A Perspective on Modern Business Cycle Theory

Nobuhiro Kiyotaki

The global financial crisis and recession that started in 2007 with the surge of defaults of U.S. subprime mortgages is having a large impact on recent macroeconomic research. The framework of modern macroeconomics that has replaced traditional Keynesian economics since the 1970s has been widely criticized. Many of the criticisms have focused on the assumptions of the representative agent and its abstraction from firm and household heterogeneity. Critics are also skeptical about the model’s ability to explain unemployment and financial crises because it abstracts from market frictions and irrationality. As a result, modern macroeconomics has often been attacked for its futility in providing policy insight in the way that traditional Keynesian economics has done.¹ Some criticisms are constructive and others are misleading. I would like to present my thoughts on what I believe are the contributions and shortcomings of modern macroeconomic theory, in particular the business cycle theory, by responding to some of these criticisms.²

1. REAL BUSINESS CYCLE THEORY

For the past few decades, real business cycle (RBC) theory has been the focal point of debates in business cycle studies.³ According to the standard

---

¹ For example, see Krugman (2009).
² Because of the limitations of space and my expertise, I will not deal with economic growth or the empirical studies of business cycles. I also admit that the references are heavily biased to my own work.
³ See Kydland and Prescott (1982) and Prescott (1986) for examples.
RBC approach, the competitive equilibrium of the market economy achieves resource allocation that maximizes the representative household’s expected utility given the constraints on resources. Although the RBC approach has often been criticized for its abstraction from firm and household heterogeneity, these charges are incorrect. Instead, it would be more accurate to view the RBC framework as one with heterogeneous firms and households all playing a part in the social division of labor under an ideal market mechanism. The real business cycle theory is a business cycle application of the Arrow-Debreu model, which is the standard general equilibrium theory of market economies.

Let us briefly outline the mechanics of an RBC model. Consider an economy with a homogeneous product that can be either consumed or invested. Labor, capital, and land are homogeneous inputs to production and total supply of land is normalized to unity. There are a number of infinitely lived households \((h = 1, 2, \ldots, H)\) and firms \((j = 1, 2, \ldots, J)\). A household’s preference is given by the discounted expected utility of consumption and disutility of work:

\[
E_0 \left\{ \sum_{t=0}^{\infty} \beta^t \left[ u_h (c_{ht}) - d_h (n_{ht}) \right] \right\},
\]

(1)

Firm \(j\)’s maximum output is a function of the factors of production: capital, land, and labor \((k_{jt}, l_{jt}, n_{jt})\) represented by the production function

\[
y_{jt} = f(k_{jt}, l_{jt}, n_{jt}; z_{jt}),
\]

(2)

where productivity of firm \(z_{jt}\) follows a Markov process. In the goods market equilibrium, aggregate output equals aggregate consumption and investment:

\[
\sum_{j=1}^{J} y_{jt} = \sum_{h=1}^{H} c_{jt} + \sum_{j=1}^{J} k_{jt+1} - (1 - \delta) \sum_{j=1}^{J} k_{jt},
\]

(3)

where \(\delta\) is the depreciation rate of capital.

Here, we assume that markets are complete, that is, there exists a complete set of Arrow securities so that state-contingent claims to goods and factors of production for every possible future state can be traded at the initial period. We also assume that capital, land, and labor can be allocated freely across firms every period and all markets are perfectly competitive. Under these assumptions, the competitive equilibrium achieves an allocation that maximizes the weighted average of all individual households’ expected utilities with constant weights \(\lambda_h\) given the resource constraints (Negishi 1960).

We define the representative household’s utility function as the weighted average of all household utilities:
The aggregate production function is defined as total output given the efficient allocation of factors of production and can be written as

\[ Y_t = A_t F(K_t, N_t) = \max_{k_{jt}, l_{jt}, n_{jt}} \sum_{j=1}^{J} f(k_{jt}, l_{jt}, n_{jt}; z_{jt}) \]

subject to the resource constraint

\[ C_t + K_{t+1} - (1 - \delta) K_t = A_t F(K_t, N_t). \]

Note that the representative household is not an assumption; it arises as an implication of constant Negishi weights under complete markets as in Negishi (1960). The aggregate production function is also constructed under the assumption that production is efficient in competitive markets without friction. Therefore, the real business cycle theory does not blindly abstract from firm and household heterogeneity. By assuming that markets are functioning “well,” we reduce an otherwise general model to one of the representative agent with an aggregate production function and analyze the business cycle phenomenon in this simplified economy.

Now I wish to discuss, in an intuitive manner, how the real business cycle theory explains the fluctuation of aggregate quantities \((C_t, N_t, Y_t, K_{t+1})\) by a shock to aggregate productivity. Suppose that aggregate productivity suddenly increases temporarily. Following this shock, marginal product of labor will increase, leading to a rise in the real wage and therefore the quantity
of labor supplied. The combined effect of higher productivity and increased use of labor will cause output to rise. But since the productivity increase is temporary, future output is expected to increase less than present output, and permanent income and consumption do not increase as much as present output. Thus, from the goods market equilibrium condition (output = consumption + investment), investment and, hence, next period capital stock will increase. This will increase next period marginal productivity of labor, labor input, and output, leading to another cycle of aggregate quantity increases and so on. Hence, we notice that a temporary shock to productivity has precipitated a persistent rise in aggregate quantities.

However, the biggest problem with the propagation mechanism described above is that short-term changes in investment have little impact on capital stock. At the same time, if there is a persistent increase of output, permanent income and consumption will increase almost as much as current income, which leaves little room for investment to rise. Therefore, we conclude that capital plays a limited role in the propagation of a productivity shock. Furthermore, the substitution and wealth effects of a productivity shock on labor supply work to cancel each other out: At a higher real wage, the representative agent is willing to supply more labor, but the higher aggregate productivity also increases the agent’s wealth, which in turn reduces the labor supply. Unless the substitution effect is very large, the overall fluctuations of labor will not be very large. As a result, we need large and persistent aggregate productivity shocks in order to explain the business cycle phenomenon. Because RBC models are missing a powerful propagation mechanism whereby small shocks to the economy amplify and produce large fluctuations, they rely on large exogenous shocks. But the question is where do these exogenous shocks come from? It is difficult to identify such shocks even with the recent global recession or the 1930s Great Depression.

2. OTHER SHOCKS

While exogenous shocks to productivity were the main source of shock in early RBC analysis, the framework was later extended to include the effects of other potential shocks. For example, in the face of a global downturn, what would be the effects of a decreasing demand for exports? We cannot address this question in a perfectly competitive economy since individual firms are assumed to make their production decisions by taking market prices as given, that is, they cannot perceive changes in demand directly. Therefore, let us assume a monopolistically competitive economy in which each firm \( j \) sells a differentiated good. The quantity of aggregate output, which can be used as either a consumption good or investment good, is a function of many differentiated goods as
\[ Y_t = \left[ \sum_{j=1}^{J} \left( x_{jt} \right)^{\frac{1}{\theta}} \right]^{\theta / (\theta - 1)}, \]

where \( \theta \) is the elasticity of substitution between differentiated goods and \( \theta > 1 \). The parameter \( x_{jt} \) is an exogenous idiosyncratic demand shock to firm \( j \)'s product. If we let \( p_{jt} \) be the price of each good, the price index that corresponds to the above aggregate output is

\[ P_t = \left[ \sum_{j=1}^{J} x_{jt} \left( p_{jt} \right)^{1-\theta} \right]^{\frac{1}{\theta - 1}}. \]

Since households and firms use differentiated goods such that their consumption and investment levels are maximized subject to their budget constraints, for a given level of aggregate output produced, aggregate demand for short, real income of each firm is given by

\[ \left( \frac{p_{jt}}{P_t} \right)^{\frac{1}{\theta}} y_{jt} = x_{jt} \left( Y_t \right)^{\frac{1}{\theta}} \left( y_{jt} \right)^{1-\frac{1}{\theta}}. \]

Export demand shocks, which shift aggregate demand \( Y_t \), will affect real income of firms, and will change production, employment, consumption, and investment levels the way productivity shocks did in the previous section. Although a monopolistically competitive economy yields inefficient equilibrium resource allocations, key features of the business cycle are not significantly different from those of a perfectly competitive economy. In fact, a monopolistically competitive equilibrium corresponds to a perfectly competitive market equilibrium with a value-added tax that redistributes the tax revenue lump sum. Therefore, simply adding monopolistic competition to an RBC model cannot account for the business cycle phenomenon, and some other source of friction such as price stickiness or a different type of shock is necessary.

Now, instead of a shock to aggregate demand, let us consider a shock to the quality of capital. Assume that a fraction \( \psi_{t+1} \) of differentiated goods becomes obsolete between periods \( t \) and \( t + 1 \), and that the capital used as inputs in the production of those goods also becomes obsolete. In other words, the idiosyncratic demand parameter \( x_{jt} \) of goods affected by the obsolescence shock becomes zero, and the corresponding amount of demand shifts toward new goods. As a result, the productive capital stock will decrease to

\[ K_{t+1} = I_t + (1 - \psi_{t+1})(1 - \delta)K_t. \]
This lower capital stock induces a decrease in output and employment. If there are no other sources of friction in the economy, however, investments will increase, encouraging the expansion of labor supply, and contribute to a quick recovery of output. This is similar to the adjustment process of an economy with initial stock of capital lower than the steady-state equilibrium level in the Neoclassical optimal growth model. Again, incorporating capital obsolescence shocks is insufficient to explain standard cases of recessions in which investment and employment are depressed instead of booming. (See Section 4 for more explanation.)

3. LABOR MARKET FRICTION

Real business cycle theory is often criticized for its lack of implications for the cyclical behavior of unemployment. This issue has been partially addressed in the Diamond-Mortensen-Pissarides framework that incorporates matching frictions that exist in the labor market between workers and firms. Matching theory assumes that it is costly and time consuming to find productive matches because workers and jobs are heterogeneous. In order to include this feature into the macroeconomic model, they introduced an aggregate job-matching function written as an increasing function of job vacancies $v_t$ and the number of unemployed workers (difference between the workforce and employment level, $\bar{N}_t - N_t$):

$$N_{t+1} = \mu_t M(v_t, \bar{N}_t - N_t) + (1 - \delta_{nt})N_t.$$  

$\mu_t$ represents the efficiency of the job matching and $\delta_{nt}$ is an exogenous parameter that measures the rate at which current job matches are destroyed. We assume firms incur a recruitment cost of $\chi$ units of the output good per vacancy. The goods market equilibrium condition is

$$Y_t = C_t + K_{t+1} - (1 - \delta)K_t + \chi v_t.$$  

After firms and workers are matched, wages are determined by Nash bargaining. We assume the Hosios condition (Hosios 1990) (under which the firm’s bargaining power is equal to the elasticity of the number of aggregate job matches with respect to the vacancies) is satisfied. Each household consists of many workers, and is therefore able to diversify labor income risk from unemployment. The competitive equilibrium of such an economy with search maximizes the expected utility of the representative household.

---

In a search and matching model, search is an investment of current resources for future returns, and we expect substantial fluctuations in unemployment only when labor productivity and demand are expected to change persistently. However, according to Shimer (2005), even with persistent labor productivity shocks, fluctuations in unemployment will be small if the marginal product of labor is significantly larger than the marginal cost of labor supply (marginal rate of substitution between labor supply and consumption $d(N)/u(C)$) and if wages are determined by Nash bargaining. Therefore, search models appear to be limited to explaining fluctuations in unemployment of young workers fresh out of school and old workers nearing retirement for whom the difference between the marginal product of labor and marginal cost of labor supply is small.

4. HETEROGENEITY AND CREDIT LIMITS

In an Arrow-Debreu economy that underlies RBC theory, credit is considered to be a particular kind of exchange: The borrower receives present goods (or purchasing power to buy goods at present) in exchange for paying the purchasing power at a future date. In this economy, there is an auctioneer who has the authority to enforce all contracts for all the contingencies, thus eliminating any failure of payment in the future. Therefore, an exchange between present goods and future goods in this market is not subject to any frictions and is no different from an exchange between two present goods. If, however, this enforcing auctioneer is absent in a decentralized market economy, then a borrower can default on his payment in the future. Anticipating the possibility of default, the creditor requires collateral for the loans and makes the amount of credit contingent on the value of the collateral. In order to analyze the business cycle in economies with such credit constraints, we assume that it takes one period to transform inputs into output. Instead of production function (2), we use

$$y_{jt+1} = f(k_{jt},l_{jt},n_{jt},z_{jt}).$$

We assume that the maturity of all outstanding debt is one period and that the debt repayment in the next period $b_{jt}$ cannot exceed a fraction $\phi$ of the expected value of the collateral, which in this model we assume to be land:

$$b_{jt} \leq \phi E_t(q_{t+1}l_{jt}).$$

Here $q_{t+1}$ is the price of the land in period $t + 1$ and $l_{jt}$ is the amount of land on collateral. We assume that the repayment amount is independent of the state of the borrower or the economy (i.e., the debt is noncontingent). So even though we rationalize the imposition of the borrowing constraint by
reference to the possibility of default, we assume that there is no default in
equilibrium. The entrepreneur’s budget constraint is

\[ c_{jt} + k_{jt} + q_{l,jt} + w_{i,n_{jt}} = \left[ y_{jt} + (1 - \psi_t) (1 - \delta) k_{jt-1} + q_{l,jt-1} - b_{jt-1} \right] + \frac{b_{jt}}{r_t}. \]  (9)

The left-hand side of the equation is the entrepreneur’s expenditure on
consumption (or dividend) and factors of production—capital, land, and labor.
(We assume the entrepreneur must buy capital and land and cannot rent their
services.) The right-hand side represents the firm’s sources of finance where
internal finance is in the square bracket—net worth that equals output plus
undepreciated capital and land from the previous period net of repayment of
old debt. The last term on the right-hand side is the external finance derived
from new debt (calculated as the present value of next period’s repayment on
loans discounted by the gross real interest rate \( r_t \)). Each entrepreneur chooses
a sequence of consumption, investment, output, and debt in order to maximize
the discounted expected utility subject to the constraints of technology, credit,
and available funds.

Now let us examine the difference between the RBC model and the econ-
omy in which producers are heterogeneous in productivity \( z_{jt} \) and are credit
constrained.\(^5\) First, if there is limited contract enforcement, then insurance
is incomplete. Because the insurance company is aware of the fact that the
insurees may not pay in the future, the company, as a precautionary measure,
demands premium payments upon entering an insurance contract. Thus, pro-
ducers and households with low net worth may not purchase insurance with
full coverage. When the economy is then hit by various shocks, the net worth
of firms and households with partial insurance coverage fluctuates, which in
turn requires an adjustment of the Negishi weights. As a result, we can no
longer maintain the assumptions of the representative household approach.
In addition, when borrowing constraints exist, firms must rely on internal
finance—their net worth—as a source of financing inputs. When the borrow-
ing constraint is binding for some firms, the marginal product of capital, land,
and labor across firms will no longer be the same. Thus, the assumptions for
the existence of a representative firm no longer hold, and an aggregate produc-
tion function such as given by equation (4) no longer exists. Now, aggregate
productivity of the economy will fluctuate endogenously with credit levels and

When productive firms borrow up to the credit limit and also use their own
net worth to finance additional investments that the loans could not cover, the

\(^5\) See Bernanke and Gertler (1989), Kiyotaki and Moore (1997a, 1997b), Kiyotaki (1998),
and Bernanke, Gertler, and Gilchrist (1999) for examples of RBC models with credit constraints.
impact of a small shock to total productivity, investment, and net worth is large. In order to explain the propagation of the effects of the shock, let's assume that net worth of all firms has declined because of the obsolescence of some of their products, and thus the capital used to produce those goods has also become obsolete. Because highly productive firms have outstanding debt from the previous period, the leverage effect of the debt will result in a sharp reduction of net worth (refer to Figure 1, point A). These productive firms will decrease their demand for capital and land because they cannot borrow more (Figure 1, point B) and aggregate productivity will fall as the share of investment of productive firms declines. Because it will take some time for the highly productive firms to recover their preshock level of net worth (Figure 1, point C), their demand for assets (capital and land) and labor will be constrained for a while and therefore aggregate productivity and aggregate demand for assets and investment are also expected to be stagnant for a while (Figure 1, point D). Under these expectations, current period asset prices drop (Figure 1, point E) and the balance sheets of the highly productive firms further deteriorate (Figure 1, point A). As a result, the small aggregate shock causes a persistent decrease of the share of investment by credit-constrained and highly productive producers, which leads to a persistent decline of aggregate productivity. Thus, with borrowing constraints, the fall in asset price is responsible for the magnified drop in output.
Joan Robinson once said, “the essence of Keynesian economics is its recognition of the central role of time in human lives. People live in the present moment which is continuously moving from an unknown future to the irrevocable past.”6 Because the demand for assets by productive firms that face credit constraints depends on each firm’s own net worth (which equals accumulated past savings), it takes time for them to recover from a negative shock to net worth (i.e., the effects of shocks are persistent as firms are held back by their past savings). Meanwhile, asset prices are driven down by expectations of a prolonged stagnation in future asset demand (i.e., expectations about the future affect present asset prices). Notice how the asset market serves as a platform on which past savings and expectations about the future interact in present time.

In the real business cycle model with no constraints on borrowing, shocks from the obsolescence of goods and capital trigger higher investment, leading to a quick recovery of capital stock and output (dotted line in Figure 2.) In contrast, in an economy where borrowing constraints exist, the obsolescence shocks significantly reduce net worth and investment of productive firms, further decreasing capital stock and output. Since it takes time for highly productive, yet credit constrained, firms to recover their net worth and investment levels, total output and productivity will both fall persistently (solid line in Figure 2).

Reinhart and Rogoff (2009) claim that, although financial crises are much like forest fires in the sense that it is difficult to predict when and where they will occur, certain conditions set the stage for crises. They present the following as indicators of an emerging crisis for a country: 1) asset prices rise rapidly, and especially the price-rent ratio for real estate increases sharply; 2) the amount of debt expands faster than aggregate output and asset values, leading to a higher leverage (the ratio of total assets to net worth); and 3) the country experiences massive capital inflows. In the presence of borrowing constraints, firm heterogeneity in productivity, and diverse investment opportunities, if an adverse shock arrives when the overall leverage level of the economy is high enough, the powerful propagation mechanism as described in Figure 1 will take effect in asset prices, credit levels, and outputs, forcing the economy into financial crisis.7

When a financial crisis is accompanied by a banking crisis, I expect that the financial system will cause additional problems that will aggravate the crisis.

---

7 Kiyotaki and Moore (1997a) explain how total output, asset price, and debt fluctuate cyclically when exposed to an exogenous shock, while Matsuyama (2008) suggests that the fluctuation can occur even in the absence of a shock. Aoki, Benigno, and Kiyotaki (2009) distinguish domestic and foreign credit limits, and show that, if the domestic economy has an underdeveloped financial system, it becomes prone to both expansion and contraction after capital account liberalization, as Reinhart and Rogoff (2009) suggest.
To analyze such crises, we need to look beyond the credit constraints of non-financial borrowers and consider the role of financial intermediaries and their financing constraints. Theories of financial intermediation have developed since Diamond and Dybvig (1983), and others such as Williamson (1987) have extended macroeconomic models to include banks. Although there is active recent research on the source of problems caused by financial intermediaries and their markets (especially “wholesale” or “interbank” financial markets), there is not yet a standard macroeconomic model for analysis of financial intermediation.\footnote{Kiyotaki and Moore (1997b, 2008) analyze the effects of productivity and liquidity shocks on aggregate production in an economy where firms are involved in both production and financial intermediation. Gertler and Kiyotaki (2011) study the moral hazard problem of financial intermediaries, the relationship between their balance sheets and business cycles, and the effects of broad monetary policies. These articles also provide more references to the literature.}

In addition, note that it is the leverage effects from debt obligations that induce the net worth and investment of highly productive firms to persistently decline in the presence of borrowing constraints. If firms issue preferred stock or other securities whose returns are contingent on the firm performance instead of taking out loans to finance their investments, the leverage effects will not materialize. Therefore, in order to justify the propagation mechanism,
we need to first explain why firms would choose to borrow and not issue contingent securities in procuring their funds. We also need to explain why firms choose not to issue common stocks in order to recover net worth when it is deteriorating.

5. CONCLUSION

In this article, I explain that business cycles in an economy of heterogeneous firms and households can be analyzed using the representative agent approach if their interactions take place in an economy without frictions and complete markets. However, in an economy where markets do not function smoothly because of frictions such as credit constraints, the representative household framework may no longer be appropriate and aggregate productivity changes endogenously with the distribution of wealth and productivity of firms. Thus, I argue that the interaction of heterogeneous firms and households in the presence of credit constraints is important for business cycle analysis. Finally, I would like to propose some questions and directions for future research.

While in the presence of borrowing constraints, capital and land does not move between firms so that the marginal products of capital and land are not equalized across firms; the allocation of capital and land will gradually adjust in a similar way that water flows downhill. For example, firms with high marginal products of capital and land do not consume or pay out dividends in excess, and hence accumulate net worth. As a result, they will eventually be less constrained by external finance constraints. Even in an economy where capital and land do not move freely, if labor can move freely between firms, the marginal product of labor will be equalized across firms. Then, the marginal product of capital and land will also become more equal across firms. One suggestion for future research is to study how the distribution of productivity and net worth of firms evolves and how persistent the differences in marginal products of inputs across firms are.

In order to obtain a deeper understanding of the importance of firm heterogeneity, we need to analyze what determines and changes the productivity of individual firms. According to Bernard et al. (2003), firm labor productivity in an industry varies widely between less than one-fourth and more than four times the average labor productivity of the industry. Differences in human and physical capital can account for only a small fraction of these productivity differences among firms. In order to explain the diverse productivity across firms, we need to consider the accumulation process of both tangible as well as intangible assets. As modern growth theory has attempted to extend its models to include endogenous technical progress in addition to the accumulation of the factors of production, perhaps it is about time for modern business cycle theories to look into the source and the propagation of the shocks by
exploring the endogenous evolution of an individual firm’s productivity in general equilibrium.

REFERENCES


