WHY IS FINANCE IMPORTANT? SOME THOUGHTS ON POST-CRISIS ECONOMICS*

YEW-KWANG NG

Winsemius Chair Professor of Economics
Nanyang Technological University
14 Nanyang Drive, Singapore 637332
ykng@ntu.edu.sg
http://www.ntu.edu.sg/home/ykng/

Published 8 October 2014

The global financial crisis of 2008 challenges some relevant aspects of macroeconomic theory such as the neutrality of money. This paper shows that this neutrality is based on the unrealistic assumption of perfect competition. Relaxing this alone (without time lags, price rigidities, menu costs and other frictions) makes money no longer necessarily neutral and hence makes financial crises and institutions much more important. The presence of increasing returns to scale at the firm level and to specialization at the economy level due to the division of labor also makes finance much more important than suggested by traditional economics.

Keywords: Finance; financial institutions; financial crisis; money supply; macroeconomics.

JEL Classification: G00, E00, D4

1. Introduction

The global financial crisis around 2008 has raised many issues including the relevance of orthodox economics (e.g., Phelps, 2010 and references cited below), not to mention the important effects on real output (e.g., Furceri and Mourougane, 2012). The issues involved are rather complicated and likely include many aspects: Psychological or behavioral (e.g., Akerlof and Shiller, 2009), leverage (Gourinchas and Obstfeld, 2012; Geanakoplos, 2014), systematic and institutional (Bezemer, 2011), globalization (Mendoza and Quadrini, 2010), regulatory (Cabral, 2013), liquidity/financial-constraint aspects (Uhlig, 2010; Cespa and Foucault, 2014), and beyond. This paper is not ambitious enough to tackle all related important aspects. Rather, it compares the orthodox result of the neutrality of money, which largely implies the unimportance of finance in general and of financial crises in particular, with their obvious importance in the real world. Our economies have evolved

*This is a revised version of an invited paper presented at the Singapore Economic Review Conference held on August 2013. Comments from participants and others are gratefully acknowledged.
into much more complex systems than reflected by the orthodox theory. For one thing, the
neutral result is based on the simple industrial organizational assumption that firms are
all perfectly competitive. Perfect competition rules out product differentiation and in-
creasing returns to scale. The omnipresence of these makes perfectly competitive firms
almost non-existent in the real world.

Section 1 below shows that the relaxation of perfect competition alone does away with
the neutrality of money, making financial crises, changes in money supply and other
changes in nominal aggregate demand possibly important for determining real economic
variables, both for the short and long runs. This is so because the behavior of a non-
perfectly competitive firm is very different from that of a perfectly competitive one. It also
explains why this is so using the simple diagrams of undergraduate microeconomics of
profit maximization. Though economists are familiar with other factors making finance
and financial crises important in the real world, this particular factor has been largely
ignored. Section 2 shows that the presence of increasing returns [both at the firm level due
to scale and at the economy level due to the economies of specialization from the division
of labor] makes entrepreneurship and financial institutions much more important than
suggested by the orthodox economic theory. The extension of economic analysis beyond
its traditional core in these directions thus suggests that finance is much more important.
Of course, the importance of finance due to the factors discussed in this paper does not rule
out other factors, which are however beyond the scope of this paper, that may also make
finance important.

Some analyses of macroeconomics with imperfect competition or market power have
in fact long been undertaken. However, they still reach the conclusion that, just the
presence of imperfection competition alone does not destroy the neutrality of money in
the absence of additional frictions like price rigidity and menu costs (e.g., Blanchard and
Kiyotaki, 1987; Dixon and Rankin, 1994). As shown below, this is not the case; even
without price rigidity and menu costs or other frictions, just the presence of market
power alone is sufficient to destroy the neutrality of money. The different conclusions are
due to the fact that these other economists implicitly assume a unique equilibrium. They
show that with an initial real equilibrium $E_1$ and money supply $M_1$, a change in money
supply to $M_2$, the original real equilibrium $E_1$ may still be an equilibrium. They then
conclude that money is thus neutral even under imperfect competition. However, as
argued below, the change in money supply and the corresponding change in nominal
aggregate demand (or the latter change due to other causes such as a financial crisis)
may shift the economy from the original equilibrium (which could still be an equilibrium
but need not actually prevail) to a different real equilibrium, making money and changes
in nominal aggregate demand possibly non-neutral. Orthodox economists recognize the
short-run non-neutrality in the presence of time lags, rigidities and/or frictions, but fail to
see that non-neutrality may prevail both in the short and long runs even in the absence of
these lags and rigidities. Our results here are thus also more consistent with the pro-
longed and large recessions in many economies in the world, European countries in
particular.
2. Why Does Non-Perfect Competition Itself Makes Money Possibly Non-Neutral?

Finance includes the use of money and the lending and borrowing of funds. It is an essential factor in the life or business of individuals, firms, banks and other organizations. Its importance seems obvious in a modern economy. However, if we look at the core of modern economic theory, the importance of finance is very limited. For one thing, a central result in modern economics is the neutrality of money. This says that, ignoring possibly very short-term effects due to time lags, money supply affects only the price level, not real economic variables like aggregate and relative output levels, employment levels and relative prices. Associated with this neutrality is the classical dichotomy between the real economic sector and the nominal monetary/financial sector, making these two sectors independent of each other.¹ If things in the financial sector do not affect the real sector, the importance of finance and financial institutions will be greatly reduced. In particular, we should not be too worried about financial crises since they should not be able to cause big reductions in output and cause mass unemployment. However, in the real world, these possible effects cannot be ignored. What cause the big difference between theory and the real world?

If we examine economic models that yield the neutrality (of money) result, they invariably involve the explicit or implicit industrial organizational assumption that firms are perfectly competitive. The reason that, together with other simplifying assumptions such as no time lags and no money illusion, the assumption of perfect competition leads to the neutrality of money may be illustrated with the simple diagram of Figure 1. As all firms are assumed to be perfectly competitive, their situation may be illustrated by one representative, one facing a horizontal demand curve for its product at the market equilibrium price \( p \).

![Figure 1. Perfect Competition Leads to Neutrality of Money](image)

¹The neutrality of money and the classical dichotomy goes together in simple models but in more complicated models, they need not always go together.
It can sell any quantity it likes at this price which thus is also its marginal revenue curve. It maximizes its profit by producing and selling at the point $q$ where its upward-sloping marginal-cost curve cuts the horizontal demand curve. At this equilibrium point, the firm does not want to sell more. (One will be hard put to find firms in the real word that do not want to sell more at the same prices. Most firms like to sell more; many incur heavy advertising/promotion costs to sell more. They do not sell more only because they have to decrease prices to sell more. If so, they face downward-sloping demand curves. This shows the rather unrealistic nature of perfect competition.) Confining first to the short run (defined by the given number of firms), the output of the firm (after multiplication with the given number of firms) also indicates the aggregate output of the economy and the price indicates the price level.

Now consider an increase in money supply that increases the nominal aggregate demand. Since firms do not want to sell more at the same prices, this higher demand pushes prices higher. For the firm illustrated in Figure 1, its demand curve moves upward to $d'$. As this firm represents all firms in the whole economy, this means that the price level $P$ also increases by the same extent. In the absence of time lags, this increases the costs faced by the firm by the same proportion, shifting its marginal-cost curve from $MC$ to $MC'$. This is so since labor supply is in accordance to real instead of money wage rates (no money illusion). Thus, the new profit-maximization equilibrium occurs at the same vertical line or the same output level. The output level is unchanged; only prices increase. Money is neutral. This is the microeconomic foundation for the macroeconomic result of the neutrality of money.

Even if we retain all other simplifying assumptions like no time lags and no money illusion, the relaxation of perfect competition alone destroys the (necessary) neutrality of money. Money may still be neutral but is no longer necessarily neutral. That money may still be neutral with non-perfect competition is illustrated in Figure 2. (Ignore the two AC

![Figure 2. Money May Still Be Neutral Under Non-Perfect Competition](image-url)
Using “non-perfect competition” is preferred to “imperfect competition” as the former may subsume other market structures like monopolies and oligopolies. The representative firm\(^2\) is initially at the equilibrium point \(A\) with MC cutting MR from below. An increase in nominal aggregate demand (as may be the result of an increase in money supply) may just shift its demand curve proportionately vertically upward, as illustrated by the demand curve \(d'\) in Figure 2. The new equilibrium at \(B\) then involves no change in output and an increase in price. Taking account of the conditions of cost and demand conditions of firms in the whole economy and their interaction, the mathematical Appendix A shows more rigorously that there are plausible conditions where this or other cases can prevail and persist as the new equilibrium.

However, for this case of non-perfect competition, the opposite (to the above Monetarist case) Keynesian case of no change in price but an increase in output following an increase in nominal aggregate demand is also possible, as illustrated in Figure 3. Here, the firm is originally at equilibrium at point \(A\). With an increase in nominal aggregate demand, the demand curve, instead of shifting proportionately vertically upward, shifts horizontally rightward and becomes more price elastic at the same price. This makes the value of marginal revenue at the same price (point \(B\) at the new demand curve \(d'\)) higher, allowing it to intersect the higher marginal-cost curve \(MC'\) at a higher output level \(q'\). The case of a

\[\text{Figure 3. Money May Be Non-Neutral Under Non-Perfect Competition}\]

\(^2\)Using a fully general-equilibrium analysis, Ng (1986, Appendix 3I) shows: (1) The hypothetical existence of a representative firm whose changes in output and price exactly represent those of the whole economy in aggregate output and the average price for any given exogenous change; (2) The actual existence of a representative firm, defined by a simple weighted average of all (or a representative sample) of firms, whose changes in output and price approximately represent those of the whole economy for any exogenous change that does not result in drastic changes in relative prices.
proportionately rightward shift in the demand curve with no change in the price elasticity of demand at $B$ (same price elasticity at $B$ than at $A$) is also possible. However, for this case to be a sustainable new equilibrium with no increase in price, the value of the new marginal cost at $q'$ must not be higher than the original marginal cost at $q$. If the marginal-cost curve does shift upward in response to a higher aggregate output level (as the case illustrated in Figure 3), the marginal-cost curve has to be downward-sloping to offset the positive shift in the whole curve, to ensure that the price does not increase. (A higher aggregate output may push up input prices and hence affect the costs of firms.)

If this offset just mentioned above is only partial so that there is some increase in price, the possibility of a partial increase in output, as may be taken to be possible using just a graphical analysis, is misleading. The increase in price of the representative firm signifies an increase in the general price level. This will shift the demand curve of the firm vertically upward over the relevant range, leading to further increases in price. A final equilibrium is reached only at the unchanged output level with only an increase in price. (To see the complete picture more rigorously, see the mathematical Appendix A.) However, if the cost responses and price elasticity situation are consistent with no increase in price, such as the case illustrated in Figure 3, the Keynesian case of no increase in price but only an increase in output is possible. Our analysis is a general-equilibrium one (though somewhat simplified to abstract away changes in relative prices due to the use of the representative-firm methodology; see footnote 2) taking account of the responses in both the cost side (effects of changes in aggregate output and the price level on the costs of the firm) and demand side (effects of changes in aggregate output, aggregate income/demand and the price level on the demand of the firm), as well as the interaction of the firm with other firms as a whole.

In fact, a case more Keynesian than the Keynesian case is also possible. This is the case of cumulative expansion/contraction. In Figure 3, if $MC'$ intersects $MR'$ to the right of $q'$, the new profit-maximization price is actually reduced. This reduction in $p$ and hence the price level will shift the demand curve of the firm downward, leading to a further decrease in the profit-maximization price and an increase in the output level. Such a cumulative expansion cannot of course go on forever. When firms have no excess capacity, its $MC$ slopes sharply upward and when the economy is approaching full employment the costs of firms respond sharply to further increases in aggregate output, making the condition for cumulative expansion or even the Keynesian case no longer applicable. However, a decrease in nominal aggregate demand may trigger a cumulative contraction. Unless input prices fall sufficiently in response to a fall in aggregate output, the cumulative contraction may be quite deep, explaining why depressions are possible.

It may be thought that the possibility of the Keynesian result of an increase in output with no change in the price level following an increase in nominal aggregate demand is only possible in the short run. For the long run, we allow free entry and exit and treat the number of firms as a variable. Adding the zero-profit condition, our analysis still shows that the various cases including both the Monetarist and Keynesian cases are still possible. The Monetarist case is already illustrated in Figure 2 with the two AC curves brought back into the picture. The Keynesian case is illustrated in Figure 4. The higher nominal aggregate demand shifts the equilibrium from point $A$ to $B$ with no change in price, only an
increase in output. This is possible even in the long run provided that the increase in the marginal cost of the firm from both its own output (from the slope of the MC curve) and the increase in aggregate output (an upward shift of MC to MC') is offset by an increase in the price elasticity of demand (at the point B in comparison to at A due to a higher degree of competition as the number of firms has increased), and the upward shift in the average-cost curve from AC to AC' (from an increase in aggregate output) is offset by the downward-sloping average-cost curve.

Is the Keynesian case more or less likely to prevail in the long run, in comparison to the short run? Our analysis suggests two opposing considerations. On the one hand, the cost curves of the firm are more likely to respond positively to the aggregate output and hence shift upward more significantly in the long run as aggregate output increases. This suggests that the Monetarist case is more likely to prevail. On the other hand, the entry of new firms (as aggregate demand increases) increases the degree of competition and makes the price elasticity of demand for the product of each firm higher in absolute value, making it less likely to require an increase in price. This makes the Keynesian case more likely in the long run.

What makes the huge difference between the case of perfect competition where money must be neutral (ignoring time lags, etc.) illustrated in Figure 1 and the case of non-perfect competition where both the Monetarist case of neutrality (Figure 2) and the Keynesian case of effectiveness (Figure 3) are possible may be explained. The crux of the difference consists in both the demand side and the cost side. On the demand side, a horizontal demand curve (as necessitated by perfect competition) cannot shift leftward or rightward, it can only shift upward or downward. However, as we are not analyzing the case of a single firm, but a firm representing the whole economy, an upward or downward shift in the demand curve or price line for the firm signifies a corresponding change in the price level.

Figure 4. Non-Neutrality of Money Even in the Long Run
In the absence of time lags, etc., a change in the price level changes the cost (including the marginal cost) level of the firm by the same proportion, dictating no change in real output but only a change in price, as illustrated in Figure 1. In contrast, for the case of non-perfect competition, the demand curve for the product of the firm is downward sloping. A downward-sloping demand curve may either shift (proportionately) upward and downward (with its intersection point with the horizontal axis, if any, unchanged); it may also shift rightward and leftward (with its intersection point with the vertical axis, if any, unchanged), or some combination of the two or even some irregular shifts. If the demand curve shifts vertically upward as in Figure 2, it favors an increase in price with no change in output; if the demand curve shifts horizontally rightward, it favors an increase in output with no change in price as in Figure 3. Thus, both the Monetarist and Keynesian cases are possible under non-perfect competition.

Which shift in the demand curve will be the case is not arbitrary but is determined by what will happen to the price level that is consistent also with the new cost condition in our simplified general-equilibrium analysis. Roughly speaking, if the new demand and cost situation necessitates a higher price level, it leads to a vertical shift; if the new demand and cost situation allows a higher output level with no change in the price level, it leads to a rightward shift. One further complication is that, in the former case there are only nominal changes and the homogeneity (of degree zero) nature of demand functions prevails, making the vertical shift perfectly proportionate. In the latter case, real output and hence real income levels change and hence may cause a change in the price elasticity of demand (see Appendix A).³

Another crucial difference between perfect and non-perfect competition concerns the cost side. As the demand curve for the product of a perfectly competitive firm is horizontal, a determinate profit-maximization equilibrium requires that its marginal-cost curve must be upward sloping. This means that the value of marginal cost will increase if output increases, making an increase in output with no increase in price impossible. On the other hand, for a non-perfectly competitive firm, the demand curve is downward sloping, making the marginal-revenue curve usually even more downward sloping. A determinate profit-maximization equilibrium is then consistent with either an upward, horizontal or even downward-sloping marginal-cost curve. This means that an increase in output need not necessarily increase the marginal cost, making the Keynesian case possible to prevail. This Keynesian case is only possible but not necessarily the case, as the effect through the change in aggregate output on the costs of firms must also be allowed for.

In addition to non-perfect competition, if we have increasing returns both at the firm, industry and at the economy levels, this makes the marginal-cost curves of firms more likely to be non-upward sloping and the cost curves of firms less likely to shift upward significantly in response to an increase in aggregate output. This makes the non-Monetarist cases⁴ more likely to prevail. The next section discusses this as well as the higher importance of finance in such situations.

---

³ If \((\eta'^\alpha + \eta'^\beta - D)\) and \((1 - \eta'^\alpha)\) are both zero, our equations show indeterminate results. Further graphical analysis and common sense then suggest that the results then depend on expectation which will then be self-fulfilling.

⁴ This possible change is abstracted away in Ng (1982).
3. Increasing Returns to Scale and Specialization and the Importance of Finance

If the real world is characterized by the neutrality of money where changes in nominal aggregate demand do not affect real output and employment, then money, finance and financial crises should not be important, as they affect only nominal variables. The horror of a depression with mass unemployment will not happen. Unfortunately, we do not live in such a world. Even before considering such complications as asset price changes under nominal rigidities, herd psychology, time lags, etc., the previous section shows that the relaxation of the unrealistic assumption of perfect competition alone changes the picture. A change in nominal aggregate demand (including that from a change in money supply) may cause changes in real output and employment. This makes money, finance and financial crises much more important. A financial crisis typically involves large falls in asset prices and aggregate demand. As these may lead to large and even cumulative fall in output and hence employment, it is not surprising that it is treated as very important in the real world. This also explains why big measures of quantitative easing of money supply and expansionary fiscal policies may be necessary upon the wake of a financial crisis.

The importance of financial matters discussed in the previous paragraph is related to non-perfect competition. There are two important factors making most firms in the real world non-perfectly competitive. On the one hand, product differentiation (including in location if not in other aspects of the product) makes demand curves downward sloping. On the other hand, the almost omnipresence of increasing returns to scale makes the average-cost curves of firms downward-sloping and their marginal-cost curves not necessarily upward sloping. A very common situation is the presence of a large fixed cost. Modern economies consist largely of services including commerce and personal services. Shops/offices providing such services have to be large enough so that customers may come in to inspect goods or to be serviced. Thus a relatively large fixed cost is incurred even if only one unit of the service is to be provided. Up to the full capacity level (when the marginal cost may increase sharply), the marginal cost of providing additional units is typically relatively low and constant. This makes the case of a horizontal marginal-cost curve with a sharply decreasing average-cost curve rectangular to the vertical axis and the horizontal MC curve very prevalent and is analyzed by the well-known Dixit-Stiglitz model (Dixit and Stiglitz, 1977). As the demand curve is downward sloping from product differentiation, long-run free entry entails the tangency of the demand curve with the AC curve at the downward sloping section, making the full capacity, upward-sloping MC not within the relevant range. The prevalence of this situation not only partly favors the possible prevalence of the non-Monetarist results, but also makes finance more important, as explained below.

If we use the perfect-competition model with Figure 1 for the firm level and the supply-demand analysis at the industry level, there is little if any need for finance. The equilibrium is determined by the intersection of the (upward-sloping) supply and (downward-sloping) industrial demand curves, with thousands of small firms each producing a minute fraction of the market output.
The situation is even more contrasting when we consider the introduction of new products. If the supply curve is wholly above the demand curve, the product is not supplied and the situation is optimal. As long as the demand curve moves up and/or the supply curve moves down by just enough to allow an intersection point with any positive output, that optimal level of output could be produced and supplied, even if the quantity transacted is small. As the supply price is below the demand price, positive profits could be made or at least no losses need be incurred; no problem.

Now consider the more realistic situation where a firm is typically faced with a downward-sloping demand curve and downward-sloping AC curve. If the AC curve is completely above the demand curve, production is not feasible. When either the demand curve moves up or the AC curve moves down by sufficiently to yield a tangency or an intersection range, some production is feasible and desirable. However, for this production to be profitable (at least not loss-incurring), the quantity transacted has to be sufficiently high. If we take into account the real-world factors like time lags (including the time taken for consumers to know the availability of a new product/firm), a firm typically cannot sell a large amount of its product when it just first introduces it to the market (unless it incurs large advertising/promotion costs which require larger sales to recover). The presence of a large amount of fixed cost or other factors causing increasing returns or downward-sloping AC curve makes the firm typically making large losses in its commencing periods. Using Figure 2 but ignoring all curves with primes, the quantity demanded has to be at least equal to \( q^o \) at a price equals \( Aq^o \) for the firm not to make losses. At lower quantities demanded due to time lags, the firm makes large amount of losses. This requires significant amount of financing to allow the firm to overcome the initial periods of loss making even for those cases where firms will eventually be successful and profitable (at least not loss making) and will be producing large amounts of surpluses to consumers. This increases the role of finance or financial intermediaries substantially in comparison of the supply-demand analysis of the previous paragraph.

Increasing returns may also operate at the economy level, particularly through the division of labor that facilitates economies of specialization. This form of increasing returns was emphasized by classical economists particularly Adam Smith as the very crucial factor that causes the wealth of nations. Modern emphases on the importance of this include those from Young (1928), Buchanan and Yoon (1994) and Sun (2005). A most elaborate analysis with rich insightful results was pioneered by Yang (e.g., Yang and Borland, 1991; Yang and Ng, 1993; Cheng and Yang, 2004 for a survey). In his tribute to Yang, Buchanan (2004) disclosed that he twice nominated Yang to the Nobel Prize Committee before the sad passing away of Yang at the age of 55 in 2004.

The central tradeoff in Yang’s analysis is that between the economies of specialization due to increasing returns at the individual level and the higher costs of transaction. Autarky (at the individual level) does away with transaction costs but does not tap the economies of specialization. The higher is the degree the division of labor, the larger is the set of goods being exchanged in the market, necessitating more transaction costs. An optimum is reached where the marginal gain from increasing the degree of specialization is offset by the additional transaction costs caused. However, this tradeoff is not continuous. Choices of
different occupations, different sets of goods, different organizational arrangements, etc. involve discrete changes. Yang devised an analytical method to compare the different “corners” of specialization, not only for an individual, but also for the whole economy. As transaction efficiency improves due to better institutions, better transport and communication, etc., the higher is the degree of specialization chosen; more goods are introduced into market transaction. The introduction of previously not marketed goods requires entrepreneurial insight as well as the ability to sustain initial periods of loss. This is especially so as the purchase of the new goods requires other individuals to readjust their degrees and patterns of specialization. In the real world, this must involve substantial time for the readjustment to materialize. The initial periods of losses could be quite long. This again increases (in comparison to the supply-demand analysis) the role of entrepreneurship and finance.

Yang’s framework also inclines us to focus on the organizational efficiency of the patterns of division of labor. For example, the indirect network externality through increasing the degree of specialization of the higher transaction efficiency generated by infrastructure improvements may explain the need, on occasions, of encouraging infrastructure investment (Ng and Ng, 2007). In the words of Buchanan (2004), “The economy of the world, as viewed through the Smithian lenses offered to us by Yang, looks quite different from the world as viewed from the perspective of conventional neoclassical economics …. Yang’s basic insights … have yet to be widely absorbed and appreciated by the disciplinary orthodoxy.”

The presence of increasing returns also makes financial/pecuniary matters more important in other ways. For example, if we use the traditional analysis of upward-sloping supply and downward-sloping demand curves with no increasing returns and the absence of distortions like external effects, the presence of pecuniary external effects has no efficiency implications. This is shown in Figure 5. An increase in demand for a product shifts the demand curve from D to D', the equilibrium point from E to E', and the price from P to P'. Pre-existing consumers of the product lose consumer surplus measured by the area P'PEB. This is a pecuniary external effect as the increase in price makes them worse off. However, this does not cause any inefficiency. The loss in consumer surplus P'PEB is more

---

**Figure 5. Efficiency-Irrelevancy of Pecuniary External Effects in the Absence of Increasing Returns**
than offset by the gain in producer surplus measured by the area \( P'PEE' \). The extra area \( \Delta BEE' \) is the additional surplus of serving the higher demand. The new equilibrium \( E' \) is efficient at the new situation, just as the original equilibrium \( E \) was efficient at the original situation. This efficiency non-relevancy of pecuniary external effects no longer applies in the presence of increasing returns, even if we allow ourselves to continue using the supply-demand analysis, as shown below. (If the increasing returns are at the industry level, the use of the supply-demand analysis is fine; if they are at the firm level, perfect competition can no longer prevail and the supply-demand analysis is no longer appropriate, strictly speaking. Then, using an analysis focusing on the firm level as in the fourth paragraph of this section may be more appropriate.)

The presence of increasing returns makes the supply curve downward sloping, as shown in Figure 6. Here, an increase in the demand curve from \( D \) to \( D' \) shifts the equilibrium point from \( E \) to \( E' \) and lowers the price from \( P \) to \( P' \). Pre-existing consumers of the product gain consumer surplus measured by the area \( PP'E'E \). This gain in consumer surplus is not offset by any loss in producer surplus. Though the price to producers is also reduced, this is achieved by lowering their average costs of producing the product. They continue to break even at \( E' \) as at \( E \), suffering no loss in producer surplus. The presence of pecuniary external effects may have efficiency implications. The new equilibrium point \( E' \) may be superior to the original equilibrium point \( E \). In the presence of increasing returns, equilibrium may not be Pareto-optimal (Ng, 2009a) and pecuniary external effects may have efficiency implications. This again makes pecuniary/financial matters more important. The further analysis of this is however beyond the scope of this paper.

4. Concluding Remarks

This paper explains how the presence of the related non-perfect competition and increasing returns at the firm and economy levels makes money possibly non-neutral and finance (including financial institutions and financial crises) much more important than suggested
by traditional economic analysis, especially in its simple version of perfectly competitive firms and supply-demand curves. In the wake of the recent global financial crisis and the related reconsideration of the adequacy of economics, the simple points discussed in this paper, though not covering many other relevant factors, may contribute partially towards a more adequate and complete picture of our complicated economy. Our analysis suggesting the higher importance of finance complements the conclusion that “economics, if it is to be relevant to reality, should stop neglecting money, wealth and debt” (Bezemer, 2011). How to incorporate these financial matters and relevant real-world institutional factors into the core of formal economic analysis will challenge the ability of economists in the foreseeable future.5

Mathematical Appendix A. The Various Equilibria May Persist

[This appendix is similar to the appendix in Ng (2009b) and is published here with corrections of some mistakes for the convenience of readers.]

This appendix supplements the largely graphical analysis of the text with a mathematical model of simplified general equilibrium. This analysis shows more rigorously that the various equilibria under different sets of conditions may persist as equilibria rather than being nullified as temporary or transitional effects. In particular, not only that the various cost conditions may be consistent with the various cases but also that the expectation regarding both prices and quantities are realized.

For the general case of either perfect or non-perfect competition, we may regard the product of each firm as a distinct good, then, from the maximization of a general utility function

$$U^i = U^i(x_1^i, x_2^i, \ldots, x_N^i),$$

where $U^i = $ utility of individual $i$, $x_g^i = $ amount of the $g$th good consumed by individual $i$. The maximization of utility specified in (A.1) above subject to a budget constraint

$$\sum_{g=1}^N p_g x_g^i = \alpha^i,$$

where $p_g = $ price of the $g$th good, $\alpha^i = $ budget constraint = nominal aggregate demand of individual $i$, we may derive the following demand functions

$$x_g^i = x_g^i(p_1, p_2, \ldots, p_N, \alpha^i), \quad i = 1, \ldots, I; \quad g = 1, \ldots, N,$$

where $I$ is the number of consumers. Summing over all consumers, we have

$$q_g = \sum_{i=1}^I x_g^i = F_g(p_1, \ldots, p_N, \alpha^1, \ldots, \alpha^I).$$

Now, to avoid a fully general equilibrium analysis, two simplifications are adopted. The first is one used by virtually all macro and aggregative studies of ignoring distributional

---

5Some interesting analyses have emerged, e.g., see Brunnermeier and Sannikov (2014).
effects by replacing the vector $\alpha^1, \ldots, \alpha^I$ by $\alpha \equiv \Sigma \alpha_i$. The second simplification is to concentrate on the representative firm and replace the price vector of all other firms by the average price. Without loss of generality, call this representative firm “Firm 1”. From (A.4), we then have

$$q_1 = f_1(p_1, P, \alpha),$$

(A.5)

where $p_2, \ldots, p_N$ has been replaced by $P$, the average price of $p_2, \ldots, p_N$. Due both to the fact that Firm 1 is representative, and the fact that it is assumed small relative to the whole economy, this average price $P$ is also the general price level of the economy. A fully general equilibrium analysis (Ng, 1986, Appendix, 3I) has been used to show that (1) for any (exogenous) change (in cost or demand) there exists (in a hypothetical sense) a representative firm whose response to the change accurately (no approximation needed) represents the response of the whole economy in aggregate output and average price, and (2) a representative firm defined by a simple method (that of a weighted average) can be used as a good approximation of the response of the whole economy to any economy-wide change in demand and/or costs that does not result in drastic inter-firm changes. These fully general-equilibrium results provide a strong methodological justification of the representative-firm approach.

Since demand functions are homogeneous of degree zero in all prices and the budget, we may divide all elements in the functional form in (A.5) by $P$ (i.e., multiply all elements by $1/P$) to obtain, dropping subscripts,

$$q = f(p/P, \alpha/P),$$

(A.6)

where the effect of $P/P$, being a constant, is defined into the functional form of $f$.

The representative firm is assumed to take the aggregate variables as given and maximize its profit with respect to its own output or price. Its profit function is

$$pf(p/P, \alpha/P) - C(q, Y, P, \epsilon^e),$$

(A.7)

where $C =$ total cost, $Y =$ aggregate output of the economy, $\epsilon^e =$ exogenous factors affect costs. The possible effects of $Y$ on $C$ may work through the input market. It may be noted that the cost function is rather general. The first-order condition for the maximization of (A.7) gives

$$\mu \equiv p\left\{1 + \frac{1}{\eta \left(\frac{C}{P}, \frac{\alpha}{P}\right)}\right\} = c(q, Y, P, \epsilon^e),$$

(A.8)

where $\eta \equiv (\partial q/\partial p)p/q$, $c \equiv \partial C/\partial q$ are respectively the price elasticity of demand and marginal cost and $\mu$ is marginal revenue. This equation simply specifies the equality of MR with MC.

From the representativeness of the firm and the requirement of equilibrium, we have

$$P = p,$$

(A.9)

$$\alpha/P \equiv A = Y = qN,$$

(A.10)

where $A =$ real aggregate demand, $N =$ number of firms. (The latter is taken as given in the short-run exercise. Otherwise, an additional equation of free entry/exit may be added, as
done below. The possible feedback of changes in profit is allowed through the effect of real income on aggregate demand in Equation (A.11) below, abstracting from distributional effects.)

The nominal aggregate demand of the economy is taken to be a function of \( P, Y \) (real income which equals real output at equilibrium) and some exogenous factors \( \varepsilon^a \) which should include the money supply.

\[
\alpha = \alpha(P, Y, \varepsilon^a); \quad 1 > \eta^{aP} > 0, \quad 1 > \eta^{aY} > 0,
\]  

(A.11)

where the restriction \( 1 > \eta^{aY} > 0 \) is similar to the case of the Keynesian cross diagram that the slope of \( C + I \) is positive but less than one to avoid an explosive system. Similarly for \( \varepsilon^a \) in (A.11) is a very general function and include the simple Keynesian and Monetarist aggregate demand functions as special cases. This completes the specification of our very simple, general, but powerful model. We turn now to the comparative-statics analysis. The total differentiation of (A.8), after substituting in the total differentiation of (A.9) and (A.10), division through by \( \mu \) or \( c \), gives

\[
(1 - \eta^{cP}) \frac{dP}{P} - (\eta^{cq} + \eta^{cY} - D) \frac{dY}{Y} = \frac{d\bar{c}}{c},
\]

(A.12)

where \( \eta^{cY} \equiv \frac{\partial c}{\partial Y} \), \( D \equiv \frac{\partial \mu}{\partial \alpha} \frac{\partial \mu}{\partial \eta} \frac{\partial \mu}{\partial \eta} \) is the proportionate effect of real aggregate demand on marginal revenue at given prices through possible effects on the price elasticity of demand, \( d\bar{c} \equiv (\frac{\partial \bar{c}}{\partial \varepsilon^a})d\varepsilon^c \) is the exogenous change in marginal cost.

The total differentiation of (A.11), after dividing through by \( \alpha \) and substituting in \( d\alpha/\alpha = dp/P + dY/Y \) from the total differentiation of (A.10), gives

\[
(1 - \eta^{aP}) \frac{dP}{P} + (1 - \eta^{aY}) \frac{dY}{Y} = \frac{d\alpha}{\alpha},
\]

(A.13)

where \( d\alpha \equiv (\frac{\partial \alpha}{\partial \varepsilon^a})d\varepsilon^a \) is the exogenous change in nominal aggregate demand.

Substituting \( dY/Y \) and \( dp/P \) from (A.13) in turn into (A.12), we obtain (A.14) and (A.15).

\[
\Delta \frac{dP}{P} = (\eta^{cq} + \eta^{cY} - D) \frac{d\alpha}{\alpha} + (1 - \eta^{aY}) \frac{d\bar{c}}{c},
\]

(A.14)

\[
\Delta \frac{dY}{Y} = (1 - \eta^{cP}) \frac{d\alpha}{\alpha} - (1 - \eta^{aP}) \frac{d\bar{c}}{c},
\]

(A.15)

where \( \Delta \equiv (1 - \eta^{cY})(1 - \eta^{cP}) + (1 - \eta^{aP})(\eta^{cq} + \eta^{cY} - D) \), \( c \equiv \partial C/\partial q \equiv \) marginal cost of the representative firm, \( C \equiv \) total cost, \( \eta^{ab} \equiv (\partial a/\partial b)b/a \) for any \( a \) and \( b \), e.g., \( \eta^{cq} \equiv (\partial c/\partial q)q/c \) is the (proportionate) change in MC as the firm itself changes its output \( q \) (the sign and magnitude of \( \eta^{cq} \) thus relates to the slope of the MC curve), \( \eta \equiv (\partial c/\partial Y)Y/c \) is the (proportionate) shift in MC as the output \( Y \) of the whole economy changes, \( \eta^{aY} \equiv (\partial a/\partial Y)Y/a \) is the (proportionate) change in nominal aggregate demand \( \alpha \) in response to an increase in aggregate output \( Y \) (which equals aggregate income in equilibrium) and is similar to the Keynesian marginal propensity to spend. Also, a “\( d \)” before a variable with a bar indicated the exogenous change in that variable; \( D \) is the (proportionate) effect of real aggregate demand on the marginal revenue of the representative firm at given prices,
through changing the price elasticity of demand. It may also be noted that, for stability and
for signing the comparative statics, $\Delta$ may be taken as positive. (This does not rule out the
possibility that $\Delta$ may be zero or even negative at a particular range, causing a continuum
of equilibria and instability. However, the negativity/instability must be reversed beyond
this range where a new equilibrium may be established.)

The two equations above indicates that the effects on the price level and aggregate
output depends on the exogenous changes in demand and costs as well as the endogenous
response variables, including the slope of the MC curve (of the representative firm), how
much MC responds to aggregate output and the price level (shifts in the MC curve),
how much the price elasticity of demand changes in response to real aggregate demand,
how much the nominal aggregate demand changes in response to real aggregate income
and the price level. An estimate of these changes and responses would then give us our
estimate of the effects on the price level and aggregate output using the above equations.
These two equations are the basic comparative-statics results which can be used to analyze
the general equilibrium effects of internal and/or external economy-wide changes in
demand and costs on the price level and aggregate output of the economy.

In particular, to analyze the effects of an exogenous change in nominal aggregate
demand as may be relevant for a financial crisis, we may abstract any exogenous changes
in costs by taking $\frac{dc}{c}$ in (A.14) and (A.15) as zero. (This is not a partial-equilibrium
analysis, as endogenous changes in costs are allowed. Rather, it is just “one thing at a
time”. ) We then have, from (A.14) and (A.15),

$$\frac{adP}{Pda} = \frac{(\eta^{cq} + \eta^{cY} - D)}{\Delta}. \quad (A.16)$$

$$\frac{adY}{Yda} = \frac{(1 - \eta^{cP})}{\Delta}. \quad (A.17)$$

These two equations specify how a change in nominal aggregate demand $\alpha$ exogenously
may affect the price level $P$ and aggregate output $Y$ proportionately.

Let us first consider the case of perfect competition. Here, we have $\eta^{cq}$ positive as the
MC curve must be upward sloping as explained in Section 3 in the text. Next, $\eta^{cY}$ is non-
negative if we ignore (as invariably done so in macro analysis) such factors as Marshallian
external economies, as then an increase in aggregate output does not decrease the marginal
costs of firms. $D$, the change in the marginal revenue of the firm as aggregate demand
changes at unchanged prices through possible changes in the price elasticity of demand (for
the product of the firm), is necessarily zero under perfect competition since the price
elasticity of demand is (minus) infinite and remain infinite and hence does not change. This
makes $(\eta^{cq} + \eta^{cY} - D)$ necessarily positive. Also, ignoring such frictions as time lags and
money illusion, marginal costs adjust proportionately to prices, making $(1 - \eta^{cP})$ zero.
Thus, from (A.16) and (A.17), we have a change in nominal aggregate demand exoge-
nously changes only the price level without affecting the output level.

If we do not include the endogenous change in nominal aggregate demand $\alpha$ (i.e., taking
$\eta^{aY}$ and $\eta^{aP}$ in $\Delta$ as equal to zero), we have, from (A.16) and (A.17), the price level $P$
decreases/increases proportionately with the exogenous decrease/increase in nominal aggregate demand. Including the endogenous changes, we have

\[
\frac{a dP}{P d\alpha} = \frac{1}{(1 - \eta^{ap})} > 1. \tag{A.16'}
\]

As \(\eta^{ap}\) is positive but smaller than one from (A.11). Thus, the endogenous change reinforces the exogenous change. This reinforcement effect through \(\eta^{ap}\) may be called the price-multiplier effect, similar to the Keynesian income-multiplier effect. An exogenous decrease in \(\varepsilon^a\) (due to say a decrease in money supply or loss in consumer and investor confidence in a financial crisis) decreases \(\alpha\) by say 10% at the initial income and price levels (see Equation (A.11)). This first leads to a decrease in the price level by 10% in the present case of perfect competition. This decrease in \(P\) further decreases nominal aggregate demand \(\alpha\) and leads to a further decrease in \(P\). If \(\eta^{ap} = 1/2\), the price multiplier equals 2 and the total decrease in \(P\) is twice as large as the initial exogenous decrease in \(\alpha\). In the present case with no effect on \(Y\), the endogenous effects through \(Y\) do not take place. For an increase (in contrast to that of a decrease mainly being concerned here) in nominal aggregate demand, a large price multiplier may partly explain some inflationary episodes. Our results are also consistent with the finding by Rotemberg and Woodford (1996) that a small change is the price of a factor of production of relatively small proportionate employment may lead to large effects on prices and output.

Now consider the more realistic and general case of non-perfect competition. Even in this case, we may still have \((\eta^{cq} + \eta^{cY} - D)\) being positive and \((1 - \eta^{EF})\) being zero, making the results similar to the case of perfect competition, as shown in Figure 2 in the text.\(^6\) However, with non-perfect competition, \((\eta^{cq} + \eta^{cY} - D)\) may no longer be positive. In particular, \(\eta^{cq}, \eta^{cY}\) and \(D\) may all be zero, which is the case illustrated in Figure 3 in the text. More generally, the values of these three variables or response parameters may combine to give a zero or even negative value for \((\eta^{cq} + \eta^{cY} - D)\). For the case of a zero value, we have an exogenous decrease in nominal aggregate demand (as may happen in a financial crisis) triggering a decrease in real output (and hence employment) with no effect on the price level, similar to the special case illustrated in Figure 3. For the case of a negative value, we have the case of a cumulative contraction (or cumulative expansion for an increase in aggregate demand) illustrated in Figure 6 in the text. In either case, a financial crisis may lead to a big fall in real output and employment and cause an economic crisis as in the case of 1929, if not reversed by appropriate counter-recession policies.

It may be noted that a negative \((\eta^{cq} + \eta^{cY} - D)\) must not be mechanically interpreted as a fall in aggregate demand leading to a decrease in prices and an increase in output. This interpretation is a similar mistake as, for the case of a steeper-than-one \(C + I\) curve in the simple Keynesian income determination diagram, taking an increase/decrease in exogenous investment as leading to a decrease/increase in income/output. A \(C + I\) curve steeper than the \(45^\circ\) line is also one of cumulative expansion/contraction.

\(^6\)These include Keynesian and cumulative expansion and the knife-edge in-between case of expectation wonderland where the output depends only on expectation and a purely price-expectational change will be self-fulfilling.
For the long run where the number of firms is allowed to change, we allow \( N \), the number of firms, to enter the demand function and also add the additional condition of zero long-run profit
\[
pf(p/P, \alpha/P) - C(q, Y, P, \varepsilon^c) = 0. \tag{A.18}
\]
It may be thought that, the zero-profit condition should only be applied to the marginal firm, not to the representative firm. However, under the condition of free entry and competition in the factor market, whatever factors that account for a positive profit of any firm will have their prices bid up to make the genuine economic (though not accounting) profit zero. With this additional condition, similarly to the derivation of (A.14) and (A.15) above, we may derive the comparative results for the long run, yielding.
\[
\nabla dP/P = \left[ M \eta^c + (E + \eta^c q) \eta^c Y - M (D + E) \right] (\partial \alpha/\partial) \\
+ (1 - \eta^c Y) (E + \eta^c q) (dC/C) + (1 - \eta^c Y) M (dc/c). \tag{A.19}
\]
\[
\nabla dY/Y = \left[ M (1 - \eta^c p) + (E + \eta^c q) (1 - \eta^c p) \right] (\partial \alpha/\partial) \\
- (1 - \eta^c p) (E + \eta^c q) (dC/C) - (1 - \eta^c p) M (dc/c), \tag{A.20}
\]
where
\[
\nabla \equiv (1 - \eta^p M \eta^c + (E + \eta^q) \eta^c Y - M (D + E)) (\partial \alpha/\partial) \\
+ (1 - \eta^c Y) M (1 - \eta^c p) + (E + \eta^c q) (1 - \eta^c p).
\]
\( M = (p - c)/p \) is the markup of price over marginal cost, \( E = (\partial \mu/\partial N) N/\mu \) at given prices (where \( \mu \equiv \) marginal revenue of the representative firm) is the effect on MR of the entry of new firms through a higher absolute price elasticity of demand due to increased competition.

The long-run effects of a shock may then be analyzed using (A.19) and (A.20) above by estimating the long-run exogenous changes in aggregate demand and costs as well as the long-run endogenous response variables contained in these two equations. Comparing (A.19) and (A.20) with (A.14) and (A.15), we may see that all cases discussed for the short run above remain relevant except that conditions determining the various cases become more complicated, with \( M \eta^c Y + (E + \eta^c q) \eta^c Y - M (D + E) \) replacing \( \eta^c q + \eta^c Y - D \) and with \( M (1 - \eta^c p) + (E + \eta^c q) (1 - \eta^c p) \) replacing \( (1 - \eta^c p) \) as the relevant values determining the various cases. This is so since, for the long run, the response of the average or total costs and the effects of entry/exit must also be taken into account.

In particular, the case with a real output adjustment and no price changes following an exogenous change in nominal aggregate demand (as illustrated in Figure 3 for a special short-run case and in Figure 4 for a long-run case) is still possible if \( M \eta^c Y + (E + \eta^c q) \eta^c Y - M (D + E) \) is zero and the cumulative contraction/expansion case is relevant if this expression is negative. Comparing the likelihood of these non-traditional cases (i.e., calling the money neutrality case as traditional) between the short and long runs, we may note that while the effects of a change in aggregate demand is more likely to have a significant and negative effects on costs, making the non-traditional cases less likely to prevail in the long run, the entry/exit effect \( E \) is positive) makes the non-traditional cases more likely to prevail in the long run.
References


