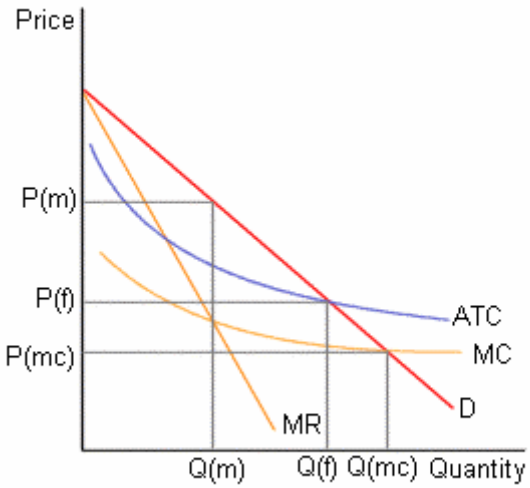
Introduction

In microeconomics, industrial organization, and public economics handbooks[1], natural monopoly is described as a situation in which, for structural reasons, only one firm finds it profitable to produce in the market; the diagrams used are similar to the following.



In the first edition of Samuelson's handbook (1948), and until the end of the 1970s, natural monopoly was considered to occur in cases of scale economies[2]; as we know, with these average costs decline when output increases throughout the entire range of market demand. In this literature, economies of scale were attributed to high fixed costs and low or zero variable costs; they were also considered a barrier to entry. Due to the monopoly power derived from it, natural monopoly was seen as a market failure, and Government intervention was required (in the forms of nationalization, regulation, or antitrust)[3].

This theory has been criticized. For example, part of the Austrian monopoly theory denies the existence of non-legal barriers to entry[4]. However, the most influential criticism came from a number of articles published in the 1970s – dealing with natural monopoly in

[1] For Microeconomics see Kreps (1990: 302); for Industrial Organization see Cabral (2000: 75), for Public Economics see Stiglitz (2000: 291).
[2] "Some of the basic factors responsible for monopoly are inherent in the economies of large-scale production" (Samuelson 1948: 40).
[3] See the interesting contribution on the history of the treatment of natural monopoly in introductory textbooks by Ulbrich (1991).
[4] Actually, monopoly theory is one of the most controversial areas in Austrian economics. In short, there are three different views: 1. Mises and Kirzner; 2. Rothbard and Armentano; 3. O'Driscoll. Although this is not the place to analyze these views, we can say that not all of them deny the existence of natural monopoly.

multiple output production[5] – and from the theory of "contestable markets"[6]. They caused a radical change in the definition of natural monopoly. Nowadays it means a situation characterized by the sub-additivity of cost functions (production costs less if it is done by one firm only), and by sustainability (entry is not profitable). Not only does this new theory demonstrate that scale economies do not help us to define natural monopoly properly, but also that they alone may not constitute a barrier to entry[7]. It implies that "if there is a free entry, this will prevent the monopolist from setting high prices, as they would trigger entry" (Motta 2004: 70). The systematization of the new viewpoint on natural monopoly was carried out by Sharkey (1982).

But apparently the story does not end here. There have also been strong criticisms of the new theory. Shepherd, for example, radically claims that the theory of contestable markets "has been mainly a detour" (1995: 299)[8]. In short, we can summarize the present state of research as follows: if there are economies of scale with sunk costs, as in the cases previously considered as natural monopolies, the behavior of the potential entrant and the reaction of the incumbent are now seen as depending on the strategic context in which they operate. In general, recent models show that monopoly power might persist under free entry, and market mechanisms might not prevent a monopolist from exercising market power (Motta 2004: 2.6)

The purpose of this article is to begin writing the history of the concept of natural monopoly, and of its policy implications. While it would certainly be very interesting to inquire into the reasons and the consequences of the change that this concept underwent in the 1970s, this paper focuses on the reconstruction of its origins, long before Samuelson's first edition. Very few studies have been devoted to this topic. We have found some synthetic reconstruction of the initial history of natural monopoly in Sharkey (1982, ch. 2)[9], Hazlett (1985), and DiLorenzo (1996); there are hints in Ekelund and Hébert (1981), O'Driscoll (1982) and Stigler (1982); we can also cite Ekelund and Hébert (2003),

[5] The question of multiproduct natural monopoly was dealt with from 1977 by Baumol, Bayley, Panzar, and Willig.

[6] The idea of contestability was dealt with in a series of articles from 1980 by Baumol, Bayley, Panzar, and Willig. See for all Baumol, Panzar and Willig (1982).

[7] There are actually two definition of entry barriers in the literature, one proposed by Bain, the other by Stigler. As Schmalensee points out "An updated Bain definition would not rule out scale economies as an antitrust barrier to entry when sunk costs are important, while the Stigler definition would" (2004: 471).

[8] And he adds: "The theory is internally inconsistent, difficult to relate to reality, and hazardous for policy treatments of market power" (Shepherd 1995: 300).

who deals with the theory of Dupuit and of Walras, and Béraud (2004), mainly focused on Walras[10]. The secondary literature is therefore rather scarce.

But the main reason why we think it is worthwhile to make a new contribution to it doesn't lie in its scarcity; it lies rather in a source of confusion that we think needs to be revised. The confusion derives from the fact that the concept of natural monopoly is composed of different elements and, as is shown in this article, every element has its own different history. We are therefore dealing with a complex concept, requiring a separate analysis of the particular paths followed by its various components. In other words, we think that an accurate historical analysis of the notion of natural monopoly cannot be written without breaking it down into all its component parts. In this paper we have identified the following features which go to make up that notion: 1. the expression itself; 2. the singling out of the concrete situations to which it is applied; 3. the inquiry into economies of scale; 4. the consideration of their compatibility with competition; 5. the drawing of the diagram; 6. the request for Government intervention. If each of these elements is not considered separately, it is hard to correctly identify original contributions, and to properly reconstruct the influence of ideas. Hence, in every section of the paper, each of the above features is separately examined from a historical perspective, highlighting the originality of economic theories in that specific respect, as well as the way those theories influenced one another.

The approach mostly followed by this paper is known as "rational reconstruction": we extract from the whole of the economists' work those parts concerning the different elements composing the traditional notion of natural monopoly, with the aim of finding out priorities and influences. We are aware of the limits of this perspective, and we sometimes suggest some interpretation and contextualization, but our main purpose here is to take only a first step, clarifying the confusion we mentioned above, and thus providing a sound basis for a further "historical reconstruction", which will be our next task.

Our investigation ends with the formulation of the concept of natural monopoly as it was in the traditional view, although the new developments of the theory after the 1970s

[9] Sharkey (1982: 15) also cites Lowry (1973).
[10] All these works will be recalled later in the paper.

will be taken into account during the examination of the economists' thought, and some reflections related to the new view will be also discussed in the conclusion.

1. The expression "natural" monopoly

Aristotle was the first to talk about monopoly (De Roover 1951: 492; Langholm 2006: 397), but who was the first to talk about "natural" monopoly? When did this expression start being used in its current sense? And why was the word "natural" employed? This is what we look at in this section. We limit our analysis here to the meaning and the definition of natural monopoly; the identification of its distinctive features by the economists under analysis will be discussed in the next sections.

Smith never uses the expression "natural monopoly", but he gives a detailed description of the characteristics of what this was to be called immediately after him: "Some natural productions require such a singularity of soil and situation, that all the land in a great country … may not be sufficient to supply the effectual demand"; the consequent "enhancements of the market price are evidently the effect of natural causes which may hinder the effectual demand from being fully supplied, and which may continue, therefore, to operate forever"(1776: I.7.24). The earliest explicit use of the term that I have found in the literature is in the essay *The Nature of Rent* by Malthus, where natural monopoly is distinguished from artificial monopoly. For Malthus there are: "peculiar products of the earth … which may be called natural and necessary monopolies" ([1815] 1969: 13). As an example of natural monopolies, he takes "certain vineyards in France, which, from the peculiarity of their soil and situation, exclusively yield wine of a certain flavour" (13-14)[11]. The expression turns up again in Bastiat, who wrote: "People who class together artificial monopoly and what they call natural monopoly … are quite blind or quite superficial" ([1850] 1864: 180). So far, we have seen that classical economists did use the expression natural monopoly, but we haven't found a definition of the concept yet. This is supplied by J.S. Mill, who explains that natural monopolies are "those which are created by circumstances, and not by law" ([1848] 149: 499). In general we can say that the expression was used to indicate those cases of monopoly deriving from natural agents supplied in fixed quantity, also including talent and location (Cairnes 1861). Economists always regarded them favorably. Hence, we think that the reason for the use of the