based on *directed* relations, was already central in the pioneering works of French (1956) and Harary (1959), empirical studies of information diffusion in social networks have generally not focused explicitly on the respective roles played by *indegree* and *outdegree*. This may be partly reflecting the fact that, although *asymmetric* information flows are the norm in opinion formation, most of the available network data sets (such as those describing friendship relations on online social networks, co-authorships of academic authors, or traffic flows) generally describe *undirected* networks. Recently, however, increasing attention has been given, both theoretically and empirically, to information flows in *directed* networks (e.g. Baños et al., 2013 and Gleeson et al., 2013).

This paper investigated a boundedly rational model of opinion formation in directed social networks that provides a simple generalization of the linear updating rules in DeGroot (1974) and DeMarzo et al. (2003). In the model, agents aggregate the information they receive from their neighbors' by using weights that may reflect their neighbors' indegree. Intuitively, when opinions are updated, relatively more importance can be attributed to more informed

individuals.

At the theoretical level, our results indicate that, in balanced networks, placing higher weight on neighbors with higher indegree is generally less efficient than placing equal weights on all neighbors. On the other hand, in unbalanced networks, it can be efficient to place higher weight on neighbors with higher indegree. Indeed, there exist unbalanced networks in which the optimal importance attributed to indegree is arbitrarily high. At the empirical level, our experimental results provide clean evidence of a causal effect of indegree on social influence. Both Bayesian updating and boundedly rational updating à la DeMarzo et al. (2003) are rejected against the alternative of a boundedly rational updating rule in which the weight placed on an agent's opinion is positively related to the number of individuals she listens to. Indeed, the importance that agents place on their neighbors' indegree is higher than would be efficient.

One possible interpretation of our findings is that agents are aware that, in the setting considered, placing a higher weight on neighbors with a higher indegree is efficient. However, in their attempt to aggregate information efficiently while retaining a simple updating rule, agents end up placing excessive weight on neighbors with high indegree. A second possible explanation is that, irrespective of any efficiency motivation, agents tend to attribute some form of authority to peers whom they perceive as better informed, and this leads them to place a higher weight on the information received from them. Another possible interpretation of our results is that the weights of the updating rule could be state-dependent. In the framework by Hegselmann and Krause (2002), for example, updating weights depend negatively on the distance between opinions. Since the beliefs of high-indegree agents are, on average, less extreme than those of low-indegree agents, they can be expected to be more similar, on average, to the beliefs of the listening agents. In this perspective, our results could be interpreted as reflecting features of beliefs, so that network structure, and more specifically indegree, would play a role only indirectly. Indegree is not necessarily the only characteristic of the local network which can influence the process of opinion formation: it is however a parsimonious statistics that can reasonably appear in a "rule of thumb" updating rule and which can be verified experimentally, on various types of networks. Moreover, in our experiment it is manipulated independently from the outdegree and the eigencentrality.

To sum up, our analysis provides causal evidence of an indegree effect that is at odds with the updating mechanisms most commonly adopted in the literature on opinion dynamics. When forming their opinion, agents do not place equal weights on all their neighbors, but use weights that are positively related to their neighbors' indegree. As a result, ceteris paribus,

subjects with higher indegree ultimately have higher social influence. This is an important finding, as it indicates that, despite their inability to fully account for the structure of their communication network, agents are able to exploit the information about its local properties. Further research should contribute to an understanding of the *mechanisms* explaining the effect of indegree on opinion formation and social influence.

Appendix 2.A Proofs

Proof of Theorem 1. Consider a strongly connected network $\mathcal{G}$: its adjacency matrix $M_{\mathcal{G}} = q_{ij}{}_{i,j\leq n}$ is necessarily *irreducible*.[17] Perfect and Mirsky (1965) have shown[18] that then there exists another matrix $P_{\mathcal{G}}$, with coefficients p_{ij} such that

- $P_{\mathcal{G}}$ is *doubly* stochastic.[19]
- $p_{ij} = 0 \iff q_{ij} = 0$.

The coefficients p_{ij} define a new linear updating rule applicable to the network $\mathcal{G}$.[20] Let us calculate the dynamics of the *average* of beliefs from one period to another according to this new updating rule:

$$\begin{aligned}\frac{1}{n}\sum_i y_i^{t+1} =& \frac{1}{n}\sum_i\sum_j p_{ij}y_j^t \\ =& \frac{1}{n}\sum_j \underbrace{\sum_i p_{ij}}_{=1} y_j^t \\ =& \frac{1}{n}\sum_j y_j^t.\end{aligned}$$

Such average is unchanged. By iterating this reasoning, we have that the average of opinions *at any time* is equal to the initial mean $\bar{x}$. When a consensus is reached, it is by definition the correct consensus. □

[17] An $n \times n$ matrix q is reducible if the set $\{1, \ldots, n\}$ can be partitioned in two subsets V_1, V_2 such that $q_{ij} = 0$ whenever $i \in V_1$ and $j \in V_2$. This implies that nodes in V_1 are not connected to nodes in V_2, and hence that the network is not strongly connected.

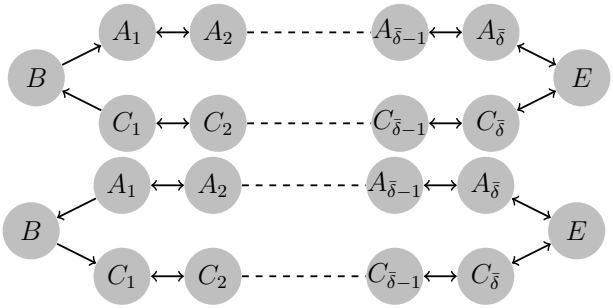
[18] It is easy to see that if a network structure is strongly connected, statement (ii) in their Theorem 1 holds. The matrix representation which they denote by $(*)$ implies that there are k nodes with no links to some other $n-k$ nodes. But since in our network each agent is by assumption linked to herself, those two sets of nodes would have to be disjoint, and partition the whole network. Hence, there would be no path from the first to the second.

[19] A matrix is doubly stochastic if all its elements are non negative, and each row and column sums up to 1.

[20] The relation between the double stochasticity of the updating matrix and the correctness of the consensus was already recognized by Harary (1959). The rest of the proof simply generalizes his Theorem 14 to a generic linear rule.

Proof of Theorem 2. Given a network $\mathcal{G}$, let $\bar{\mathcal{G}}$ be the corresponding *undirected* network (with adjacency matrix $\bar{q}_{ij} = 1 \iff q_{ij} + q_{ji} > 0$), and ${\mathcal{G}_i}^{\delta}$ the subnetwork of $\mathcal{G}$ restricted to nodes which are distant at most $\delta \in \mathbb{N}$ from i in $\bar{\mathcal{G}}$. Since the weights $\bar{\pi}_{ij}$ only depend on local properties of the network, there must exist a $\bar{\delta} \in \mathbb{N}$ such that they are only determined based on $\mathcal{G}_i^{\bar{\delta}}$. Consider then the networks in figure 2.7:

Figure 2.7: Locally similar networks for node E



Notice that $\mathcal{G}_E^{\bar{\delta}}$ is identical in both networks, and hence the vector of weights $\bar{\pi}_E$ adopted by E must be identical too. Assume, without loss of generality, that $\bar{\pi}_{EA_{\bar{\delta}}} \geq \bar{\pi}_{EC_{\bar{\delta}}}$. In the second network, E gets to know the opinion of agents $B, C_1, \ldots, C_\delta$ *only* through the link coming from C_δ. Hence, she is weighting the opinions of those $\bar{\delta} + 1$ nodes *less* than the opinions of the other $\bar{\delta}$ nodes $A_1, \ldots, A_{\bar{\delta}}$. This rule *cannot* lead to the correct consensus on such network. □

Example 1. *Consider a network $\mathcal{G}$, and (recalling that agents are numbered from 1 to n) execute the following steps:*

1. *start from agent 1: since the network is strongly connected, there must be a path from 1 to 2: call it s_1, and assume without loss of generality that it has no cycles;*

2. *again, since the network is strongly connected, there must be a path from 2 to 3: assume without loss of generality that it has no cycles, and call s_2 the union of s_1 with such a path;*

3. *by repeating the step above, for each $i < n$, a path s_i is constructed, which goes from 1 to i and passes through any $i' < i$: let s_n be the union of s_{n-1} with a path (again, without cycles) from n to 1: s_n is a cycle which passes through each node at least once and at most n times;*

4. *for each pair* (i, j) *with* $j \neq i$*, define* π_{ij} *as* $\frac{1}{n}$ *multiplied by the number of times that* s_n *passes through the link from* j *to* i *(possibly zero). For each* i*, define* π_{ii} *as* $1 - \sum_{j \neq i} \pi_{ji}$*.*

The updating rule having such π_{ij} *as updating weights is a valid linear rule. If we consider* s_n *as a weighted network where the weight of a link is given by* π_{ij}*, it is strongly connected, and its adjacency matrix is doubly stochastic. Hence, the resulting updating rule leads to the correct consensus.*

Proof of Lemma 2. Consider the influence weights for the opinion of *a given* agent as a vector in the standard $n-1$-simplex, ω_{ij}^t. For instance,

$$\omega_i^0 = (\underbrace{0, \ldots, 0}_{i-1}, 1, \underbrace{0, \ldots, 0}_{n-i})$$

that is, before receiving any information from neighbors, the opinion of each agent is entirely formed by her initial signal. With a linear belief updating rule, the evolution of such vectors is simply described as a straightforward rewriting of (2.1):

$$\begin{aligned} \omega_i^t &= \sum_{j=1}^{n} q_{ij} \pi_{ij} \omega_j^{t-1} \\ &= \pi_{ii} \omega_i^{t-1} + \sum_{j \neq i} q_{ij} \pi_{ij} \omega_j^{t-1}. \end{aligned}$$

Since the network is anonymous, all neighbors of a given node are equivalent; and since the rule is anonymous, each node places an equal weight on each neighbor. So the above can be rewritten as

$$\omega_i^t = \pi_{ii} \omega_i^{t-1} + \frac{(1 - \pi_{ii})}{d_i} \sum_{j \neq i} q_{ij} \omega_j^{t-1}, \tag{2.A.1}$$

where d_i is agent i's indegree. Notice that $y_i^t = \omega_i^t \cdot x$, and hence the beliefs are uniquely determined once π_{ii}^t is fixed for each i.

For what concerns point 2, let us define the system as *biased* at a given time $\bar{t}$ if for some h

$$\sum_{i=1}^{n} \omega_{ih}^{\bar{t}} \neq 1.$$

Let us assume without loss of generality that $\bar{t}$ is the first time for which this happens. Notice that $\bar{t} > 0$, since

$$\sum_{i=1}^{n} \omega_{ih}^0 = \omega_{hh}^0 = 1,$$

and that

$$\sum_{i=1}^{n} \omega_{ih}^{\bar{t}} = \sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij} \pi_{ij} \omega_{jh}^{\bar{t}-1}$$
$$= \sum_{j=1}^{n} \omega_{jh}^{\bar{t}-1} \sum_{i=1}^{n} q_{ij} \pi_{ij}.$$

Since the network and the rule are anonymous, the value of the nested sum must be the same for all j, and since the sum of $q_{ij}\pi_{ij}$ across j *and* i is n, the value of such nested sum must be $\frac{n}{n} = 1$. So

$$\sum_{j=1}^{n} \omega_{jh}^{\bar{t}-1} = \sum_{i=1}^{n} \omega_{ih}^{\bar{t}} \neq 1.$$

But this contradicts the hypothesis that $\bar{t}$ is the first time at which the system is biased. □

Proof of Theorem 3. One implication is obvious: given any agent i, let $\bar{d}$ be the (equal, by assumption) indegree of all $d_i + 1$ agents in $S(i)$. Then, by applying Equation (2.3), we have that for each $j \in S(i)$,

$$\pi_{ij} = \frac{\bar{d}^{\rho}}{(d_i + 1)\bar{d}^{\rho}} = \frac{1}{d_i + 1},$$

that is, the updating weights do not depend on the value of ρ - hence the consensus belief and social influence weights do not either. For the reverse implication, notice that if a network $\mathcal{G}$ is not regular, there is at least a pair of agents i, j with different indegree and such that i listens to j (if this is not the case, it is easily to show by induction that the network is regular). Define now as H_k the set of agents h such that there exist $j', j'' \in S(h)$ with $k = d_{j'} < d_{j''}$. Intuitively, we are considering all agents i on whose updating weights ρ does matter, because the neighbors have different indegree, and classifying them based on the *lowest* indegree of a neighbor: the underlying idea of the remaining of the proof is that this will allow us to identify an agent who is necessarily *disadvantaged*, in terms of social influence, by a strictly positive value of ρ. Let $\bar{k}$ be the smallest k such that H_k is non-empty - the non-regularity assumption means precisely that there exists at least one such k. Let $\bar{i} \in H_{\bar{k}}$, and $\bar{j} \in S(i)$ such that $d_{\bar{j}} = \bar{k}$. Notice that, for each $i' \in S^{-1}(\bar{j})$, we have that $\bar{j}$ *must* have smaller or equal indegree than all other agents in $S(i')$ (otherwise, we would have found a non-empty $H_{k'}$ with

$k' < \bar{k}$). As a consequence, $\pi_{i'\bar{j}}$ will be *weakly smaller* with $\rho > 0$ than with $\rho = 0$, and $\pi_{\bar{i}\bar{j}}$ will be *strictly* smaller. But since we know that

$$w_{\bar{j}} = \sum_{i=1}^{n} \pi_{i\bar{j}} w_i,$$

this means that $w_{\bar{j}}$ will strictly decrease *as a function of its neighbors*, and hence that at least *some* w_i will be affected by a change of ρ. □

Lemma 3. *In a balanced network, when the GBR rule is applied with $\rho = 0$, the social influence weight of each agent is proportional to her degree (including the self-link) $d_i + 1$.*[21]

Proof of Lemma 3. Recall from Equation (2.7) that the vector of social influence weights is the unique element w in $\mathbb{R}^n$ satisfying $\sum_{i=1}^{n} w_i = 1$ and

$$w_i = \sum_{j \in N^{-1}(i)} \pi_{ji} w_j \tag{2.A.2}$$

for each i; when $\rho = 0$, the above translates to

$$w_i = \sum_{j \in N^{-1}(i)} \frac{1}{d_j + 1} w_j. \tag{2.A.3}$$

Assuming the network is balanced, if the social influence weights can be written as $w_i = \alpha(d_i + 1)$, where α is a constant, then the right hand side of Equation (2.A.3) becomes

$$\sum_{j \in N^{-1}(i)} \frac{1}{d_j + 1} \cdot \alpha(d_j + 1) = |N^{-1}(i)| \cdot \alpha = \alpha(d_i + 1) = w_i$$

that is, Equation (2.A.2) is satisfied for each i. To guarantee that the sum of social influence weights adds up to 1, it is sufficient to set α accordingly:

$$\frac{1}{\alpha} = \sum_{i=1}^{n} d_i + n = \sum_{i=1}^{n} (d_i + 1)$$

(notice that such α is the inverse of the total number of links). □

[21] This result was already proved by DeMarzo et al. (2003) (as part of their Theorem 6) for the specific case of undirected networks.

Lemma 4. *Given a network $\mathcal{G}$, let $\mathcal{E}_{\mathcal{G}} : \mathbb{R} \to [-1, 0]$ be the function mapping each real number to the efficiency of the GBR rule played with such value of ρ. If $\mathcal{G}$ is balanced, $\frac{\partial \mathcal{E}_{\mathcal{G}}}{\partial \rho}(0) < 0$.*

Proof of Lemma 4. Lemma 3 tells us that on a balanced network, when $\rho = 0$, influence weights are an increasing function of degree. Now consider a generalization of the GBR model in which *each node* i adopts a different ρ_i, and let $\mathcal{E}^*_{\mathcal{G}} : \mathbb{R}^n \to [-1, 0]$ be the generalization of $\mathcal{E}_{\mathcal{G}}$. We want to study $\mathcal{E}^*_{\mathcal{G}}$ around $(0, \ldots, 0)$. For small positive variations of a single $\rho_{\bar{i}}$, the elements of $S^{-1}(\bar{i})$ will be affected by a change in their social influence weight which is an increasing function of degree (and hence of social influence weights themselves), so that the sum of (squares of) absolute differences from the mean, $\frac{1}{n}$, will increase. Now, it may be that the *indirect* changes (i.e. on agents connected to agents in $S^{-1}(i)$) affect social influence weights in a way that is *not* an increasing function of degree. However, it is easy to see that the entity of such indirect changes will be *smaller*, in absolute value, than the direct change they originate from. So the same will hold for the squares of such differences, and as a result

$$\frac{\partial \mathcal{E}^*_{\mathcal{G}}}{\partial \rho_{\bar{i}}}(0, \ldots, 0) < 0.$$

Now, the derivative of $\mathcal{E}_{\mathcal{G}}$ in ρ coincides with the directional derivative of $\mathcal{E}^*_{\mathcal{G}}$ along the vector $(1, \ldots, 1)$. So it is also negative. □

Proof of Theorem 4. It is easy to verify that $\mathcal{E}_{\mathcal{G}}$ is smooth. Assume then that there exists some $\rho^* > 0$ such that $\mathcal{E}_{\mathcal{G}}(\rho^*) > \mathcal{E}_{\mathcal{G}}(0)$: since (by Lemma 4) $\mathcal{E}'_{\mathcal{G}}(0) < 0$, it must be that $\mathcal{E}_{\mathcal{G}}$ has a local minimum $\rho_* \in [0, \rho^*]$. However, since $\rho_* > 0$, the derivative in ρ of a given term $(w_i - \frac{1}{n})^2$ is larger the larger w_i. So if the linear combination of such derivatives is 0 in ρ_*, it must be positive in a right neighborhood of ρ_*, that is, $\mathcal{E}''_{\mathcal{G}}(\rho_*) < 0$. So ρ_* is *not* a local minimum. □

Proof of Theorem 5. Given a natural number K, consider a network having the following binary tree-like structure:

- an agent A_1 listens to two other agents $A_{2,1}$, $A_{2,2}$,
- each agent $A_{k,i}$ listens to two other agents $A_{k+1,2i-1}, A_{k+1,2i}$, for each $k < K$,

- each "leaf" agent $A_{K,i}$ listens to A_1 and to her "close relatives" $A_{K,i-1}$ and $A_{K,i+1}$ ($A_{K,1}$ listens to $A_{K,2^K}$ and $A_{K,2}$, while $A_{K,2^K}$, listens to $A_{K,2^K-1}$ and $A_{K,1}$).

Notice that,

- the structure is perfectly symmetric, in the sense that all the agents positioned on a given "layer" will exhibit the same vector of updating weights (which we will hence denote for simplicity as $\pi_{k-1,k}$ and $\pi_{k,k}$ rather than $\pi_{A_{k-1,i}A_{k,j}}$ and $\pi_{A_{k,i}A_{k,i}}$, respectively) and the same social influence (which we will hence denote as w_k rather than $w_{A_{k,i}}$),

- for most of the layers of this structure, the updating weights are independent from ρ; namely, for any k such that $1 \leq k < K-1$,

$$\pi_{k,k} = \pi_{k-1,k} = \frac{2^\rho}{3 \cdot 2^\rho} = \frac{1}{3}$$

and hence

$$\begin{aligned} w_k =& \pi_{k,k} w_k + \pi_{k-1,k} w_{k-1} \\ =& \frac{1}{3} w_k + \frac{1}{3} w_{k-1} \\ =& \frac{1}{2} w_{k-1}; \end{aligned}$$

- since the number of agents on a given layer *doubles* at each level, the sum of social influences of all agents in a given layer, which we will denote as W_k, is the same for any k from 1 to $K-1$.

Now consider the social weight of A_1. The updating weights of a leaf (which has indegree 3, rather than 2) can be calculated as:

$$\pi_{K,1} = \frac{2^\rho}{2^\rho + 3^\rho + 3^\rho + 3^\rho} = \frac{2^\rho}{2^\rho + 3^{\rho+1}}; \qquad \pi_{K,K} = \frac{3^\rho}{2^\rho + 3^{\rho+1}}.$$

and since 2^K leafs listen to A_1,

$$\begin{aligned} w_1 =& \pi_{1,1} w_1 + 2^K \pi_{K,1} w_K \\ =& \frac{1}{3} w_1 + 2^K \frac{2^\rho}{2^\rho + 3^{\rho+1}} w_K \\ =& 2^K \frac{3 \cdot 2^{\rho-1}}{2^\rho + 3^{\rho+1}} w_K \\ \implies W_1 =& \frac{3 \cdot 2^{\rho-1}}{2^\rho + 3^{\rho+1}} W_K. \end{aligned}$$

Assume the optimal level of ρ is bounded above by some $\hat{\rho}$. This means that for $K \to \infty$, this last ratio will also be bounded. That is, asymptotically,

$$W_K \sim W_1 = W_2 = \cdots = W_{K-1}$$

and hence, since the sum of all w_i is 1, $w_1 = W_1$ will converge to 0 asymptotically as $\frac{\alpha}{K}$, where α is a constant. Now, it easy instead to verify that since, for given $\bar{K}$,

$$2^{\bar{K}} \frac{3 \cdot 2^{\rho-1}}{2^{\rho} + 3^{\rho+1}} \overset{\rho\to\infty}{\to} 0$$

and such term is continuous in ρ, we can define $\rho_{\bar{K}}$ such that

$$2^{\bar{K}} \frac{3 \cdot 2^{\rho_{\bar{K}}-1}}{2^{\rho_{\bar{K}}} + 3^{\rho_{\bar{K}+1}}} = 1;$$

moreover, it is straightforward to verify that $\rho_{\bar{K}} \overset{\bar{K}\to\infty}{\to} \infty$. When the GBR rule is applied with such ρ_K, we have, by definition, that $w_1 = w_K$. That is,

$$\frac{W_K}{2^K} = W_1 = W_1 = W_2 = \cdots = W_{K-1}.$$

The sum of the weights still sums up to 1, but now the nodes with the maximum influence are 2^K+1 (all leafs, and A_1) so now each influence weight will converge to 0 as $\frac{1}{2^K}$ (or faster), rather than as $\frac{\alpha}{K}$.

Now, observe that the *correct* weights converge to 0 as $\frac{1}{n} = \frac{1}{2^{K+1}-1}$. Hence, the sum of square deviations in the case of any finite ρ will converge towards at least $\left(\frac{\alpha}{K}\right)^2 = \frac{\alpha^2}{K^2}$, while in the case of $\rho = \rho_K$ it will converge towards at most

$$2^{K+1} \cdot \left(\frac{1}{2^K}\right)^2 = \frac{2^{K+1}}{2^{2K}} = \frac{1}{2^{K-1}} \overset{K\to\infty}{<} \frac{\alpha^2}{K^2}.$$

Hence the most efficient ρ for $K \to \infty$ *must* also tend to ∞. □

Theorem 6. *Consider two strongly connected networks $\mathcal{G}_1$, $\mathcal{G}_2$, with adjacency matrices q^1, q^2, identical except for an element:*

$$0 = q^1_{\bar{i},\bar{j}} \neq q^2_{\bar{i},\bar{j}} = 1.$$

Let w^1 and w^2 be the vectors of social influence weights resulting from the GBR updating rule implemented with $\rho = 0$ on $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively. Then, $w^2_{\bar{j}} > w^1_{\bar{j}}$.

Let us define $\Delta_i = \frac{w_i^2 - w_i^1}{w_i^1}$, the (relative) *increase in the social influence of an agent i when adding the link from $\bar{j}$ to $\bar{i}$*: Theorem 6 simply states that $\Delta_{\bar{j}}$ is positive. In order to prove it, we prove the following stronger result.

Lemma 5. *Consider $\mathcal{G}_1$, $\mathcal{G}_2$, w^1, w^2, as above. Then, $\Delta_{\bar{j}} \geq \Delta_i$ for all $i \neq \bar{j}$.*

Proof of Lemma 5. Let π_{ij}^1 and π_{ij}^2 be the weights attributed by i to j in $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively, and let $S_1(i)$, $S_2(i)$ be the listening sets of i in the two networks. Notice that, whenever $i \neq \bar{i}$, then π_{ij}^1 and π_{ij}^2 coincide. Equation (2.7) allows us to express the social influence of an agent as a linear combination of the social influence weights of the agents she talks to. By plugging it in the definition of Δ_i, the same can be done for what concerns the *relative change* of influence; for any $i \neq \bar{j}$, the *listeners set*

$$S^{-1}(i) = \{j : i \in S(j)\}$$

is unchanged, and so we have

$$\Delta_i = \frac{w_i^2 - w_i^1}{w_i^1} = \sum_{j \in S_1^{-1}(i)} \frac{\pi_{ji}^2 w_j^2 - \pi_{ji}^1 w_j^1}{w_i^1}.$$

Let us define $\hat{\Delta}_i$ as:

$$\hat{\Delta}_i = \sum_{j \in S_1^{-1}(i)} \frac{\pi_{ji}^1 (w_j^2 - w_j^1)}{w_i^1} = \sum_{j \in S^{-1}(i)} \frac{\pi_{ji}^1 w_j^1 \Delta_j}{w_i^1}.$$

We can observe that:

- $\Delta_{\bar{j}} > \hat{\Delta}_{\bar{j}}$, since $S_2^{-1}(\bar{j}) = S_1^{-1}(\bar{j}) \cup \{\bar{i}\}$, while $\pi_{i\bar{j}}^1 = \pi_{i\bar{j}}^2$ for all $i \neq \bar{i}$,
- $\Delta_j < \hat{\Delta}_j$ for any other $j \in S_1^{-1}(\bar{i})$, since $\pi_{\bar{i}j}^1 < \pi_{\bar{i}j}^2$, while $\pi_{ij}^1 = \pi_{ij}^2$ for all $i \neq \bar{i}$,
- $\Delta_j = \hat{\Delta}_j$ for any $j \notin S_2^{-1}(\bar{i})$.

Now, assume Lemma 5 is false. That is, there exists $j' \neq \bar{j}$ such that $\Delta_{j'} > \Delta_{\bar{j}}$. Let us assume, without loss of generality, that $\Delta_{j'} \geq \Delta_i$ for all i.

We can then write:

$$\begin{aligned}\Delta_{j'} \leq \hat{\Delta}_{j'} &= \sum_{i \in S^{-1}(j')} \frac{\pi_{ij'} w_i^1 \Delta_i}{w_{j'}^1} \\ &\leq \sum_{i \in S^{-1}(j')} \frac{\pi_{ij'} w_i^1 \Delta_{j'}}{w_{j'}^1} \\ &= \frac{\Delta_{j'}}{w_{j'}^1} \sum_{i \in S^{-1}(j')} \pi_{ij'} w_i^1 \\ &= \Delta_{j'} \frac{w_{j'}^1}{w_{j'}^1} = \Delta_{j'}.\end{aligned}$$

We can see that both inequalities must be binding. The first implies that $j' \notin S_2^{-1}(\bar{i})$, and hence $j' \notin S_1^{-1}(\bar{i})$; the second that $\Delta_i = \Delta_{j'}$ for all $i \in S^{-1}(j')$. By applying the same process recursively to any such i, and exploiting the strong connectedness of the network, we can show that $\Delta_{\bar{j}} = \Delta_{j'}$, which contradicts the initial assumption.[22] □

Proof of Theorem 6. The sum of social influence weights is by definition 1 in any network:

$$\sum_{i=1}^{n} w_i^1 = 1 = \sum_{i=1}^{n} w_i^2.$$

As a consequence, the *weighted* sum of percentage changes Δ_i must be 0:

$$\sum_{i=1}^{n} w_i^1 \Delta_i = 0$$

(with all weights w_i^1 strictly positive). Since $\hat{\Delta}_i$ is a linear combination of Δ_j for different j, if we had $\Delta_j = 0$ for all j, then we would have $\hat{\Delta}_i = 0$ for all i. Instead we know that $\Delta_j < \hat{\Delta}_j$ for some j. So the maximum Δ_i, which is guaranteed by Lemma 5 to be $\Delta_{\bar{j}}$, must be strictly positive. □

Notice that $\Delta_{\bar{i}}$ is not guaranteed to be negative. The influence of $\bar{i}$ will decrease in *relative* terms (that is, compared to $\hat{\Delta}_{\bar{i}}$), but the increase in influence of $\bar{j}$ may more than compensate this effect if there is a short path from $\bar{i}$ back to $\bar{j}$.

[22] If the chosen path from j' to $\bar{j}$ passes through $S_1^{-1}(\bar{i})$, the contradiction will arise even *before* reaching $\bar{j}$.

Similarly, $\Delta_{\bar{j}}$ is not guaranteed to be positive if $\rho > 0$. Again, social influence will increase in relative terms, but for sufficiently high values of ρ the increase in influence of $\bar{i}$ will be large enough to make the influence weights of all other agents decrease.

Appendix 2.B Experimental Instructions

[Translated from Italian]

Welcome and thank you for taking part in this experiment. During the experiment talking or communicating with other participants is not allowed in any way. If you have a question at any time, raise your hand and one of the assistants will come to answer your question. By carefully following the instructions you can earn a sum of money that will depend on the choices made by you and the other participants. On top of that amount, you will receive in any case 5 € for the participation in this experiment.

General Rules

- 24 subjects will take part in this experiment.
- The experiment takes place in 4 phases of 8 rounds each, for a total of 32 rounds.
- At the beginning of the experiment 6 groups of four subjects will be randomly and anonymously formed by the computer.
- You will be assigned to one of the 6 groups. You will interact only with those in your group, without knowing their identity. The composition of each group will remain unchanged throughout the experiment.

The development of a phase

- In the first round of each of the four phases, in all groups, each subject will be randomly and anonymously assigned a different role: A, B, C, and D.
- The computer will randomly generate four integers that we will refer to as *signals*. Each component of the group will be shown only one of the four signals. Signals will be denoted as x_A, x_B, x_C, and x_D.
- In each of the 8 periods of the phase, each subject will be asked to guess the mean of the four signals extracted by the computer for that phase: $\bar{x} = \frac{(x_A+x_B+x_C+x_D)}{4}$.
- For making each guess, there is a maximum time of 120 seconds (which will be shown by a counter in the top right corner of the screen).
- At any moment, it is possible to open a calculator by simply clicking its icon, in the bottom left corner of the screen.

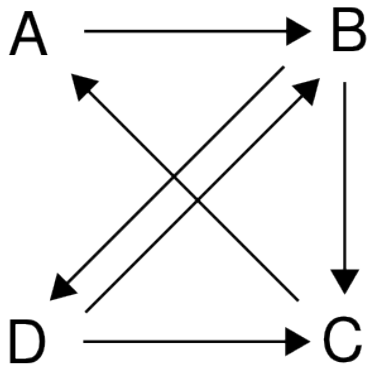
How earnings are determined

- Individual earnings will depend on how close the guess comes to the value of $\overline{x}$:
  - At the end of the experiment, the computer will randomly extract one of the 32 periods.
  - The earnings will be equal to 15 euros minus the difference (in absolute value) between $\overline{x}$ and the guess made in the selected round.
  - If this difference turns out to be negative, the subject will earn 0 euros.
- Examples:
  - if $\overline{x} = 1424$ and the guess is 1424, the difference is 0 and earnings are 15 euros.
  - if $\overline{x} = 308$ and the guess is 311.5, the difference is 3.5 and earnings are 11.5 euros.
  - if $\overline{x} = 803.25$ and the guess is 792, the difference is 11.25 and earnings are 3.75 euros.
  - if $\overline{x} = 62.5$ and the guess is 30.5, the difference is 32 and earnings are 0 euros, since $15 - 32 < 0$.

  In each of this cases, the participant will also receive 5 euros for participating in the experiment.
- In each round, the optimal guess (which allows to get the maximum earnings) depends on the information that each subject has on the signals:
  - if she knows only her own signal, the optimal choice is her own signal,
  - if she knows or can infer two signals, her optimal choice is the mean of the two signals;
  - if she knows or can infer three signals, her optimal choice is the mean of the three signals;
  - if she knows or can infer four signals, her optimal choice is the mean of the four signals.

Information

- In each of the tree phases
  - In the first round, each subject knows his own signal.
  - From the second round onwards, before making his choice, each subject will be informed by the computer of the choices made in the previous rounds by some of the components of his group, based on the structure represented in the following figure:



- Therefore, before making his choice
  - A will be informed of the choices made by C.
  - B will be informed of the choices made by D.
  - C will be informed of the choices made by A and B.
  - D will be informed of the choices made by A and C.
- The roles (A, B, C, D), the signals (x_A, x_B, x_C, x_D) and by consequence their mean will change at each phase: the computer will generate them randomly before the first period of the phase.

Feedback e payments

- At the end of each phase the computer will show to each subject the four signals of his group, their mean, and the choices made.
- At the end of the experiment each subject will be shown the round the computer has selected to determine payments, the value of x for his group, the choice she made and the corresponding amount earned in euro.

- The experiment will terminate and the amount earned by each subject will be paid in cash.

Control questions

1. If you knew only your signal (a), what would be your optimal guess? .....
2. If you knew your signal (a) as well as the one of another member of your, group (b), what would be your optimal guess? .....
3. If you knew your signal (a) as well as the ones of two other members of your group (b and c), what would be your optimal guess? .....
4. If you knew your signal (a) as well as the ones of three other members of your group (b, c, and d), what would be your optimal guess? .....

Appendix 2.C Additional material

2.C.1 Complete estimations results

Table 2.7 reports the OLS estimates for all regressors featured in Equation (2.10), under the assumption that $\sum_j w_j = 1$, and that coefficients for dummies in each category sum up to 0. There is no obvious interpretation for the significance of the coefficient for the first session, since the only change with respect to other sessions was the spatial disposition of the network. Moreover, the estimate of such coefficient seems to be strongly affected by outliers, and while the coefficient is large in absolute size, its explanatory variable is lower than the one of signals, which have a much larger variability. Notice the estimate for μ is mostly non-significant - that is, there appears to be no overall systematic bias in subjects guesses. Still, its negative sign and the fact that the coefficients for phases appear to be increasing (although not significantly so) may suggest the presence of a learning effect due to repeating the same task multiple times. This hypothesis can be verified by regressing the *absolute difference* of the (final) guesses from true mean over the phases dummies. The result is shown in Table 2.8: although coefficients for phases may suggest a downwards trend in variance, it is once again not significant. While this may seem counter-intuitive, since the task is repeated four times, it must be recalled that subjects are assigned each position only once, so this may limit their learning ability.

Table 2.7: Estimated social influence weights, overall

	0 %	1 %	5 %	10 %
Signal A	0.268***	0.263***	0.269***	0.255***
	(0.031)	(0.031)	(0.032)	(0.027)
Signal B	0.294***	0.299***	0.301***	0.303***
	(0.013)	(0.014)	(0.015)	(0.014)
Signal C	0.214***	0.213***	0.205***	0.205***
	(0.017)	(0.017)	(0.018)	(0.017)
Signal D	0.224***	0.224***	0.226***	0.238***
	(0.023)	(0.023)	(0.023)	(0.017)
nsession==1	-10.930***	-10.700***	-11.867***	-5.544**
	(3.823)	(3.749)	(3.590)	(2.562)
nsession==2	4.631	5.036	3.748	-0.454
	(4.212)	(4.133)	(4.224)	(3.946)
nsession==3	-0.982	-2.035	-0.041	-1.357
	(4.699)	(4.195)	(3.590)	(3.680)
nsession==4	7.281	7.699	8.159	7.355
	(4.971)	(4.966)	(5.247)	(5.110)
phase==1	-10.770*	-12.008*	-10.323	-3.836
	(5.908)	(6.164)	(6.336)	(5.145)
phase==2	0.190	0.563	-0.463	-2.910
	(4.209)	(4.311)	(4.288)	(3.925)
phase==3	8.130*	8.542**	8.924**	6.857**
	(4.195)	(4.160)	(3.973)	(3.377)
phase==4	2.450	2.903	1.862	-0.112
	(3.953)	(3.982)	(4.074)	(4.013)
μ	-4.165	-4.509*	-3.606	-1.419
	(2.574)	(2.468)	(2.358)	(2.025)
Number of observations	384.00	380.00	364.00	344.00

Note: First column: extended version of column (1) of Table 2.3; other columns: sample restricted by eliminating groups with most extreme deviations from optimal beliefs (1%, 5%, 10%, respectively).

Table 2.8: Decomposition of deviation from true mean

	0 %	1 %	5 %	10 %
Signal A	0.002	0.004	-0.008	0.023
	(0.033)	(0.033)	(0.029)	(0.019)
Signal B	0.017	0.014	0.015	0.004
	(0.015)	(0.014)	(0.014)	(0.013)
Signal C	-0.040*	-0.039*	-0.027	-0.033**
	(0.021)	(0.021)	(0.017)	(0.014)
Signal D	0.021	0.021	0.020	0.007
	(0.022)	(0.022)	(0.020)	(0.014)
phase==1	9.625*	10.138**	9.411*	4.506
	(5.005)	(5.114)	(5.239)	(2.892)
phase==2	1.975	1.822	2.812	3.080
	(3.953)	(3.921)	(3.696)	(3.005)
phase==3	-7.007**	-7.177**	-8.491***	-6.121**
	(3.299)	(3.348)	(3.085)	(3.082)
phase==4	-4.593	-4.783	-3.732	-1.465
	(3.361)	(3.401)	(3.638)	(3.300)
Constant	21.892***	22.032***	21.237***	18.413***
	(2.893)	(2.920)	(2.842)	(2.695)
Number of observations	384.00	380.00	364.00	344.00

Note: Dependent variable: absolute distance from true mean (at period 8). First column: all observations; other columns: sample restricted by eliminating groups with most extreme deviations from optimal beliefs (1%, 5%, 10%, respectively).

2.C.2 Per-period parameter calibration

The analysis in Section 2.4 was mainly developed with regards to the beliefs expressed by subjects in the final period, considered as an approximation of the consensus belief after convergence. In principle, the same analysis could be employed instead to analyze beliefs from each period: in Table 2.6, we have already reported the estimates of updating weights, based on data from all rounds.

Figure 2.8: Efficiency and fit with data of the generalized model, for all periods.

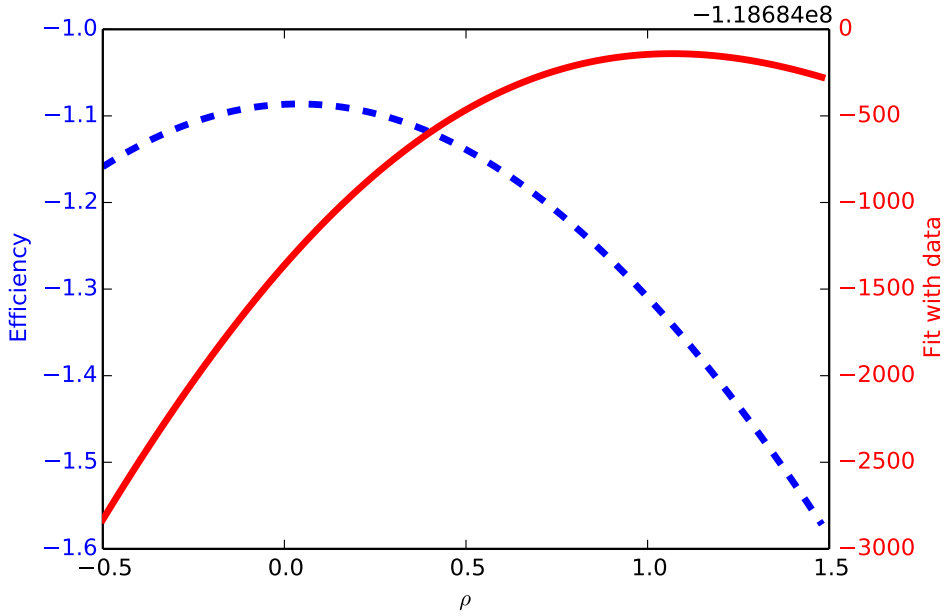


Figure 2.8 compares the fit of the model as a function of ρ (dashed line, left scale), as maximized in Equation (2.11), with its theoretical efficiency (solid line, right scale), as defined in Equation (2.9). This time, however, both are computed over all the 8 rounds of data (respectively, over 8 periods of simulation). The two scales are not directly comparable. Still, the figure makes it clear that also in this formulation of the model, the $\hat{\rho}$ which best agrees with observed data (peak on the right of the figure) is not only substantially larger than 0, but also substantially larger than the theoretical most efficient ρ^* (peak on the left). This at least qualitatively confirms the results of the convergence analysis.

Chapter 3

Not *That* Fundamental: Bubbles in Financial Networks

3.1 Introduction

The way in which financial risk is typically conceived in the literature has dramatically changed since the beginning of the new millennium, from a vision where the main unit of study was the *single* firm or institution, to a generalised approach including the *environment* in which such firms or institutions operate, bringing terms such as *"systemic risk"* in the common usage of researchers and policymakers.

It is true that in the description of the *"Financial Instability Hypothesis"* by Minsky (1992) one can already read *"In the modern world, analyses of financial relations and their implications for system behaviour cannot be restricted to the liability structure of businesses and the cash flows they entail. [They...] have liability structures which the current performance of the economy either validates or invalidates."* However, in more recent times this point of view has gained many supporters, for at least two reasons. The first, at least in chronological terms, is represented by the studies providing analytic models of how risk may *spread* in entangled financial and economic systems: a simplified but explanatory example is provided by Allen and Gale (2000b) (see also Allen et al., 2009 for an extensive literature review). Their work, and in general the literature on *financial contagion*, studies the way in which otherwise *sane* economic entities, be them firms or local economies, can face an increased risk of extended failures due to the interconnections among them. The present paper presents a network approach to the problem, focused on the international financial market considered as a closed system, and on the consequences, for the stability of such network, of finan-

cial speculation. The second, and probably determinant, cause is represented by the financial crisis which has started in 2007. In particular, in the light of the collapse of Lehman Brothers in September 2008, the fear for the fate of Citybank and other financial institutions, and the consequent reduction in liquidity flows all around the word, the issues of systemic risk and financial contagion were brought at the centre of the research agenda. The frequently cited "too big to fail" criterion, characterising financial entities which assume a systemic importance (and the resulting possibility of moral hasard), has been declined by many as "too *interconnected* to fail"[1], highlighting the relevance of interbank claims as channels for financial risk.

It is worth observing that in a same financial market, *contagion* can have at least two related but conceptually different forms. The term can refer to the phenomenon by which a bank going bankrupt causes other banks, linked by credit relationships, to register losses in their books, possibly causing a chain of failures (as in the works of Elliott et al., 2013, Acemoglu et al., 2013, and Allen et al., 2010), or at least a loss of value for otherwise "sane" organizations (as in the work of Cabrales et al., 2013). But it can also be related to *information* and *expectation* dynamics which can affect the market in disruptive ways by simply changing the perceived *robustness* of organisations, as in the works by Battiston et al. (2012) and Allen et al. (2012). It is intriguing to observe that the current financial crisis may have had the failure of a huge number of financial and economic actors worldwide more as a *consequence* than as a *cause*. Whatever kind of contagion is considered, the existence of a financial network offers single banks the possibility of *risk shifting* - that is assuring against idiosyncratic risk, without considering the negative externality in terms of increased *aggregate* risk (see Zawadowski (2013)). But the phenomenon of informational contagion evidences the importance that *expectations* can have on the performance of the financial market, and more precisely on the risk and extent of contagion. This is the main motivation for the present paper, which aims at modelling the effect of *self-fulfilling collective beliefs* on the aggregate dynamics (and in particular on the aggregate fragility) of the financial market, by introducing speculation in the context of a model based on the one by Elliott et al. (2013). In particular, the collective belief under study will be the one of *financial bubbles* - cases in which an asset is paid a price which is higher than its fundamental value.

The literature on financial bubbles has always faced the demanding request for coherent explanations to a phenomenon which has often been seen

[1]The expression was popularised by Ben Bernanke, in its "Financial Reform to Address Systemic Risk" 2009 speech, http://www.federalreserve.gov/newsevents/speech/bernanke20090310a.htm

as *behaviourally* motivated. While the two points of view are not exclusive (in particular because *out of equilibrium* actions, when taken into account, end up influencing optimal strategies), a stream of literature has focused on modelling the ways in which market imperfections may cause the emergence of bubbles. Feiger (1976) attributes the possibility of bubbles to the incompleteness of markets, Kreps (1977) and Harrison and Kreps (1978) to heterogeneity of expectations, Allen and Gale (2000a) claim that *"bubbles are caused by agency relationships in the banking sector"*, while Zawadowski (2013) focuses on uncertainty over fundamentals, and moral hazard: similar issues were introduced in a minimal model proposed by Allen and Gorton (1993). A specific stream of literature, however, has developed focusing on *rational* bubbles, which are compatible with the absence of market imperfections. Its basis were laid by the positive and negative results presented by Blanchard and Watson (1983), Tirole (1982), Diba and Grossman (1988c). For instance, for rational bubbles to emerge, it is sufficient that different *generations* of agents interact in the market.

Such kinds of speculative phenomena are due to self-fulfilling expectations, and can be seen as the symmetric of *bank runs*. In principle, both issues can be considered as the consequence of the interplay of psychological and rational motives, which end up causing herding phenomena.[2] In both, the *fundamentals* of an asset are assumed to matter, but the central role for explaining the dynamics of prices and actions is attributed to a phenomenon of self-fulfilling beliefs (see Diamond and Dybvig, 1983; He and Xiong, 2009 analyse in detail the relative importance of fundamentals and beliefs in a dynamic context). In line with these streams of studies, I will not claim that *real* bubbles are unaffected by uncertainty concerning fundamentals, as well as from agency problems and other market imperfections: rather, my aim is to suggest that, since bubbles are compatible with a purely rational model, they could be much more pervasive than it is usually thought. As stated by Blanchard and Watson (1983), *"It is hard to analyze rational bubbles. It would be much harder to deal with irrational bubbles."*

This study presents a model of rational bubbles in financial networks. The main technical obstacle it overcomes consists in providing a formalisation of speculative bubbles which does not require neither the assumption of irrationality of any market participants, nor the assumption of an incoming flow of wealth/agents in the economy. The basic intuition behind such formalisation is that even sophisticated market participants may want to ad-

[2]The word "herding" implies here no pejorative meaning, just as in the original vision of the stock market as a *beauty contest* proposed by Keynes (1937) in Chapter 12 of his *General Theory*. In particular, herding may not be a sign of irrationality.

here to self-fulfilling beliefs concerning the price of given assets, and deviating from the fundamental value, as long as the aggregate offer of such assets is limited - a condition which however is intrinsic in the fact that they represent *speculative* investments. A model of financial networks based on the static one by Elliott et al. (2013), but with a minimal microfoundation in order to allow for portfolios to evolve, is then presented, together with some stylised facts about the potential effect of such kind of speculation on the structure of the network, and ultimately on the resilience of the system.

The following section presents formally and extends the model of financial networks by Elliott et al. (2013). Section 3.3 then introduces the model of asset price bubbles, and discusses some of its implications. Finally, Section 3.4 sketches the possible consequences of such bubbles on the robustness of financial networks, and Section 3.5 discusses the main results.

3.2 The model of financial networks

The fundamental building block for the present model of financial networks was provided by Elliott et al. (2013). The financial market is composed by n organisations $\{1, \dots, n\}$, and m *primitive assets*[3] $\{1, \dots, m\}$, with prices respectively $p_1, \dots, p_m$. Each organisation i can hold an amount $D_{ik} \geq 0$ of each asset k, and a share $C_{ij} \geq 0$ of each other organisation j. By assumption, $C_{ii} = 0$, and $\hat{C}_{ii} := 1 - \sum_{j \in N} C_{ji}$, which is the share of organisation i held by *outside* shareholders, is positive. The matrices C with coefficients c_{ij} and D with coefficients d_{ik} define the structure of the financial market. In particular, C can be interpreted as the adjacency matrix of a weighted network: the *financial network*.

3.2.1 Values and timing

Let F be the vector of *values of fundamental assets*, that is, the total value of external assets held by each organisation: $F_i = \sum_k D_{ik} p_k$. Elliott et al. (2013) define the *equity value* V_i of organisation i recursively as the sum of F and of the total value of positions in other organisations:

$$V_i = F_i + \sum_j C_{ij} V_j$$

[3]The presence of different types of assets does not play any relevant role in the present discussion, and is kept only for coherence with the original paper.

which in matricial notation becomes

$$V = F + CV \implies V = (I - C)^{-1}F. \tag{3.2.1}$$

The rationale for this definition is obvious: if we assume for simplicity that owning a share $x\%$ of an organisation i legally gives right at any time to exchange it for $x\%$ of its fundamental assets *plus* $x\%$ of each of its positions in the other organisations, it is clear that having

$$V_i < F_i + \sum_j C_{ij}V_j \tag{3.2.2}$$

would lead to arbitrage opportunities. Any owner of $x\%$ of organisation i would be able to replace them with the corresponding share of F_i and of the positions in the other organisations, gaining a value of

$$\frac{x}{100}\left(F_i + \sum_j C_{ij}V_j - V_i\right). \tag{3.2.3}$$

in the operation.[4] So this would not be a coherent definition of "value". However, the opposite:

$$V_i > F_i + \sum_j C_{ij}V_j \tag{3.2.4}$$

would cause no such arbitrage opportunities (as long as the amount of shares of organisation i is fixed). In the present study, I will be interested in the following, more general definition of values:

$$V_B = BF + CV_B \implies V_B = (I - C)^{-1}BF, \tag{3.2.5}$$

where B is an $n \times n$ matrix. Notice that the vector of values V defined in Equation 3.2.1 (and which will be referred to as *"fundamental value"*) is V_I. Also notice that B must be such that Equation 3.2.2 is false for every i: a sufficient condition for this to hold is that $B \geq I$ element-wise. If $B \neq I$, however, V_B is not supported simply by the possibility of exchanging shares with a corresponding amount of other assets (be them fundamental assets or other shares). So for it to be a reasonable definition of value, it must be

[4] As shown by Brioschi et al. (1989) and Fedenia et al. (1994), V_i is not a proper definition of *market value* to outside investors, since for instance $\sum V_i > F_i$ whenever C is non-trivial. They define instead the market value as $v_i = C_{ii}V_i$. This distinction is inessential in the present context, since no external asset is used as numeraire, but should be taken into account whenever C_{ii} was assumed to change, that is, whenever some form of interaction with an external economy is introduced.

either imposed by an authority, or enforced by a system of prices, recognised as such by market participants. Exploring the second possibility resorts to having shares in financial organisations considered as *investments*, and hence in order to proceed forwards a temporal structure will be introduced. A discrete time model is assumed, where values will be determined by B^t for $t = 1, 2, \ldots$, and the financial network structure similarly represented by C^t. For simplicity of exposition, in the scope of the present study D will instead assumed to be constant.[5] Similarly, the financial network fill be studied as a closed system, so considering C_{ii} as fixed for every i.

The following paragraphs will be devoted to the identification of *consistent* deviations from the fundamental value - that is, *rational bubbles*.

3.3 Rational bubbles

It is helpful to introduce bubbles in a simplified setting. Consider a market with n risk neutral, infinitely patient market participants $1, 2, \ldots, n$, each with an initial amount of wealth $w_i^1 = \frac{1}{n}$. Assume that only one kind of asset is present, its total quantity present on the market amounts to 1, and its price b_t follows the stochastic process:

$$b^t = \begin{cases} \underline{b} & \text{at } t = 1, \\ \eta^t b^{t-1} & \text{if the bubble does not burst at time } t > 0, \\ b_f^t & \text{if it does.} \end{cases}$$

For the time being, it will be assumed that $b_f^t \equiv 0$. At each time t, each participant i can declare a demand $d_i^t < w_i^t$ of the asset: let $D^t = \sum d_i^t$ be the total demand, and $W^t = \sum w_i^t$ the total wealth. The bubble bursts at the first t such that $D^t < \eta^t b^{t-1}$: let us denote it as T. For each $t < T$, agents having bought the asset at time $t-1$ can sell it and receive back the investment, plus an interest of $\eta^t - 1$. At time t, the bubbly investment is hence convenient in expected terms if:

$$\begin{aligned} b^t \leq \mathbb{E}_t[b^{t+1}] =& \mathbb{E}_t[b^{t+1} | b^{t+1} \neq 0] \cdot \mathbb{P}_t\{T \neq t+1\} \\ =& \eta^{t+1} b^t \cdot \mathbb{P}_t\{T \ \neq t+1\}, \end{aligned}$$

that is:

[5] Considering D as fixed removes the possibility of arbitrage opportunities on the behalf of "overpriced" financial organisations, such as selling positions in other organisations and buying at the same time fundamental assets, keeping the value unchanged but generating a positive cash flow. A more general model allowing this choice of assets to be endogenous should consider a variation of Equation 3.2.5 such as $V_B = B(F + CV_B)$.

$$\eta^{t+1} \geq \frac{1}{1 - \mathbb{P}_t\{T = t+1\}} \tag{3.3.1}$$

(the possible inverse relation between the premium implicit in a series of bubbly prices and the risk of the bubble bursting was already suggested by Blanchard, 1979). Notice that if Equation (3.3.1) is *falsified with certainty* for some $\hat{t}$, then

$$D^{\hat{t}} = 0 \implies \mathbb{P}_{\hat{t}-1}\left\{T = \hat{t}\right\} = 1 \implies \eta^{\hat{t}} \not\geq \frac{1}{1 - \mathbb{P}_{\hat{t}-1}\left\{T = \hat{t}\right\}}$$

(it is falsified also for $\hat{t} - 1$), and by backward induction, it is easy to see that the bubble will not start in the first place. This in particular implies that for the bubble to be consistent, both $\eta^{t+1} \geq 1$ and $b^t \leq W^t$ must hold for each t. From now on, it will be implicitly assumed that this is the case.

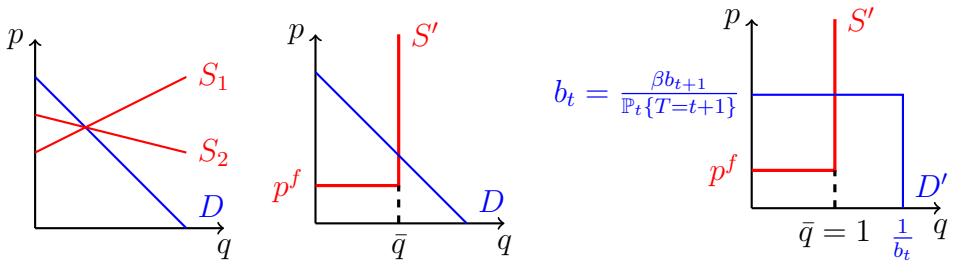
Assume that for a given time $t > 1$, Equation (3.3.1) was strict: because of the fact that agents are risk-neutral, they would invest as much as possible in the asset. That is, demand would be 1, larger than the total value of the asset. Hence, the *present* price b_t of the bubble would adapt upwards, so that $\frac{b^t}{b^{t-1}} > \eta_t$ (it could be assumed, alternatively, that market forces would someway cause a correction downwards of b^{t+1}, so that $\frac{b^{t+1}}{b^t} < \eta^{t+1}$). For the price history defined by the series of η^t to be consistent, it must hence be that, in expected terms, Equation (3.3.1) is an *equality*, except at most for $t = 1$.

3.3.1 Speculative goods

Tirole (1982) states a well known impossibility result for infinite horizon bubbles with a finite number of agents and a finite amount of wealth. Since the aim of the present section is to provide an example of such bubbles, it is instructive, before proceeding in the description, to analyse the departures from his assumptions.

One first difference is that in this model, individuals are in fact *indifferent* between investing in the bubble and not doing so, while Tirole (1982) requires for the bubble to exist that agents have a strictly positive incentive to invest in it, and as stated by Allen and Gorton (1993), *"a theory which critically depends on people's behaviour when they are indifferent is not very satisfactory"*. A practical justification for the current setting is that by forbidding the equality in Equation (3.3.1) we would make the problem unbounded, at no benefit for the economic intuition, and that the existence results will be shown to be qualitatively robust with respect to the choice of

Figure 3.1: Demand/price curves



Left: ordinary good; centre: limited quantity good with fundamental value p^f; right: the speculative good under analysis. The demand curve for generic speculative goods can be a combination of D' (for small quantities) and D (i.e. due to preference for diversification).

mixed strategies adopted, and of time preferences of individuals. A more conceptual defence is the following: if a market mechanism is characterised by an infinity of equilibria (depending on an infinite number of possible symmetric strategies, as will be clear later), of which only one - $b^t \equiv \underline{b}$ - corresponds to the non-existence of bubbles, then the indifference should rather lead to *exclude* it. A second difference is that in the context of Tirole (1982), individuals are risk-averse. This is a conceptually important difference, but it could easily neutralised by introducing a moderate degree of risk-aversion. Unfortunately, this would make the exposition and search for a closed form example more difficult.

The crucial difference (which distinguishes the present paper also from the work of Diba and Grossman, 1988c, among others) is however another: that the market is not assumed to clear. This is not by chance, and is a direct consequence of the particular nature of the good under study. The classical assumption of market clearing comes from the underlying idea that the marginal utility of consumption from some good is strictly decreasing in its quantity, while the marginal cost for producing it is increasing (or less sharply decreasing). However, in the case of speculative goods, both assumptions may fail (see Figure 3.1), so that the final price may be such that the demand is not satisfied (but *not necessarily* such that individuals are willing to pay a higher price for the same quantity). This means that the model will need to include a *rationing function* matching the exceeding demand with the fixed supply of the asset:

$$R(d_1, \ldots, d_n) \to (d'_i, \ldots, d'_n) \text{ with } \sum_{i=1}^{n} d'_i = b^t.$$

For simplicity of exposition, it will be assumed from now on that each

individual gets an amount of the asset proportional to her demand:

$$R(d_1, \dots, d_n) = b^t \frac{(d_1, \dots, d_n)}{\sum_{i=1}^n d_i},$$

and the value of the asset effectively acquired by i will be denoted as R_i.

The crucial additional ingredient for the presence of bubbles is represented by the demand functions of agents. For instance, if each market participant demands at each time period a quantity

$$d_i^t = w_i^t,$$

the aggregate demand is always $D^t = W^t$. This is compatible with a price of $b_t \equiv 0 \implies \eta_t \equiv 1$: it implies that $\mathbb{P}\{T = t\} \equiv 1$, and Equation (3.3.1) is trivially satisfied (as an equality). This is the bubble-free equilibrium. However, for any $\underline{b} \leq W^1$, the following prices series is also possible: $\eta_t \equiv 1$, $b_t \equiv \underline{b} > 0$, which corresponds to a positive, constant and everlasting bubble.

Notice that in general the wealth owned by each individual at time $t > 1$ is equal to the value of the asset owned plus the amount of wealth not invested, and in the simple symmetric setting considered until now, it can be calculated as:

$$w_i^t = R_i^t \cdot b^t + \left(w_i^1 - R_i^1 \cdot \underline{b}\right) = \frac{1}{n} \cdot b^t + \frac{1}{n} - \frac{1}{n}\underline{b} = \frac{1}{n}(1 + b^t - \underline{b}) \geq 1.$$

It should not come as a surprise that the average (nominal) wealth owned by individuals can in principle be higher than $w_i^1 = \frac{1}{n}$, although this is a closed economy: the increase in value of the asset itself would be the cause (and even the essence) of inflation.

3.3.2 A non-deterministic example

The equilibria described so far are formally valid but practically of limited interest, since they represent bubbles that *do not grow*, and *never burst*. On the other hand, an equilibrium such that $\mathbb{P}\{T = \bar{t}\} = 1$ for some given $\bar{t} > 1$ cannot exist: for the backward induction reasoning already exposed, the bubble would not start in the first place. So to have a "realistic" bubble, some level of indeterminacy must be introduced. In what follows, indeterminacy will be the result of *mixed strategies*. Again, this is a formally valid choice (since with Equation (3.3.1) being an equality, the decision of investing or not in the bubbly asset is indifferent in terms of expected utility), but for the scope of higher adherence to real world financial markets, other solutions could be conceived, such as random rationing rules, random budget

constraints, or the randomly determined possibility of asset holders to keep their shares from one period to the other.

Let the demand d_i^t of agent i at time t be uniformly distributed over $[0, w_i^t]$. The aggregate demand will have a probability distribution which is the convolution of individual demands:

$$\begin{aligned} f_D^t(x) &= \int_0^{\sum_{i=1}^n w_i^t} f_1(y_1) \int_0^{\sum_{i=2}^n w_i^t} f_2(y_2) \dots f_n \left(x - \sum_{i=1}^{n-1} y_i \right) dy_{n-1}, \dots, dy_1 \\ &= \int_0^{\sum_{i=1}^n w_i^t} \frac{1}{w_1^t} \int_0^{\sum_{i=2}^n w_i^t} \frac{1}{w_2^t} \dots \frac{\left(x - \sum_{i=1}^{n-1} y_i\right)}{w_n^t} dy_{n-1}, \dots, dy_1. \end{aligned}$$

Notwithstanding the complicated formulation of this probability distribution, which is defined on the interval $[0, W^t]$, it is easy to show that it is continuous, symmetric, single-peaked and strictly positive everywhere except than on the extrema of its domain. Moreover, it is well known (see Feller, 2008) that for large n it is approximated arbitrarily well by a normal distribution of mean $\frac{W^t}{2}$ and variance $\frac{W^t}{12}$. Let F_D^t be its cumulative distribution function: a price history compatible with this strategy must be such that

$$\eta^{t+1} = \frac{1}{1 - \mathbb{P}_t\{D^{t+1} < b^{t+1}\}} = \frac{1}{F_D^{t+1}(b^{t+1})} \qquad \forall t. \tag{3.3.2}$$

At time $t = 1$, with the price set at $b^1 = \underline{b} < 1$, the above condition becomes:

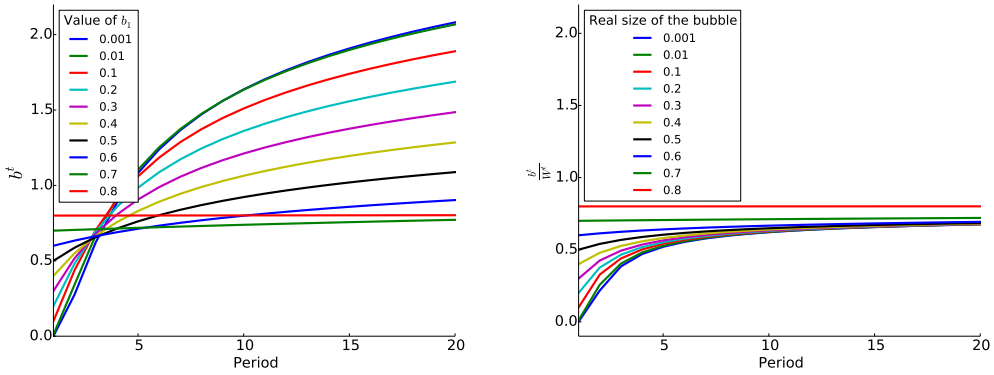
$$b^2 \cdot F_D^2(b^2) = b^1. \tag{3.3.3}$$

Consider $g(x) := x \cdot F_D^2(x)$; it is obvious that

$$\begin{aligned} g(0) &= 0 \cdot F_D^2(0) &&= 0 < b^1, \\ g(W^2) &= W^2 \cdot F_D^2(W^2) = W^2 \geq W^1 &&= 1 > b^1, \end{aligned}$$

and since F_D^2 is increasing and continuous on the interval $[0, W^t]$, $g(x)$ is strictly increasing on that same interval. So there is one (and only one) x in that interval such that $g(x) = b^1$: let it be b^2. By definition, Equation (3.3.3) is satisfied. Notice that $F_D^2(b^2) < 1$ and hence $b^2 > b^1$, as expected. By iterating this reasoning, one can see that a series of η^t compatible with $b^1 = \underline{b}$ exists and is uniquely determined, as long as b^t is lower than W^t. However, this condition is trivially satisfied, since the quantity $W^t - b^t$ is constant in time (the available wealth grows *only* as a consequence of the bubble). Figure 3.2 (left) shows an approximation of the price series for large n, when different values are adopted as b^1. Interestingly, the dynamics of the bubble are (in nominal terms) strongly sensible to initial conditions: if the price

Figure 3.2: Simulated price histories



Price of the asset over time, for different starting values (assuming that the bubble does not burst before). Left: nominal price; right: real price (b^t/W^t).

starts at 0.8, it has scarce possibilities of growing given the available wealth in the economy. If it starts at 0.1, however, by the time it reaches 0.8 the wealth of the economy will be larger, and the potential to grow further much increased. These differences however disappear if the bubble is considered in real terms (Figure 3.2, right). The different starting points then only correspond to different *stages of the development* of the bubble, as will be clarified later.

3.3.3 Alternative assumptions

It is worthwhile to consider which of the assumptions of the model are fundamental for the qualitative results it yields. The choice of the uniform distribution for the mixed strategy is not. The only characteristics of the aggregated distribution F_{D^t} which are exploited are its continuity (which is guaranteed as long as the strategies of individuals are defined in the form of probability distribution functions), and the fact that it takes value 0 and 1 respectively at the two extrema of its domain (which is a simple consequence of it being a cumulative distribution function). Notice, however, that different probability distributions could introduce strong path dependence even with $n \to \infty$, in the sense that the form of F_D^t could be largely affected by the current distribution of wealth to individuals (rather than just the total amount), while in the case considered it is always approximately a normal distribution. For similar reasons, the initial homogeneity of individuals is not a fundamental assumption, as long as n is large is enough (that is, no indi-

vidual has relevant market power). The formulation of the rationing function can also vary: again, different rationing functions (i.e. giving priority to individuals with larger demands) could introduce forms of paths dependence. Finally, the assumption that the asset looses all its value when the bubble bursts could be relaxed in two ways: by setting a fundamental value $b_f^t \gtrsim 0$ which is anyway guaranteed to investors (as will be shown later), or by establishing that another bubble immediately forms, with starting price to be determined (possibly randomly): both solutions could be easily incorporated in a new version of Equation (3.3.1), and the proof of existence would not be hindered.

All along the description of the model, it has been implicitly assumed that short-selling is impossible. A non-zero probability of short-selling in the strategies which are adopted (i.e. the individual demand being uniformly distributed over $[-w_i^t, w_i^t]$) could limit, and possibly neutralise, pricing bubbles, artificially inflating the quantity of asset available in the market. It is easy to see, however, that going short is not (strictly) profitable. That is, bubbly equilibria remain valid even under the possibility of taking negative positions. They are even compatible with moderate amounts of short-selling, although the extent of the bubble will be reduced. More interestingly, *unexpected increases* in the short-selling behaviour could lead the bubble to burst - that is, coalitions of agents (since the model assumes no single agent has significant market power) could, by coordinating, cause, and benefit from, the burst of the bubble. On the other side, another assumption implicitly made, that agents are financially constrained, clearly limits the extent to which bubbles can grow, which depends on the total amount of wealth that market members *manage*, whether or it is directly owned by them.

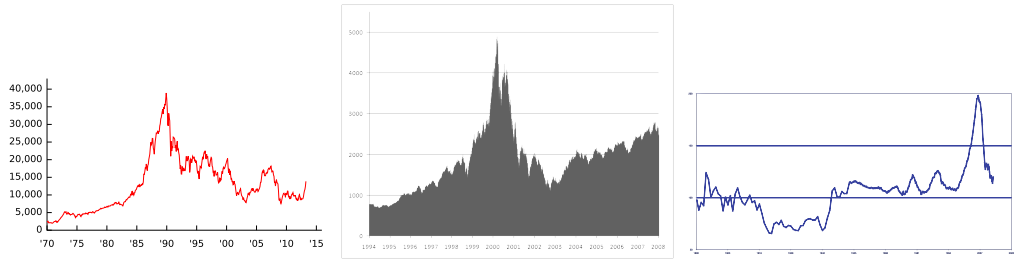
The assumption that individuals are infinitely patient and risk-neutral is more delicate. Let us start by observing that the price series b^t is unbounded. If a limit b^∞, with wealth W^∞, existed, it would be such that (recall Equation (3.3.3)):

$$b^\infty \cdot F_D^\infty(b^\infty) = b^\infty \implies F_D^\infty(b^\infty) = 1 \implies b^\infty = W^\infty.$$

This however is not possible, since we know that $\Delta := W^t - b^t$ is constant (this is reassuring, since $\frac{b^t}{W^t}$, the share of total wealth invested in the bubble, increases and tends to 1, which means that, *in real terms*, the bubble is positive *and finite*: this is not, in other words, a purely inflationary bubble as in the work of Diba and Grossman, 1988b). With this in mind, the introduction of a time discounting factor $\beta < 1$ would transform Equation (3.3.3) in

$$\beta b^t \cdot F_{D_t}(b^t) = b^{t-1},$$

Figure 3.3: Price histories of recent well-known economic bubbles



Left: the Nikkei bubble of the 80's; centre: the dot-com bubble which burst in 2001; right: the housing market bubble in the USA which peaked in 2007. *Source: Japanese Government, NASDAQ, Standards&Poors; re-elaboration: Wikipedia; images covered by GFDL license.*

and for the proof of existence to proceed as usual, it would require that $\beta W^t > b^t$, that is,

$$\beta W^t > W^t - \Delta \implies \beta > 1 - \frac{\Delta}{W^t},$$

which is false for $W^t \to \infty$. If however $\beta = \frac{1}{1-r}$, where r is the interest rate in the surrounding economy, $b_f^t = (1-r)b^{t-1}$ and $\eta^t > 1 + r$ (in particular, we will have $\eta^t \overset{t\to\infty}{\to} 1 + r$), then the choice of buying the asset can be seen as an investment in a purely speculative asset, with zero expected returns, bundled with a safe investment characterised by "normal" yields (one can think of η^t as being composed by the capital, the *dividends* r and the bubbly component $\eta^t - 1 - r$). And assuming that the asset under consideration is represented by shares of a financial organisation (as will be the case in Section 3.4), guaranteeing such yields resorts simply to investing available resources in the market, at the normal interest rate. With all this said, the asset would *still* feature bubbly dynamics, since $\eta^t > 1 + r$.

3.3.4 Confronting with evidence

It is tempting to compare Figure 3.2 (left) with some "famous" financial bubbles, such as the ones depicted in Figure 3.3. The difference is striking: while those real bubbles feature prices which, at least for a period of time preceding the crash, are a roughly *convex* function of time (prices are often said popularly to have "*increased exponentially*"), in the example just presented they are (at least after some period of time) *concave*. This is not a consequence of the particular mixed strategy considered, but a general feature of the model, and a consequence of the economic system being

closed. Several authors, starting with Blanchard and Watson (1983), have produced models of rational bubbles in which prices follow a convex curve. This is possible only assuming that the bubble is due to an incoming flow of wealth/participants (in their case, an overlapping generations model with young individuals buying the asset from the old at a price increasing from one generation to another), a condition which also allows for the probability of the bubble bursting to remain constant or even increase (rather than decreasing) over time. More to the point, although a convex bubble can be sustained for some time even by rational individuals with perfect knowledge and purely speculative interests, this must be done with the belief that other individuals, either irrational or motivated by other interests (like, in the case of a housing bubble, the need of a house in which to live), will in the end bear the cost (this is the essence of the backward induction reasoning previously exposed). The specificity of the present model is that it features dynamics which are compatible with any level of sophistication of all market participants.

The immediate interpretation of the apparent discrepancy with real world evidence is that the present model is not realistic. An alternative one would be that the model describes kinds of bubbles which, relying on weaker assumptions on the behalf of agents, and giving raise to less extreme dynamics, can be more pervasive and nevertheless generally go unnoticed. This possibility calls for empirical investigations.

3.4 Bringing the bubble to the network

Traditional models of rational bubbles are based on diverse assumptions of openness of the economic system, such as in the overlapping generation model by Blanchard and Watson (1983). They are hence unsuitable for investigating the possibility of speculation in the international financial market considered as a closed system. In Section 3.3, instead, evidence of bubbly equilibria in a closed economy was presented. The aim of the present section is to sketch the consequences of such speculative dynamics when nested in the model of financial networks described in Section 3.2. Providing a comprehensive microfoundation of such model is a daunting task, and is outside of the scope of the current work: the analysis which follows will instead be based on some simplistic assumptions and on the main intuitions presented by Elliott et al. (2013).

For simplicity, it will be assumed that for each t the matrix B^t is diagonal. Until no bubble bursts, H^t will be the (still diagonal) matrix such that $H^t \cdot V_{B^{t-1}} = V_{B^t}$. Assume that the fundamental asset number 1 is *cash*, and

set $p_1 = 1$ and $\sum_i D_{i1} = 1$.[6] Market participants will be allowed to buy and sell shares in other organisations at the current market price, defined as $V_{B^t i}$. A straightforward adaptation of the model presented in Section 3.3 to the model of networks would bring the unwanted consequence of *extreme volatility*: once controlling for the value of each organisation, the structure of the network at time $t+1$ would be uncorrelated with the structure at time t. For simplicity of exposition, I will hence make the opposite assumption: that positions are permanent, *in nominal terms*. In other words, $C^t_{ij} \cdot V^t_j$ will be non-decreasing in time, for any $i \neq j$: this will come as a consequence of the particular formulation of mixed strategies. At each time $t > 1$, each organisation i will only use available cash to buy additional shares - for each $j \neq i$, it will demand an amount d^{t+1}_{ij} uniformly distributed over the following interval:

$$\left[0, D^t_{i1} \frac{(H^{t+1}_{ii} - 1)V_{B^t_i}}{\sum_{j \neq i}(H^{t+1}_{jj} - 1)V_{B^t_j}}\right],$$

that is, proportional to the nominal *offer* in shares of organisation j.

It is easy to verify that the total demand $\sum_{j \neq i} d^{t+1}_{ji}$ of the asset i is distributed over $[0, \sum_{j \neq i} D^t_{j1}]$.

3.4.1 Failures and contagion

The analysis of *financial contagion* by Elliott et al. (2013) lies on the assumption of *failure costs*, in which an organisation i incurs when its value falls under a given level $\underline{v}_i$. This introduces non-linearities into the system, in terms of reaction to negative shocks. In turn, such non-linearities reverberate on organisations owning shares of i, which may fail themselves as a consequence, and transmit the failure further on: this is the essence of contagion. While in their model shocks are exogenously determined, in the present one, the burst of a bubble can represent an *endogenous* shock for the underlying financial organisation, pushing $V_{B^t i}$ under $\underline{v}_i$. The analysis of expected profits will hence need to include the possibility of failure due to contagion from other failing organisations. Although a closed form description of the system becomes prohibitive (in particular because organisations can now loose *part* of their value without their bubble bursting, either because $B^t_{ii} \equiv 1$, or because the demand of asset i is able to compensate for the loss), it is still clear that prices must satisfy the condition of indifference, namely:

$$V_{B^t i} = \mathbb{E}[V_{B^{t+1} i}].$$

[6] Considering cash as the numeraire good is a convenience assumption, without any particular economic consequence, and the same holds for the total amount of money in the system: other prices will simply adapt as consequence.

Hence, if $H_{ii}^{t+1} > H_{jj}^{t+1}$, it is a sign that the risk associated with holding a share of organisation i is higher than the one associated to a share of organisation j. However, all else equal,[7] H_{ii} is also positively correlated with systemic importance, since the sum of weights of all links to organisation i (or in other terms, the average relative importance of i in the balance sheet of other organisations) is precisely $V_{B^t i}$. This allows us to identify one first channel through which the presence of bubbles interacts with the topology of the financial network.

> **Stylised result 1:** If prices of organisations shares are affected by pricing bubbles, the risk of failure of such bubbles is positively correlated with the systemic importance of organisations.

Consider now two organisations i, j, and assume that they are ex-ante identical, in the sense that $D_{ih} = D_{jh}$ for each $h \notin \{i, j\}$, $F_i = F_j$, $H_i^2 = H_j^2$ and $C_{ih}^1 = C_{jh}^1$ for each $h \notin \{i, j\}$. Moreover, let l be a third organisation such that the diagonal element $H^t{}_{ll}$ is larger than any other $H^t{}_{hh}$, for all t. If $C_{il}^2 > C_{jl}^2$, that is, organisation i happens to buy a larger amount of shares of l, then, all else equal, its value will also grow faster during the period of growth of the bubble for asset l: $\mathbb{E}[V_{B_i}^t] > \mathbb{E}[V_{B_j}^t]$ for $t > 1$. The immediate consequence of this is that the financial network will grow (in the sense of becoming more interconnected) according to a model of preferential attachment (see Barabási and Albert, 1999). A deeper analysis suggests however that the propensity of a node i to attract future links will not just depend on the *current* volume of node i, but also on the future volume of organisations of which i owns shares, inducing a form of *"preferential preferential attachment"*, the ultimate effect of which will be a strong *clustering* effect of larger players.

The analysis of systemic risk of Elliott et al. (2013) revolves around the two crucial concepts of *integration* and *diversification*, which are both shown to have non-monotonic effects in terms of the propensity for cascades of failures to form. The former, defined as *"how much [of an organisation] is cross-held by other organisations"*, in unaffected by the presence of bubbles in the present model, since by assumption $\hat{C}_{ii}$ is considered as fixed. The second instead, defined as the fact that a *"typical organisation [is] held by many others, or by just a few"*, is strongly affected by the clustering effect mentioned above, which leads to a *concentration* of investments. Concentration can have in principle a beneficial effect on systemic stability: if organisation i has (relevant) financial connections to only a few organisations, then it will

[7] More precisely, the underlying assumption is that the share $\hat{C}_{ii}$ of each organisation held by *outside* shareholders is uncorrelated with the presence and entity of the pricing bubble.

not be affected by failures of others ones (as an extreme case, if the financial network contains two separate components, no cascade will ever transmit from one to the other). This beneficial effect is however cancelled in the present setting, since investments will be concentrated on *more risky* organisations, which are in turn the most central ones. Tihs can be summarised in what follows.

> **Stylised result 2:** The presence of pricing bubbles induces a decrease of diversification which is detrimental to systemic stability.

Notice that this result is unambiguous only as long as all shocks are assumed to be due to the burst of pricing bubbles. If exogenous shocks are considered, and their probability and size assumed to be uncorrelated with the size of the organisation affected, then large organisations could be able to absorb them more easily, and in the end be "safer neighbours" than small ones: since pricing bubbles tend to make organisations grow larger, this could counterbalance the negative effects of concentration. [8]

Finally, it is worth observing that although the whole mechanism of pricing bubbles is due to a *multiplicity of equilibria*, à la Diamond and Dybvig (1983), the issue of *bank runs* was not investigated (as in the work of Elliott et al., 2013, who show that the multiplicity issue goes *beyond* the pure bank run phenomenon, the current analysis focuses on the *best-case* equilibrium): the possibility of agents reversing their expectations of future prices could be an evident source of new risk. In order to take it into account, however, the exchange of information between agents would need to be modelled.

3.5 Discussion and conclusions

Pricing bubbles have traditionally been associated to *psychologic* motives, or to heterogeneity of agents, in terms of timing, preferences, or utilities. In contrast, phenomena of self-fulfilling expectations can make pricing bubbles possible even in closed systems of homogeneous and perfectly rational agents, as long as the assets at stake are inherently *limited*. In the light of this result, the possible effects of bubbles are studied on a microfoundation of the model of financial networks by Elliott et al. (2013). Not only does the presence of bubbles represent an endogenous source of risk: their dynamics

[8] Cabrales et al. (2013) show however, in a similar setup, that once the spreading of value shocks, rather than just the probability and extent of cascades of failures, is taken into account, optimal configurations are characterised by perfect assortativity in size.

also alter the topology of the financial network in such a way as to make it more fragile with respect to cascades of failures. The famous vision of the financial market by Keynes, as a *"beauty contest"* in which the main criterion for purchasing an asset is the expected demand for it, becomes then compatible with any level of sophistication of market participants, and this observation has alarming implications for analyses of inherent volatility and systemic risk in the international financial market. Speculation may not be, after all, a symptom of (*microeconomic*) market imperfections or exogenous asymmetries.

Several open issues are left to further research. The main assumptions underlying the present model are that once the future history of prices is set, any coordination between market players is absent, and that strategies are symmetric in nature. Those assumptions are clearly unrealistic, and the effect of strategic heterogeneity and coordination on the possibility of pricing bubbles would be an interesting issue of research. Moreover, the assumption of mixed strategies could itself be dispensed with, as long as some indeterminacy is introduced at other levels: for instance, market frictions which affect purchases and sales in a non-deterministic way could also make bubbly prices *strictly convenient* in expected terms, rather than simply indifferent. This would in turn make the existence of rational bubbles in real world markets even more plausible.

While this study only provides qualitative evidence in terms of risk changes due to bubbly prices, the actual consequences for the probability and expected damage of financial crises could be estimated either by finding closed form equilibria, or with simulations. Most importantly, analyses of data from financial markets could shed light on the presence of such rational bubbles in the real world, by identifying systematic *concavities* of prices time series in periods of growth (analogously to the search of *explosive* rational bubbles by Diba and Grossman, 1988a and Lux and Sornette, 2002 among others).

Chapter 4

The Impact of Peer Pressure on Tax Compliance: a Field Experiment

4.1 Introduction

The literature on fiscal compliance has developed from the seminal model of Allingham and Sandmo (1972). In their work (as in similar studies by Kolm, 1973, and Singh, 1973), expected utility maximizing agents choose the income level to be reported to the fiscal authority, considering the probability of being audited and the size of the fine. At the empirical level, however, researchers have faced a major puzzle: in all advanced economies, the level of tax compliance is far higher than the one predicted by such theories (Graetz and Wilde, 1985, Alm et al., 1992). This paper deals with such puzzle by proposing and implementing a new experimental design.

A stream of literature has approached the discrepancy by extending the original model with more realistic specifications of the context in which tax declaration choices are made. In this context, financial strain (Wärneryd and Walerud, 1982) and the broad category of *opportunities* have been analyzed, among other factors. The role of third-party reporting, which limits the possibility of employees to evade taxes, has been widely discussed (Andreoni et al., 1998) and tested experimentally (Slemrod, 2007 and Kleven et al., 2011). Even these studies, however, recognize that the high level of compliance which is observed empirically cannot be fully explained without taking into account behavioral factors. This view, which is nowadays widespread in the literature, is the starting point for the present study.

Several extensions of the basic model have been proposed, which take into consideration non-monetary motives. Bordignon (1993) embeds fairness-based evaluations into the utility function, while Gordon (1989) introduces non-pecuniary *stigma* costs associated with tax evasion. Weigel et al. (1987) and Groenland and Van Veldhoven (1983) provide a *social and psychological model*, which represents a broader approach to the several conditions which influence fiscal behavior, such as personality (Lewis, 2011). Studies on behavioral aspects of tax compliance are rooted in the wider stream of literature about the social aspects of deterrence (see for instance Grasmick and Bursik Jr, 1990 and Paternoster et al., 1983).

We contribute to such literature by testing, through a field experiment, the salience, for the fiscal compliance of *sellers*, of *direct peer pressure*. A relevant feature of our experimental settings is that such peer pressure can be cleanly isolated from the ordinary commercial transaction, also in the economic sense. That is, there is no reasonable interpretation for the request of the client, apart from an act of unselfish *nudging* towards fiscal compliance.

There are, on the other hand, multiple channels through which such pressure can influence the fiscal behavior of sellers; for ease of exposition, we will regroup them in three classes: *honesty*, *opportunity*, and *conformism*. A vast

empirical literature points at the importance of *social norms* in regulating human behavior: in this sense, peer pressure can signal the status of fiscal compliance as a social norm, pushing the seller towards honesty, in order to avoid the psychic cost related to fiscal evasion, or the sense of shame if being caught. Then, even if a seller is not *personally* concerned with the social norm, the fact that some clients do favor honesty could still incentivize a form of purely opportunistic compliance targeted at not loosing part of the customer base. Finally, the mere knowledge, or feeling, that fiscal compliance is *the typical behavior* in the community of reference could raise the perceived probability of audits, by representing a signal of the probability perceived by others. While empirically disentangling such effects is a hard task, a consistent literature has identified the first as particularly relevant for the taxpayer decision (see Kirchler, 2007 pp. 64-65, and the more recent work by Galbiati and Zanella, 2012). Erard and Feinstein (1994) merge the complementary approaches of utility maximization and tax morale by exploring the consequences of an (exogenously given) share of *honest* taxpayers on the audit rate for other citizens - and hence on their fiscal behavior. In equilibrium, even purely selfish citizens end up paying more than in the basic framework. The goal of our experiment is not to isolate one specific channel, but rather to measure empirically the overall effect that peer pressure has on the compliance choice, a measure which has relevant policy implications.

Studies of tax compliance have been historically confronted with a lack of data that is particularly hard to overcome, as effectively summarized in Cowell (1991): *"Data from official investigations are hardly ever available and data from other sources may be suspect: if you could directly observe and measure a hidden activity, then presumably it could not really have been properly hidden in the first place."* Weigel et al. (1987) considered as fundamental for future fiscal research the development of *creative methods* for attaining objective estimates of tax evasion behavior. The issue is still open, as reported in recent years by Halla (2012). In particular, the frequent use of survey data, where individuals self-report their tax behavior, has since long been perceived as a crucial issue (Weigel et al., 1987, Elffers et al., 1987), because of the possible misreporting. Therefore, a growing stream of literature has focused on *experiments* aimed at reproducing the economic and psychological reasoning behind tax compliance. This stream can be traced back to Reis and Gruzen (1976) and Kidder et al. (1977); more recent attempts in this direction are those of Alm et al. (1992) and Cummings et al. (2006). In his exhaustive review of the field, Torgler (2002) acknowledges the relevance of experiments in that tax enforcement, tax rate and income levels can be controlled.

The effect of social norms and social disapproval on tax compliance has

been approached experimentally for instance by Bosco and Mittone (1997). Their design allows to test two separate hypotheses, concerning the effects of either *subjective* or *collective* moral constraints on tax compliance. Subjective moral constraints are manipulated as follows: while in the control group money collected through taxes is just taken away from participants, in the experimental treatment there is a partial redistribution of the collected amount. In order to test for the second hypothesis, instead, a treatment is run in which the identity of individuals who are caught cheating is publicly revealed, and evaders hence run the risk of being identified as such by other participants. The authors find significant evidence only in favor of the first hypothesis. More recent examples of experimental studies on the effect of social pressure on the compliance choice are for instance the works by Cummings et al. (2001), Alm et al. (2007), and Fortin et al. (2007).

However, Torgler (2002) casts doubts on the fact that laboratory experiments can be considered informative about actual tax compliance behavior. This view is shared by Halla (2012), who suggests that individuals react to experimenters' stimuli differently than with real tax authorities. Indeed, social norms are part of the *culture* of any society, of which a laboratory experiment allows to study only schematized traits, and at the same time they are a fundamental ingredient of the compliance decision (Posner, 2000) because they "constitute constraints on individual behavior beyond the legal, information and budget constraints usually considered by economists" (Fehr et al., 2002).

Although their number has been recently increasing, relatively few attempts have been made to identify the size and the determinants of tax evasion through the use of field experiments. This is due in part to the typical reluctance of national fiscal authorities toward randomized actions (other than budget-motivated randomized audits such as those described by Erard et al., 2002), which are supposed to go against the principle of equity.[1] The approach of Schwartz and Orleans (1967), later adopted by Wenzel (2001), is based on surveys sent to taxpayers some time before they file their tax declaration. The questions asked vary from group to group: this enables the authors to identify the reduction in evasion due to *"conscience"* versus the one due to *"sanctions"*, by arousing respectively the feeling of *guilt* related to the social loss or the *fear* of detection. While they find that the relative importance of the two motives depends on the social and economic status of individuals, overall they report that *"conscience appeals are more effective*

[1] Randomized setups are characterized precisely by the fact that they treat equal citizens differently, rather than shaping enforcement actions deterministically on observable variables.

than sanction threats". Slemrod et al. (2001), through threat-of-audit letters, identify the response of taxpayers to an increase in audit probability, and report mixed evidence. They find an increase in amounts declared by low and middle-income taxpayers, but a *decrease* in amounts declared by high-income ones. This result is attributed to the particular wording used in the letters, together with the heterogeneity of beliefs and of information that individuals have about the fiscal authority. Kleven et al. (2011) bring into the picture the effects of an audit itself on subsequent tax declarations, as an indicator of undeclared income. Their main conclusion is that fiscal evasion is severely hindered by third-party reporting. Still, they acknowledge the evidence of behavioral factors: even though audits do not imply a higher audit probability in the future, they have a positive deterrence effect for the following fiscal year. Finally, Fellner et al. (2013), in addition to independently testing the effect of a *threat* (a message directed at changing the perceived sanction risk) and of a *moral appeal* (stressing that evasion is an act against *fairness*, which harms honest taxpayers), introduce the innovative element of *social information*. A subsample of their subjects is informed of the compliance rate for the specific TV license fees on which the experiment is based. The authors show that the effect of such new information goes in the direction of *conformity*, by increasing (decreasing) the compliance of individual with lower (higher) prior expectations on the compliance rate.

While Fellner and coauthors interestingly bring into the picture the effect of social pressure, their study has in common with the other field experiments previously cited that the treatment comes from the interaction of citizens with *institutions* - in particular, it is determined in the context of the surveys, audits, or threat-of-audit letter that these institutions implement. Instead, to the best of our knowledge, no field experiment on fiscal compliance has been previously implemented focusing on the *direct* effect of social pressure between peers, as in the tradition of experiments on peer pressure started by the seminal work of Asch (1955) (also see Falk and Fischbacher (2002) and Falk and Ichino (2005)). The present paper tries to fill this gap by exploiting the particular case of tax evasion among shop sellers in Italy, a country where non-compliance is relatively widespread (as confirmed both by official reports, and by our experimental data). The vast majority of Italian shops are obliged by the law to release a tax receipt for each sale.[2] The Italian law imposes some strict technical requirements: receipts must be released by approved cash registers, they must feature specific markings, they must be numbered progressively, starting from 1 at the beginning of each day, and

[2]Supermarkets and kiosks are among the few exceptions, which however are irrelevant in the present context.

while they must be released to the clients, an authoritative copy is kept in the cash registers, and must then be stored in the event of tax inspections. The total sum of receipt amounts represents the revenues of a shop, and receipts themselves can constitute a proof for the fiscal authority that sales occurred. Both value added tax and income tax are then calculated on the basis of such revenues. As a consequence, the omitted release of a receipt is an act of fiscal evasion, allowing the seller to evade both value added and income tax. Interestingly, this act of tax evasion is not only common, but also, at least in the case under analysis, committed *openly*, making it trivial for a purchaser to ascertain non-compliance. Although it would also be trivial for the purchaser to actively *fight* tax evasion - by simply requesting the receipt when it is not released - this behavior is far from being widespread. As will be clarified later, when such a request is made, it is an unambiguous act of social pressure. Moreover, although any client who does not receive the receipt for a given sale can denounce the shop to the fiscal authority, this becomes impossible once the receipt has been (requested and) released.

According to Kirchler (2007) social norms are a function of an individual's perceived expectation that one or more relevant referents would approve a particular behavior. In the prevailing literature, such "relevant others" are members of the group where the principle of reciprocity applies; the concept of *strong reciprocity*[3] and its importance for the enforcement of social norms are studied in detail by Fehr and Gächter (1998) and Fehr et al. (2002). It could be questioned whether the occasional customer who requests a receipt, without having any stable tie with the subject studied (the seller), could be considered as a *relevant other*. While the effect we find is also compatible with the two other purely utilitarian channels already proposed for the effect of social pressure - *opportunity* and *conformism* - it still proves that the ethical views of even occasional customers *do* influence the vendor, who evidently sees them as representatives of the "common view" of the whole Italian, or local, society.

Our experiment is based on an ordinary interaction between clients and sellers, except for the fact that clients exert social pressure on non-compliant sellers, by asking them to release a tax receipt. Hence, the positive effect on tax compliance that we find has relevant policy implications, since it highlights the role that an aware population can have in the fight against tax evasion. We also feature evidence of the fact that non-compliance is higher when the client is of the same gender of the seller, suggesting that tax evasion

[3] "A person is a strong reciprocator if she is willing to sacrifice resources (a) to be kind to those who are being kind (strong positive reciprocity) and (b) to punish those who are being unkind (strong negative reciprocity)."

is *easier* when the decision maker feels his peers as socially closer.

The remainder of the paper is structured as follows. Section 4.2 presents the theoretical framework for our analysis. Section 4.3 describes our experimental design. Section 4.5 presents our results, which are discussed in Section 4.6. Finally, Section 4.7 summarizes our conclusions.

4.2 Theoretical background

Bordignon (1993) bases his explanation of "excessive" tax compliance on *fairness-based* motives: citizens do not only care about their own net income, but also about social welfare. While this approach certainly brings into the picture an interesting aspect of taxpayers' decision making, its possible implications in terms of social pressure are non obvious.[4]

In the model of Gordon (1989), instead, the utility of the taxpayer takes the general form

$$u(C, E)$$

where C is increasing in ordinary consumption C, and decreasing in E, the amount of undeclared income (embodying the concept of *stigma*). Consumption in turn depends on the amount of income concealed:

$$C = Y(1 - \tau) + x\tau E$$

with Y representing disposable income, and τ the tax rate. $x = -s$ if the cheater is caught, $x = 1$ otherwise. This simple formulation clearly implies a trade-off between the *stigma cost* and the cost of compliance: under given separability assumptions, Gordon rewrites the utility function as

$$u(C, E) = U(C) - vE$$

where vE is the *private psychic cost of evasion.* v is assumed to be distributed in the population according to some distribution $F(v)$.

The choice of relying on the concept of *stigma* has the interesting implication that individuals may pay more taxes than what would be optimal for their balance sheet, *in particular* if they are more afraid to be caught (even keeping fixed the expected fine).

[4]Notice that in principle the effect of social pressure could partly overlap with fairness motives. It could be interpreted by the potential evader as a signal of private information on the trustworthiness of the public government - and hence of the good use of public revenues. This could be a driving factor toward higher compliance for a fairness-motivated seller.

We can extend the model by enriching the utility function, assuming that the psychic cost for the taxpayer depends on the *ethical tastes* of the peers:

$$\tilde{u}(C, E) = U(C) - f(\pi)E$$

where π captures the extent to which the individual is concerned with the social attitude against tax non-compliance, and f is a strictly increasing function. Without loss of generality, we set $f(0) = v$.[5]

When $\pi = 0$, the optimal level of tax evasion e^* is, by definition, the same as in the original model by Gordon. If instead $\pi > 0$, we can write

$$\frac{\partial \tilde{u}}{\partial E}(e^*) = \frac{\partial u}{\partial E}(e^*) - (f(\pi) - v).$$

By definition of e^*, we have that $\frac{\partial u}{\partial E}(e^*) = 0$, and hence

$$\frac{\partial \tilde{u}}{\partial E}(e^*) = -(f(\pi) - v) = -(f(\pi) - f(0)) < 0;$$

that means e^* now represents a suboptimally high level of evasion: the new optimal level will be $\tilde{e}^* < e^*$. It is easy to verify that the starting condition $\pi > 0$ is both sufficient and necessary for this conclusion, and that in general $\tilde{e}^*$ decreases for increasing values of π. According to this formulation, the aim of our experiment is precisely to verify empirically if $\pi > 0$ - that is, if sellers are concerned with the social attitude towards tax evasion.

Moreover, the experiment may shed some light on an additional hypothesis suggested by the model: "individuals cheat less if they are observed by *more* peers". Notice that the model does not give an unambiguous prediction concerning this hypothesis: the effect of peers simply standing by while an individual makes her compliance choice is unclear: it depends on her belief concerning the "ethical tastes" of those peers.

4.3 The experiment

The particular category of businesses which we study - bakeries - exhibits some features that make it suitable for our experiment. First, the good at sale, bread, is relatively standardized, making it easier to compare different

[5] For π to be 0, it is sufficient that the society does not take a stand against tax evasion, *or* that it does, but the seller is unconcerned or unaware. In this latter case, a request for the receipt could represent an information on previously *unknown* ethical tastes, or updated information about a recent *change* in such tastes.

shops. Second, it has a low cost, which implies that the profit obtained by evading is generally not the object of bargaining between the seller and the buyer.[6] Therefore, the act of requesting the receipt does not affect the utility of the buyer: it can instead be considered as an uninterested act of social pressure. Third, the volume of daily sales is typically high, so even assuming that *occasional* clients are a minority, the presence of two of them in a time span of 12 minutes is far from exceptional (see Section 4.6.1 for an estimate of the number of sales).

The experimental sample consisted of 108 bakeries located in the city of Milan. For each bakery, the time line of the experiment was articulated in two periods. In period 1, an agent entered the shop and bought a loaf of bread. If the receipt was not released, the agent would ask for it:[7] this request was our treatment. In period 2, twelve minutes after the first agent left the shop, another agent entered the same bakery. Following the same procedure as in period 1, the agent bought a loaf of bread of a different type. The role of this second agent was to assess if the receipt was now given. Whatever was the behavior of the seller, no request for a receipt took place at this time.

In all cases, the purchase was paid with an amount of money higher than its cost,[8] so that the agent had to wait for the change. This design choice was made because a client standing still after having paid and received the bread would have probably influenced the behavior of the seller. In this way instead, the moment in which the change was given (with or without the receipt) represented unambiguously the end of the transaction. The choice of the twelve minutes time span was made because it is absolutely unlikely that any client would spend such an amount of time in a bakery. This means that when the second agent entered, the first one, as well as the clients present in the shop when the request for the receipt had taken place, had already left the shop. In this way, any change in the behavior of sellers can be attributed uniquely to a reaction of the sellers themselves to the request, as opposed to *indirect* pressure, or to the presence of the client who proved to be particularly "picky".

Note that the request for a receipt was made only for non-compliant

[6] Bargaining is known to be far from being unrealistic in other sectors, in which the amount of evasion benefits *per purchase* is much higher and the buyer is often offered a discount conditional on not receiving the receipt.

[7] The receipt was always requested using the same wording ("*Vorrebbe essere così gentile da rilasciarmi lo scontrino?*" , which roughly translates to "*Would you be so kind as to give me the receipt?*").

[8] For the sake of homogeneity, banknotes were never used, and the amount given was always lower than 2 €.

bakeries, and to *all* of them. In principle, it would have been possible to randomly select *ex ante* one half of the bakeries to be used as a control, by *not* asking them the receipt, in any case. This was not our choice for three reasons. First, the population of non-compliant bakeries at a given time in a given city is intrinsically limited, and by using half of them as control we would have halved the power of any test based on the experimental data. Second, receiving a receipt from a shop allows us to extract some interesting information on the shop emitting it - for instance, the number of receipts emitted since the beginning of the working day. Third, and most importantly, our identification strategy *does not rely* on the presence of a control for the identification of a causal effect, as explained in Section 4.4.

For the later interpretation of the results, two aspects of our experimental design are worth stressing. First, although explicit requests for receipts are presumably rare in Italy, all other aspects of the interaction between the buyers and the sellers in the experiment are absolutely ordinary. The change of agent and of type of bread being asked from one pass to the other, together with the fact that bakeries are characterized by a high number of low volume sales, make it virtually impossible that any seller noticed anything unusual - apart obviously from the rebuke, when there was one. Second, although the seller can have, as already mentioned, an opportunistic response, the request for the receipt itself can *only* be interpreted as an act of gratuitous social pressure, signaling adherence to the social norm of fiscal compliance, rather than self-interest. Bread is not covered by any warranty for which the receipt could serve as a proof of the purchase. The Italian legislation does not envisage sanctions for clients unable to show the receipt of a purchase just made.[9] And although in principle a client can denounce a seller for not releasing the receipt relative to a purchase, this becomes impossible precisely after the receipt has been requested and released.

Since the experiment consisted of buying twice from each of the bakeries and then comparing the results in terms of compliance, we also made sure that each of the two passes had, on average, exactly the same characteristics (except for the treatment), in order to remove any potential confounding factor. Therefore, (a) the entry order of the two agents (one male and one female, both around 25 years of age) was randomized; (b) the types of bread purchased were randomized;[10] most importantly (c), the second agent *did*

[9]Such sanctions were theoretically present, although very rarely implemented, until 2003, when a legislative change left only the existing sanctions on sellers.

[10]Each time, one agent asked for a type of bread and the other one asked for another, resorting to a third and then to other types if the requested one was not available. The three types chosen are comparable in weight, size, cost, and all of them are usually sold by any bakery.

not know if the first had been spontaneously given the receipt and hence if the bakery had been treated. This is important for the causal interpretation of the results: it is clear that if a seller suspects a client of being a tax inspector in plainclothes, he will release a receipt spontaneously. So in principle the probability of a tax receipt being spontaneously released may strongly depends on physical or dressing characteristics of agents which my raise such suspects on behalf of sellers. By randomizing the order of the agents, however, this effect can be easily controlled for (see Section 4.6.2). The fact that the two sales happen in a short time span also guarantees that the result does not depend from the timing/day of the purchase (i.e. if tax inspectors are expected to arrive mostly in the morning). Finally, it is worth remarking that the release of the receipt, even if as a consequence of an explicit request, is the best guarantee for the seller that the client is not a tax inspector and will not denounce the seller to the authority - after all, the receipt *was* emitted.

Before proceeding to the analysis of our data, it should be pointed out that the two agents could possibly meet two different sellers inside the shop. This is apparently not a frequent event: the characteristics (gender, apparent age and ethnicity) of the vendor which were recorded by agents show a mismatch between the two passes only in the 14% of cases. While indeed in the 40% of cases at least two vendors were present during the purchase, the size of the shop and the number of other clients were typically such that *any* seller would notice the request for the receipt. Finally, even if our estimates are *within bakery* rather than *within seller*, they are what matters for policy implications (social pressure is still relevant even if felt indirectly by another seller of the shop) and presumably represent a lower bound for the *within seller* effect.

In order to obtain a preliminary assessment of tax compliance, a single pass was carried out in January 2012, with 177 bakeries being investigated. Of these, 22% did not release a receipt to the agents.[11] Given the purely descriptive purpose of this first investigation, no treatment was administered. The field experiment took place shortly after, in the first 2 weeks of March 2012. At that time, all the bakeries that had not released the receipt during the preliminary investigation were inspected, together with 70 other bakeries randomly selected. In total, 108 shops were included in the experiment: 21% did not release the receipt during the first pass, and hence were treated. Of these, 13 were treated by a female agent, 10 by a male agent - each agent

[11] Additionally, at least 15 other bakeries (8.5%) released a receipt to the agents but *not* to some other client present at the same time, for a total of at least 30.5% of bakeries observed in a non law-abiding behavior.

had entered as first in exactly 50% of the total bakeries. Among the treated bakeries, one type of bread had been asked in 11 cases, and the other type in 12 cases. A summary of the compliance outcomes which will be relevant for the analysis is presented in Table 4.2.

It should be noted that the conditions of our field experiment are not entirely controlled: i.e. the agents may not represent faithfully the typical client of bakeries analyzed, in particular in terms of loyalty, and hence the theoretical framework described in section 4.2 applies only for specific subgroups of the population of clients. This observation calls for additional research, i.e. it is reasonable to hypothesize that a request made by an occasional client influence the behavior of the seller mainly when facing occasional clients; at the same time, it is also reasonable to assume that a request made by a loyal client would affect the behavior of the seller to a larger extent.

4.4 Methods

We are interested in estimating the causal effect of social pressure on tax compliance. The population of interest for our experiment is the set of all bakeries in the city of Milan. The treatment (denoted by the Boolean variable D_i) consists of requesting the receipt if not released spontaneously: therefore, only non-compliant bakeries are treated. In the terminology of the classic treatment-effect framework, the aim of our analysis is the estimation of the Average Treatment on the Treated (ATT), answering the following question: "does exerting social pressure *on non-abiding sellers* affect their propensity to tax compliance?" In fact, it is hardly possible to measure an ATE (Average Treatment Effect), since it is difficult to imagine treating bakeries where the receipt is spontaneously given.[12]

Table 4.1 introduces the moments of data on which our analysis is based. The compliance status observed in a generic bakery during pass j ($j = 1, 2$) is denoted as c_i^j. This is a Boolean variable, equal to 1 if and only if the bakery did release the receipt. $\mathbb{P}_D$ denotes probabilities for treated bakeries, while $\mathbb{P}$ represents the "natural" probability, in absence of any treatment.

[12]In principle, an experiment could be ran in which social pressure is exerted *at the start* of the transaction, for instance with agents stating explicitly they want the receipt at the moment of asking the loaf of bread. This experimental setup, by randomizing the sample of treated bakeries, would indeed produce an ATE. This was not our choice for two reasons. First, the measurable effects would have been largely diluted. Second, the practical implementation of our experiment is much more realistic: it does happen, although it is certainly not the norm, that clients not receiving the receipt ask for it, while it is much less common that a client asks for the receipt beforehand.

Table 4.1: Relevant probabilities

		Second pass (c^2)		
		0	1	Total
First pass (c^1)	0	$\mathbb{P}_D\{c^1=0,c^2=0\}$	$\mathbb{P}_D\{c^1=0,c^2=1\}$	$\mathbb{P}\{c^1=0\}$
	1	$\mathbb{P}\{c^1=1,c^2=0\}$	$\mathbb{P}\{c^1=1,c^2=1\}$	$\mathbb{P}\{c^1=1\}$

The identification of the ATT amounts to finding a causal relationship between D_i and c_i^2, restricting the attention to bakeries with $c_i^1=0$. However, this cannot be done by simply regressing c_i^2 on D_i, since in our sample D_i is perfectly collinear with c_i^1. In what follows, we describe instead our identification strategy. By definition, the Average Treatment Effect on the Treated can be expressed as:

$$ATT = \mathbb{E}\left\{c^2|c^1=0,D=1\right\} - \mathbb{E}\left\{c^2|c^1=0,D=0\right\}. \tag{4.4.1}$$

Notice that the only difference between the two terms on the right hand side is the treatment status, so that we can rewrite the above as:

$$\begin{aligned} ATT =& \frac{\mathbb{P}_D\{c^1=0,c^2=1\}}{\mathbb{P}\{c^1=0\}} - \frac{\mathbb{P}\{c^1=0,c^2=1\}}{\mathbb{P}\{c^1=0\}} \\ =& \frac{\mathbb{P}_D\{c^1=0,c^2=1\} - \mathbb{P}\{c^1=0,c^2=1\}}{\mathbb{P}\{c^1=0\}}. \end{aligned} \tag{4.4.2}$$

We mentioned in Section 4.3 that the two passes were designed in order to be perfectly comparable.[13] Intuitively, this means that the "natural" compliance rate - in the absence of any request - would be the same for the first and second pass; formally, $\mathbb{P}\{c^2=1\}=\mathbb{P}\{c^1=1\}$. A straightforward consequence of this is that, still in the absence of any treatment, the probability of the event "compliance only at the first sale" ($c^1=1$, $c^2=0$) is the same as the probability of "compliance only at the second sale" ($c^1=0$, $c^2=1$):[14]

$$\mathbb{P}\{c^1=0,c^2=1\} = \mathbb{P}\{c^1=1,c^2=0\}. \tag{4.4.3}$$

[13] While an obvious difference between the two passes is that 12 minutes elapsed in between, our data does not support the hypothesis that compliance is systematically increasing as time passes.

[14] Formally,

$$\begin{aligned} \mathbb{P}\{c^1=1\wedge c^2=0\} =& \mathbb{P}\{c^1=1\} - \mathbb{P}\{c^2=1\wedge c^2=1\} \textit{ and} \\ \mathbb{P}\{c^1=0\wedge c^2=1\} =& \mathbb{P}\{c^2=1\} - \mathbb{P}\{c^2=1\wedge c^2=1\}. \end{aligned}$$

The two coincide if and only if $\mathbb{P}\{c^2=1\}=\mathbb{P}\{c^1=1\}$, which is true by assumption.

We can estimate the second term of equation (4.4.3) from data for bakeries that released the receipt at the first pass ($c^1 = 1$), and hence were not treated. As a consequence, equation (4.4.2) can be rewritten as

$$ATT = \frac{\mathbb{P}_D\{c^1 = 0, c^2 = 1\} - \mathbb{P}\{c^1 = 1, c^2 = 0\}}{\mathbb{P}\{c^1 = 0\}}. \tag{4.4.4}$$

All terms appearing in (4.4.4) can be estimated from our data.

4.5 Results

Table 4.2 summarizes the experimental data, providing the empirical realizations corresponding to the probabilities of Table 4.1. Notice that the frequency of compliance during the second pass ($c^2 = 0$) was 82%, higher than during the first pass.

Table 4.2: Summary of data

		Second pass (c^2)		
		0	1	Total
First pass (c^1)	0	7	16	23 (21 %)
	1	12	73	85 (79%)
	Total	19 (18%)	89 (82%)	108 (100%)

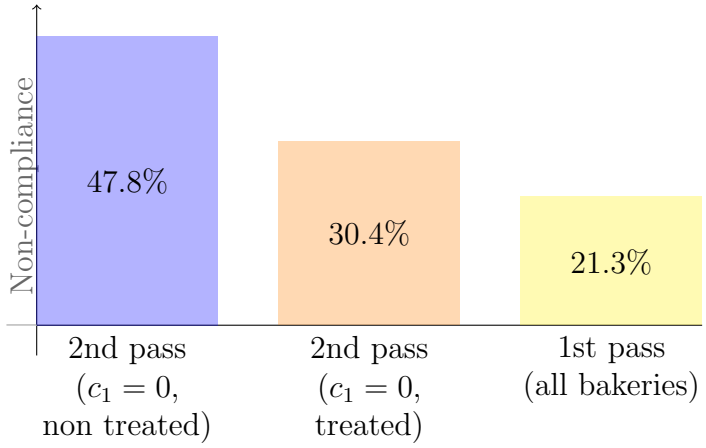
By estimating (4.4.4) we obtain:

$$ATT \approx \frac{\frac{16}{108} - \frac{12}{108}}{\frac{23}{108}} = 17.4\%.$$

For bakeries which did not release the receipt, the treatment lowers the probability of not receiving the receipt in the second pass by 17.4 points: it decreases from 47.8% to 30.4%. Notice that both values are higher than the non-compliance rate we expect from a generic bakery, which is 21% (see Table 4.2, and recall that for an *untreated* bakery the two passes are perfectly equivalent). Hence, our data show that bakeries which are non-compliant in the first pass are *intrinsically* different from the others, thus providing strong evidence of illegal behavior persistence. This effect is so strong that non-compliance in the second pass is higher in bakeries which were non-compliant in the first pass, *even though* those are precisely the bakeries which were treated.[15]

[15]Running an exact Fisher test for Table 4.2 allows us to reject the null hypothesis

Figure 4.1: Compliance rates



Persistence of non-compliance without (left) and with treatment (center), compared to the probability of non-compliance of a generic bakery in the first round (right).

Interestingly, the effect of the treatment seems to be stronger among bakeries which were the object of the preliminary investigation discussed at the beginning of Section 4.3 (and which were at the time non-compliant): in such a sub-sample, the estimation of the ATT is 36.36%. This seems to suggest that the request for the receipt has a stronger impact on bakeries which are *frequently* non-compliant. A plausible interpretation for this is that frequently non-compliant bakeries are precisely the ones where customers less frequently request for a receipt - and in which hence one request has a stronger impact.[16]

In order to assess if the estimated ATT is statistically significant, we run a one-sided test of $\mathcal{H}_1 : ATT > 0$ against the null hypothesis $\mathcal{H}_0 : ATT = 0$. Since the small sample makes asymptotic distributional assumptions unlikely, we refrain from using McNemar's test to reject the possibility of a type I error.[17] As suggested in Sheskin (2004), we instead observe that, assuming $\mathcal{H}_0$ holds, the number N_{01} of bakeries releasing the receipt only at the second pass is distributed according to a binomial:

$$N_{01} \sim \mathcal{B}(p, N_0)$$

that the first and the second pass are independent ($p = 0.003$), hence proving that fiscal behavior is characterized by persistence, *despite* our treatment effect.

[16]We thank an anonymous referee for this intuition.

[17]We also refrain from using a Fisher exact test, since we do not have a complete contingency table (we do not observe $\mathbb{P}\{c^2 = 0\}$).

where p is the probability of changing compliance status by pure chance (the *"natural"* switching rate), and N_0 is the total number of bakeries not releasing the receipt at the first pass. Denoting as $\hat{p}$ our estimate of p, the probability of a type I error is therefore calculated as:

$$\mathbb{P}\left\{N_{01} \geq 16 | N_{01} \sim \mathcal{B}(\hat{p}, N_0)\right\} = 7\%$$

(see Appendix 4.A for details). We are hence able to identify a causal effect of the treatment on the treated (ATT) at a significance level of $\alpha = 10\%$.

4.6 Discussion

In Section 4.1, we already referred to some of the channels, both psychological and purely utilitarian, through which social pressure could be affecting the compliance decisions. For instance, the seller may be *ashamed* of having received a rebuke. Alternatively, he may feel *embarrassed* by the discovery that he is acting *unjustly* (and possibly, that a sense of justice is more widespread than he used to think)[18]. These two different approaches can be seen as corresponding to the concepts of *collective* and *subjective* moral constraints studied in laboratory experiments by Bosco and Mittone (1997). Based on the structure of our experimental data, if the observed effect was related to *collective* moral constraints (*shame*) we would expect to find a higher impact of a rebuke when it is enacted in the presence of other clients. However such effect is not significant, possibly because our sample is not large enough to detect it, or because the number of clients is correlated with other characteristics of the shop, which influence the fiscal compliance. Our results are instead consistent with the idea that the intimate feeling of injustice plays a central role, an interpretation also supported by the experimental work of Bosco and Mittone (1997).

4.6.1 A social fiscal multiplier

Consider a client not receiving a receipt and asking for it. The *ATT* measures the effect of this event on the probability that, approximately 12 minutes later, another client receives a receipt. A more informative figure for policy implications would be the *number* of receipts which can be expected to be released, overall, as a consequence of that single request. In this paragraph,

[18]The seller could also expect that this goes hand in hand with an increase of fiscal controls on behalf of the authority.

we consider how through a back-of-the-envelope calculation it is possible to reconstruct an estimate for such number.

The number of receipts released by a bakery in the 12 minutes considered can be easily calculated - for bakeries which are compliant at the second pass - from the sequential number which is reported on each receipt:[19] in our sample, it was measured to be, on average, 5.5. Let η be the number of clients making a purchase in a given bakery in the 12 minutes after the rebuke, and recall from Section 4.5 that the observed rate of tax evasion 12 minutes after a request is 30.4%. Assuming that the correlation between compliance and the number of clients is negligible, we can write

$$\eta \cdot (1 - 30.4\%) = 5.5 \qquad \text{which gives} \qquad \eta = 7.9.$$

Now, if we assume that the effect of the rebuke is constant in time, then we expect that at least

$$\eta \cdot ATT \approx 1.38$$

additional receipts are released in the subsequent 12 minutes.

Given the approximations involved, such estimation should be considered only as an attempt in grasping the order of magnitude of the effect. In particular, there are at least two reasons why it could be downward biased. First, sales occurring *before* the 12th minute are expected to be affected *even more* by the rebuke (assuming its effect decreases with time).[20] Second, the calculation above ignores the possible effects on sales occurring *after* the 12th minute.

In principle, by adding 1 (the "direct" effect of the rebuke) to 1.38 we estimate hence a lower bound for the *social fiscal multiplier*, in the sense that each time a client asks for a receipt, at least 2.38 are released, on average, as a consequence. However, when considering a *general* effect, some further issues must be kept in mind. The effect of the treatment may be *local*: for instance, a seller who has been rebuked by a young client may, in the future, increase compliance when facing young clients. Moreover, the agents had no ambitions of representing the average client of a bakery in terms of observable characteristics and loyalty to the shop. While it is reasonable to think that a rebuke coming from a loyal customer will presumably have an *even higher* psychological impact on the seller, we have no way of predicting the effect of

[19] This is a legal obligation: the sequential numbers restart from 1 each day. We do verify that the correlation between the sequential number and the observed propensity to release the receipt is positive, but this clearly does not imply a correlation with the number of clients.

[20] This is true even when taking into account that it biases the reconstructed number of clients upwards.

other variables, such as age. This could be an interesting topic for further research. Finally, the sample under observation is not purely random, in the sense that 38 of the 108 bakeries were chosen for their non-compliant behavior during the January 2012 preliminary investigation. Further research could be devoted at studying the interplay between persistence of illegal behavior and reaction to social pressure.

It is also worth studying *to what extent* our findings can be generalized to countries with less widespread - at least public - fiscal evasion than Italy. An exhaustive answer to this question crucially depends on the relative importance of the different channels, considered in Section 4.1, through which social pressure can influence the fiscal behavior.

4.6.2 Robustness

When the client and the vendor are of the same gender, the probability of the receipt being released drops by 13.6% and the difference is statistically significant at the 5% level.[21] The explanatory power of this interaction variable is far higher than the effect of the mere gender of the client or of the vendor, which are non-significant (p-values of 30.4% and 62.21%, respectively).

In order to rule out the possibility that our main results concerning the treatment effect are driven by the higher frequency of coincidence of genders in the first pass, we disaggregate our data as shown in Table 4.3.

Table 4.3: Data disaggregated on coincidence of genders *in the first pass*.

	Compliance status			
	N_{00}	N_{01}	N_{10}	N_{11}
Coincidence	5	11	5	33
Non-coincidence	2	5	7	40

N_{ij}: compliance i (1 = compliant, 0 = non-compliant), at period 1 and j at period 2

In line with the effect just mentioned, we find that, in the "coincidence" case, $N_{01} > N_{10}$, while the opposite holds for the "non-coincidence" case.[22] In the absence of an effect of requests for the receipt, we would expect the magnitude of the two differences to be the same: instead, in the "coincidence" case it is three times higher than in the other one. This discrepancy is precisely what is expected in virtue of the treatment.

[21] These figures are calculated using data from both passes.

[22] Notice that the gender of the vendor in the first and in the second pass is generally unchanged (93.10% of cases): hence, if genders coincide in the first round they almost certainly differ in the second.

We compute the ATT also for different sub-samples. We find consistent results restricting the attention both to bakeries visited in the morning hours (54% of the sample) and to those visited in the afternoon: the ATT is always positive. Although it is higher in the morning (0.214) than in the afternoon (0.111), the difference is not statistically significant.[23] We find similar results when disaggregating on the (apparent) age of the vendor in the first pass:[24] the effect of the treatment is always positive, and no significant difference in its magnitude is found. When restricting to the sub-sample of bakeries which had already been visited in the January 2012 preliminary investigation (and had not, at the time, released the receipt), we obtain a p-value of 0.015; the results are not significant instead if we restrict to the few bakeries *not* visited in the preliminary pass.

4.7 Conclusions

We estimate, through a field experiment, the causal effect of social pressure on tax compliance of shop sellers. In our experiment, social pressure takes the form of a request for the receipt, made to bakery sellers who do not spontaneously release it (fiscal non-compliance was observed in at least 30% of the bakeries studied). Through the request, we manipulate the perception of the seller concerning the "common stand" of the Italian society towards fiscal evasion. Our results support the established hypothesis according to which *"compliance cannot be explained entirely by the level of enforcement"* (Torgler, 2002), but rather it also depends on behavioral factors. In particular, we are able to show that *direct* social pressure increases by 17.4% the propensity of sellers to release the receipt in the near future, and the result is significant at the 10% level. This finding also suggests the existence of a "social fiscal multiplier": every request for a receipt causes the seller to release approximately 1.38 additional ones.

We also find that the probability of receiving a receipt is significantly lower when a client is of the same gender than the seller (-13.6%, $\alpha = 5\%$). Since the gender of the agent was chosen independently of any characteristics of the bakery, the effect of the coincidence of genders has a causal interpretation. The explanatory power of this interaction variable is far larger than the effect of the mere gender of the client or of the vendor, which are non-significant. This finding suggests that *illegality feeds out of complicity*, the

[23] We also observe that the rate of compliance on the first pass is homogeneous across hours of the day.

[24] Vendors were recorded as "young" when they were attributed 30 years or less (this measure has clearly no ambition of absolute precision).

latter being scarcer when individuals belong to different social groups (in this case, defined by gender), and provides additional evidence in favor of Torgler's point of view presented above. A word of warning is however required: the experiment involved *only one* agent of each gender. Additional evidence based on experiments involving more actors would be required to confirm that what we observe is indeed a consequence of the gender matching, rather than of individual characteristics of the agents. Moreover, studies conducted on a larger scale could analyze the effect of the coincidence of genders on the observed ATT.

The policy implications of our study consist in a strong support for awareness campaigns and other instruments aimed at influencing the behavior of sellers through the strengthening of social norms and the diffusion of best practices.

Our experimental setting allows us to measure the *short-term* effect of social pressure. This is a unique feature among field experiments on fiscal compliance, which makes the results particularly interesting from the point of view of the psychology of compliance decisions. However, the design could easily be extended also to the study of longer term effects: experiments conducted with more than 2 agents acting consecutively, after predetermined intervals of time, may shed some light on the persistence of the effect of social pressure, a very relevant issue for policy implications. Further research could also be devoted to measuring the sensitivity of the results to changes in the location of the experiment. While we expect the results to be quite sensitive to the city or country where the experiment is run (the effect of peer pressure may for example depend upon the initial level of compliance), the combined results of studies coming from several towns could yield a more complete picture on the phenomenon. Other design choices worth experimenting with are the type of shop, and most importantly the characteristics of the agents.

Appendix 4.A Significance

Assuming $\mathcal{H}_0 : ATT = 0$ holds, N_{01} is distributed according to a binomial where the number of draws is equal to the total number of treated bakeries, and therefore to the number of bakeries not releasing the receipt at the first pass (N_0), while the probability of each bakery changing compliance status by pure chance is p:

$$N_{01} \sim \mathcal{B}(p, N_0).$$

In order to estimate

$$p = \frac{\mathbb{P}\{c_1 = 0, c_2 = 1\}}{\mathbb{P}\{c_1 = 0\}} = \frac{\mathbb{P}\{c_1 = 1, c_2 = 0\}}{\mathbb{P}\{c_1 = 0\}}$$

we use the known sample statistics:

$$\hat{p} = \frac{N_{10}}{N_0} = \frac{12}{23} = 0.52 \implies N_{01} \sim \mathcal{B}(0.52, 23).$$

Observing a value of $N_{01} = 16$, we can therefore calculate the probability of a type I error as:

$$\mathbb{P}\left\{N_{01} \geq 16 | N_{01} \sim \mathcal{B}(\hat{p}, N_0)\right\}.$$

The final estimate for the p-value is hence:

$$\sum_{k=N_{01}}^{N_0} \binom{N_0}{k} \hat{p}^k (1-\hat{p})^{N_0-k} = \sum_{k=16}^{23} \binom{23}{k} 0.52^k \cdot 0.48^{23-k} = 0.0707$$

corresponding to the blue area in Figure 4.2.

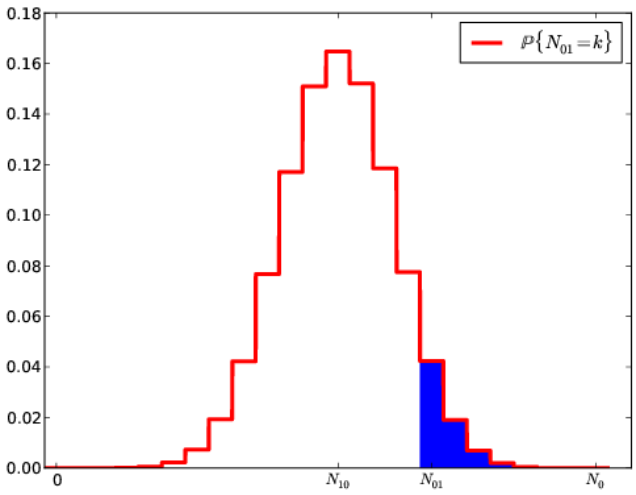


Figure 4.2: Probability distribution $\mathcal{B}(\hat{p}, N_0)$.

Appendix 4.B Reconstructing the DiD

We can trace back our experiment to a Difference in Differences model. We do not have a unified time break; instead, for every bakery we have a "before" (c^1) and an "after" (c^2). In the meanwhile, some bakeries are treated - that is, rebuked.

Let us start from the general form for the DiD:

$$\begin{aligned} ATT = &\big(\mathbb{E}\{c^2|D=1\} - \mathbb{E}\{c^1|D=1\}\big) \\ &- \big(\mathbb{E}\{c^2|D=0\} - \mathbb{E}\{c^1|D=0\}\big) \end{aligned} \tag{4.B.1}$$

Notice that, given two bakeries characterized by the same (unobservable) level of compliance, if the first releases the receipt at some time but the other does not, then the temporal evolution of fiscal compliance *will be different*, in the sense that the first can only become less compliant, while the second can only become more compliant.

This means that, for the Difference in Difference method to give unbiased results, we must imagine that also all the untreated bakeries refrained from releasing the receipt on the first pass.

Under this assumption, we have $\mathbb{E}\{c^1|D=0\} = 0$.

We instead lack an observational value for $\mathbb{E}\{c^2|D=0\}$; we can however resort to exploiting assumption (4.4.3), according to which $\mathbb{E}\{c^2(0)|D=0\} =$

p.

Given these premises, (4.B.1) will be evaluated as

$$\begin{aligned} ATT \approx & \left(\frac{N_{01}}{N_0} - 0 \right) - (\hat{p} - 0) \\ = & \frac{N_{01}}{N_0} - \hat{p} \end{aligned}$$

as above.

In general, the *selection bias* could be calculated as

$$\mathbb{E}\{c^1 | D = 1\} - \mathbb{E}\{c^1 | D = 0\}$$

which in our data would correspond to $0 - 1 = -1$. However, it is important to remark that this number sums up two different effects:

1. bakeries being treated have, on average, a lower *intrinsic* propensity to release receipts,

2. on top of that, among *occasionally* unlawful bakeries, the one being treated are "caught" in a *contingent* "non-complying moment".

Effect 1 is the one taken into consideration when deriving the ATT, since effect 2 does not affect the results from the second pass in any way.

Bibliography

Acemoglu, D., K. Bimpikis, and A. Ozdaglar (2010). Dynamics of information exchange in endogenous social networks. Technical report, National Bureau of Economic Research.

Acemoglu, D., M. A. Dahleh, I. Lobel, and A. Ozdaglar (2011). Bayesian learning in social networks. *The Review of Economic Studies 78*(4), 1201–1236.

Acemoglu, D. and A. Ozdaglar (2011). Opinion dynamics and learning in social networks. *Dynamic Games and Applications 1*(1), 3–49.

Acemoglu, D., A. Ozdaglar, and A. Tahbaz-Salehi (2013). Systemic risk and stability in financial networks.

Adamic, L. A., X. Wei, J. Yang, S. Gerrish, K. K. Nam, and G. S. Clarkson (2010). Individual focus and knowledge contribution. *CoRR abs/1002.0561*.

Aksnes, D. W. (2003). A macro study of self-citation. *Scientometrics 56*(2), 235–246.

Allen, F., A. Babus, and E. Carletti (2009). Financial crises: theory and evidence. *Annu. Rev. Financ. Econ. 1*(1), 97–116.

Allen, F., A. Babus, and E. Carletti (2010). Financial connections and systemic risk.

Allen, F., A. Babus, and E. Carletti (2012). Asset commonality, debt maturity and systemic risk. *Journal of Financial Economics 104*(3), 519–534.

Allen, F. and D. Gale (2000a). Bubbles and crises. *The economic journal 110*(460), 236–255.

Allen, F. and D. Gale (2000b). Financial contagion. *Journal of political economy 108*(1), 1–33.

Allen, F. and G. Gorton (1993). Churning bubbles. *The Review of Economic Studies 60*(4), 813–836.

Allingham, M. and A. Sandmo (1972). Income tax evasion: A theoretical analysis. *Journal of public economics 1*(3-4), 323–338.

Alm, J., G. H. McClelland, and W. D. Schulze (1992). Why do people pay taxes? *Journal of Public Economics 48*(1), 21–38.

Alm, J., G. H. McClelland, and W. D. Schulze (2007). Changing the social norm of tax compliance by voting. *Kyklos 52*(2), 141–171.

Andreoni, J., B. Erard, and J. Feinstein (1998). Tax compliance. *Journal of economic literature 36*(2), 818–860.

Asch, S. E. (1955). Opinions and social pressure. *Scientific American 193*(5), 31–35.

Azoulay, P., T. Stuart, and Y. Wang (2014). Matthew: Effect or Fable? *Management Science 60*(1), 92–109.

Baccini, A. (2010). *Valutare la ricerca scientifica*. Bologna: Il Mulino.

Baccini, A. and L. Barabesi (2010). Interlocking editorship. a network analysis of the links between economic journals. *Scientometrics 82*(2), 365–389.

Bala, V. and S. Goyal (1998). Learning from neighbours. *The Review of Economic Studies 65*(3), 595–621.

Bala, V. and S. Goyal (2000). A noncooperative model of network formation. *Econometrica 68*(5), 1181–1229.

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: A network-analytic model. *American Sociological Review 63*(6), 829–846.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013). The diffusion of microfinance. *Science 341*(6144).

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics 107*(3), 797–817.

Baños, R. A., J. Borge-Holthoefer, and Y. Moreno (2013). The role of hidden influentials in the diffusion of online information cascades. *arXiv preprint arXiv:1303.4629*.

Barabási, A.-L. and R. Albert (1999). Emergence of scaling in random networks. *science 286*(5439), 509–512.

Battiston, S., D. Delli Gatti, M. Gallegati, B. Greenwald, and J. E. Stiglitz (2012). Liaisons dangereuses: Increasing connectivity, risk sharing, and systemic risk. *Journal of Economic Dynamics and Control 36*(8), 1121–1141.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy 100*(5), 992–1026.

Blanchard, O. J. (1979). Speculative bubbles, crashes and rational expectations. *Economics Letters 3*(4), 387 – 389.

Blanchard, O. J. and M. W. Watson (1983). Bubbles, rational expectations and financial markets.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology 2*(1), 113–120.

Bordignon, M. (1993). A fairness approach to income tax evasion. *Journal of Public Economics 52*(3), 345–362.

Bosco, L. and L. Mittone (1997). Tax evasion and moral constraints: some experimental evidence. *Kyklos 50*(3), 297–324.

Brioschi, F., L. Buzzacchi, and M. G. Colombo (1989). Risk capital financing and the separation of ownership and control in business groups. *Journal of Banking & Finance 13*(4), 747–772.

Buechel, B., T. Hellmann, S. Klößner, et al. (2012). Opinion dynamics under conformity. *Institute of Mathematical Economics Working Paper* (469).

Cabrales, A., P. Gottardi, and F. Vega-Redondo (2013). Risk-sharing and contagion in networks. Working papers, EUI, Department of Economics.

Cainelli, G., M. A. Maggioni, T. E. Uberti, and A. De Felice (2010). The strength of strong ties: Co-authorship and productivity among italian economists. Working paper, Dipartimento di Scienze Economiche "Marco Fanno".

Campbell, P. (2005). Not-so-deep impact. *Nature 435*(77045), 1003–1004.

Corazzini, L., F. Pavesi, B. Petrovich, and L. Stanca (2012). Influential listeners: An experiment on persuasion bias in social networks. *European Economic Review 56*(6), 1276–1288.

Cowell, F. A. (1991). Tax-evasion experiments: an economist's view. In Webley (Ed.), *Tax Evasion: An Experimental Approach*, pp. 123–127. Cambridge University Press.

Cummings, R. G., J. Martinez-Vazquez, and M. McKee (2001). Cross cultural comparisions of tax compliance behavior. *International Studies Program Working Paper Series, at AYSPS, GSU*.

Cummings, R. G., J. Martinez-Vazquez, M. McKee, and B. Torgler (2006, December). Effects of tax morale on tax compliance: Experimental and survey evidence. Working paper series, Berkeley Olin Program in Law & Economics.

de Solla Price, D. (1965). Networks of scientific papers. *Science 149*, 510–515.

de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science 27*(5), 292–306.

De Stefano, D., V. Fuccella, M. P. Vitale, and S. Zaccarin (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association 69*(345), 118–121.

DeMarzo, P. M., D. Vayanos, and J. Zwiebel (2003). Persuasion bias, social influence, and unidimensional opinions. *The Quarterly Journal of Economics 118*(3), 909–968.

Diamond, D. W. and P. H. Dybvig (1983). Bank runs, deposit insurance, and liquidity. *The journal of political economy*, 401–419.

Diba, B. T. and H. I. Grossman (1988a). Explosive rational bubbles in stock prices? *The American Economic Review 78*(3), 520–530.

Diba, B. T. and H. I. Grossman (1988b). Rational inflationary bubbles. *Journal of Monetary Economics 21*(1), 35–46.

Diba, B. T. and H. I. Grossman (1988c). The theory of rational bubbles in stock prices. *The Economic Journal 98*(392), 746–754.

Dittmer, J. C. (2001). Consensus formation under bounded confidence. *Nonlinear Analysis-Theory Methods and Applications 47*(7), 4615–4622.

Dutta, B. and S. Mutuswami (1997). Stable networks. *Journal of Economic Theory 76*(2), 322 – 344.

Elffers, H., R. H. Weigel, and D. J. Hessing (1987). The consequences of different strategies for measuring tax evasion behavior. *Journal of Economic Psychology 8*(3), 311–337.

Elliott, M., B. Golub, and M. Jackson (2013). Financial networks and contagion. *Available at SSRN 2175056*.

Ellison, G. and D. Fudenberg (1993). Rules of thumb for social learning. *Journal of Political Economy*, 612–643.

Erard, B. et al. (2002). Compliance measurement and workload selection with operational audit data. In *Internal Revenue Service research conference, George Washington University, Washington, DC, June*, pp. 11–12.

Erard, B. and J. S. Feinstein (1994). Honesty and evasion in the tax compliance game. *The RAND Journal of Economics 5*(1), 1–19.

Falk, A. and U. Fischbacher (2002). "Crime" in the lab-detecting social interaction. *European Economic Review 46*(4), 859–869.

Falk, A. and A. Ichino (2005). Clean evidence on peer effects. *Journal of Labor Economics 24*(1), 39–57.

Fang, F. C. and A. Casadevall (2011). Retracted science and the retraction index. *Infection and immunity 79*(10), 3855–3859.

Fedenia, M., J. E. Hodder, and A. J. Triantis (1994). Cross-holdings: estimation issues, biases, and distortions. *Review of Financial Studies 7*(1), 61–96.

Fehr, E., U. Fischbacher, and S. Gächter (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human nature 13*(1), 1–25.

Fehr, E. and S. Gächter (1998). Reciprocity and economics: The economic implications of *Homo Reciprocans*. *European economic review 42*(3), 845–859.

Feiger, G. (1976). What is speculation? *The Quarterly Journal of Economics*, 677–687.

Feller, W. (2008). *An introduction to probability theory and its applications*, Volume 1. John Wiley & Sons.

Fellner, G., R. Sausgruber, and C. Traxler (2013). Testing enforcement strategies in the field: Threat, moral appeal and social information. *Journal of the European Economic Association 11*(3), 634–660.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics 10*(2), 171–178.

Fortin, B., G. Lacroix, and M.-C. Villeval (2007). Tax evasion and social interactions. *Journal of Public Economics 91*(11), 2089–2112.

Fortunato, S. (2004). Universality of the threshold for complete consensus for the opinion dynamics of deffuant et al. *International Journal of Modern Physics C 15*(09), 1301–1307.

French, Jr, J. R. (1956). A formal theory of social power. *Psychological review 63*(3), 181.

Friedkin, N. E. and E. C. Johnsen (1990). Social influence and opinions. *Journal of Mathematical Sociology 15*(3-4), 193–206.

Galbiati, R. and G. Zanella (2012). The tax evasion social multiplier: Evidence from italy. *Journal of Public Economics 96*(5), 485–494.

Gale, D. and S. Kariv (2003). Bayesian learning in social networks. *Games and Economic Behavior 45*(2), 329–346.

Galeotti, A. (2006). One-way flow networks: the role of heterogeneity. *Economic Theory 29*(1), 163–179.

Galeotti, A., S. Goyal, and J. Kamphorst (2006). Network formation with heterogeneous players. *Games and Economic Behavior 54*(2), 353 – 372.

Garfield, E. (1955). Citation indexes for science. a new dimension in documentation through association of ideas. *Science 122*(3159), 108–111.

Garfield, E. et al. (1965). Can citation indexing be automated? In *Statistical association methods for mechanized documentation, symposium proceedings*, pp. 189–192.

Gleeson, J. P., D. Cellai, J.-P. Onnela, M. A. Porter, and F. Reed-Tsochas (2013). A simple generative model of collective online behaviour. *arXiv preprint arXiv:1305.7440*.

Goeree, J. K., A. Riedl, and A. Ule (2009). In search of stars: Network formation among heterogeneous agents. *Games and Economic Behavior 67*(2), 445 – 466.

Golub, B. and M. O. Jackson (2010). Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics 2*(1), 112–149.

Gordon, J. (1989). Individual morality and reputation costs as deterrents to tax evasion. *European Economic Review 33*(4), 797–805.

Goyal, S., M. Van Der Leij, and J. Moraga-González (2006). Economics: An emerging small world. *Journal of Political Economy 114*(2), 403–412.

Graetz, M. and L. Wilde (1985). The economics of tax compliance: fact and fantasy. *National Tax Journal 38*, 355–363.

Grasmick, H. G. and R. J. Bursik Jr (1990). Conscience, significant others, and rational choice: Extending the deterrence model. *Law and Society Review 24*(3), 837–861.

Groenland, E. A. and G. M. Van Veldhoven (1983). Tax evasion behavior: A psychological framework. *Journal of Economic Psychology 3*(2), 129–144.

Gross, P. L. K. and E. Gross (1927). College libraries and chemical education. *Science*, 385–389.

Halla, M. (2012). Tax morale and compliance behavior: First evidence on a causal link. *The BE Journal of Economic Analysis & Policy 12*(1).

Haller, H. (2012). Network extension. *Mathematical Social Sciences 64*(2), 166–172.

Haller, H., J. Kamphorst, and S. Sarangi (2007). (non-) existence and scope of nash networks. *Economic Theory 31*(3), 597–604.

Harary, F. (1959). A criterion for unanimity in French's theory of social power. In D. Cartwright (Ed.), *Studies in social power*. University of Michigan.

Harrison, J. M. and D. M. Kreps (1978). Speculative investor behavior in a stock market with heterogeneous expectations. *The Quarterly Journal of Economics 92*(2), 323–336.

He, Z. and W. Xiong (2009). Dynamic bank runs. In *AFA 2010 Atlanta Meetings Paper, available at: ssrn. com/abstract*, Volume 1358672.

Hegselmann, R. and U. Krause (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation 5*(3).

Hegselmann, R. and U. Krause (2005). Opinion dynamics driven by various ways of averaging. *Computational Economics 25*(4), 381–405.

Herings, P. J.-J., A. Mauleon, and V. Vannetelbosch (2009). Farsightedly stable networks. *Games and Economic Behavior 67*(2), 526 – 541.

Jackson, M. O. (2010). *Social and economic networks*. Princeton University Press.

Jackson, M. O. and L. Yariv (2010). Diffusion, strategic interaction, and social structure. In J. Benhabib, A. Bisin, and M. O. Jackson (Eds.), *Handbook of Social Economics*. Elsevier.

Jadbabaie, A., P. Molavi, A. Sandroni, and A. Tahbaz-Salehi (2012). Non-bayesian social learning. *Games and Economic Behavior 76*(1), 210–225.

Jadbabaie, A., P. Molavi, and A. Tahbaz-Salehi (2013). Information heterogeneity and the speed of learning in social networks. *Working paper*.

Keynes, J. M. (1937). *The General Theory of Employment, Interest and Money*. Palgrave Macmillan.

Kidder, L. H., G. Bellettirie, and E. S. Cohn (1977). Secret ambitions and public performances: The effects of anonymity on reward allocations made by men and women. *Journal of Experimental Social Psychology 13*(1), 70–80.

Kirchler, E. (2007). *The economic psychology of tax behaviour*. Cambridge University Press.

Kirchsteiger, G., M. Mantovani, A. Mauleon, and V. Vannetelbosch (2011, February). Myopic or farsighted? an experiment on network formation. CEPR Discussion Papers 8263.

Kleven, H., M. Knudsen, C. Kreiner, S. Pedersen, and E. Saez (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica 79*(3), 651–692.

Kolm, S.-C. (1973). A note on optimum tax evasion. *Journal of Public Economics 2*(3), 265–270.

Kreps, D. M. (1977). A note on "fulfilled expectations" equilibria. *Journal of Economic Theory 14*(1), 32–43.

Labbé, C. (2010). Ike antkare one of the great stars in the scientific firmament. *International Society for Scientometrics and Informetrics Newsletter 6*(2), 48–52.

Lewis, A. (2011). The social psychology of taxation. *British Journal of Social Psychology 21*(2), 151–158.

Lux, T. and D. Sornette (2002). On rational bubbles and fat tails. *Journal of Money, Credit, and Banking 34*(3), 589–610.

Merton, R. K. (1968). The matthew effect in science. *Science 159*(3810), 56–63.

Minsky, H. P. (1992). The financial instability hypothesis.

Möbius, M., T. Phan, and A. Szeidl (2010). Treasure hunt. Technical report, working paper.

Moed, H. F. and E. Garfield (2004). In basic science the percentage of "authoritative" references decreases as bibliographies become shorter. *Scientometrics 60*(3), 295–303.

Olfati-Saber, R. and R. M. Murray (2004). Consensus problems in networks of agents with switching topology and time-delays. *Automatic Control, IEEE Transactions on 49*(9), 1520–1533.

Paternoster, R., L. E. Saltzman, G. P. Waldo, and T. G. Chiricos (1983). Perceived risk and social control: Do sanctions really deter? *Law and Society Review 17*(3), 457–479.

Perfect, H. and L. Mirsky (1965). The distribution of positive elements in doubly-stochastic matrices. *Journal of the London Mathematical Society 1*(1), 689–698.

Posner, E. A. (2000). Law and social norms: The case of tax compliance. *Virginia Law Review 86*(8), 1781–1819.

Pratelli, L., A. Baccini, L. Barabesi, and M. Marcheselli (2012). Statistical analysis of the hirsch index. *Scandinavian Journal of Statistics 39*(4), 681–694.

Radicchi, F., S. Fortunato, B. Markines, and A. Vespignani (2009, Nov). Diffusion of scientific credits and the ranking of scientists. *Physical Review E 80*, 056103.

Redner, S. (1998). How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems 4*(2), 131–134.

Reis, H. T. and J. Gruzen (1976). On mediating equity, equality, and self-interest: The role of self-presentation in social exchange. *Journal of Experimental Social Psychology 12*(5), 487–503.

Schwartz, R. D. and S. Orleans (1967). On legal sanctions. *The University of Chicago Law Review 34*(2), 274–300.

Sheskin, D. (2004). *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall/CRC.

Singh, B. (1973). Making honesty the best policy. *Journal of Public Economics 2*(3), 257–263.

Slemrod, J. (2007). Cheating ourselves: The economics of tax evasion. *Journal of Economic Perspectives 21*(1), 25–48.

Slemrod, J., M. Blumenthal, and C. Christian (2001). Taxpayer response to an increased probability of audit: evidence from a controlled experiment in Minnesota. *Journal of Public Economics 79*(3), 455–483.

Smith, L. and P. Sørensen (2000). Pathological outcomes of observational learning. *Econometrica 68*(2), 371–398.

Sridharan, M. and K. Parthasarathy (1972). Isographs and oriented isographs. *Journal of Combinatorial Theory, Series B 13*(2), 99–111.

Tirole, J. (1982). On the possibility of speculation under rational expectations. *Econometrica*, 1163–1181.

Torgler, B. (2002). Speaking to theorists and searching for facts: Tax morale and tax compliance in experiments. *Journal of Economic Surveys 16*(5), 657–683.

Wärneryd, K.-E. and B. Walerud (1982). Taxes and economic behavior: Some interview data on tax evasion in Sweden. *Journal of Economic Psychology 2*(3), 187–211.

Weigel, R. H., D. J. Hessing, and H. Elffers (1987). Tax evasion research: A critical appraisal and theoretical model. *Journal of Economic Psychology 8*(2), 215–235.

Wenzel, M. (2001). Misperceptions of social norms about tax compliance (2): A field-experiment. *Centre for Tax System Integrity Working Paper*.

West, J., T. Bergstrom, and C. Bergstrom (2010). The eigenfactor metrics$^{\text{TM}}$: A network approach to assessing scholarly journals. *College & Research Libraries 71*(3), 236–244.

Zawadowski, A. (2013). Entangled financial systems. *Review of Financial Studies 26*(5), 1291–1323.

List of Figures

List of Tables

Thanks: to my family for supporting me no matter what; to Luca Stanca, for teaching me to (almost) never be satisfied; to Giulia for (almost) never being satisfied; to Simona and Lorenzo for making it fun to study, even DSGE models; to Lia, Valentina and all the friends who shared or tolerated the idea that "vacations together" has a different meaning when there is a deadline.