# NETWORK EFFICIENCY AND THE BANKING SYSTEM

Nicola Giocoli*

*University of Pisa*

giocoli@mail.jus.unipi.it

**Abstract**

Inspired by the Coasean "market vs firm" dichotomy, we offer a new definition of efficiency by applying the notions of network cost and network efficiency as developed in complex network theory. Network analysis is relevant for every system of interconnected exchanging agents. One such system is the banking sector. It is showed that the notions hereby presented may improve upon the predictions of Allen & Gale's standard model, where agents exchange liquidity and where troubles in a local area of the network may lead to systemic collapse.

**JEL Codes:** D02, D23, G21

**Keywords:** network theory; efficiency; banking system; systemic risk

**Word Count:** 8,139 (including footnotes and references)

FINAL VERSION: JUNE 10TH, 2014

* I thank Giacomo Luchetta and this journal's editor and anonymous referees for their useful comments. Financial support from INET 2011 Grant "Free from what? Evolving notions of 'market freedom' in the history and contemporary practice of US antitrust law and economics" is gratefully acknowledged. I am of course responsible for any remaining mistake.

# NETWORK EFFICIENCY AND THE BANKING SYSTEM

## §1. Introduction: from Coase to banks

In a classic statement, Ronald Coase declared: "The main reason why it is profitable to establish a firm would seem to be that there is a cost of using the price mechanism", the latter being "[t]he cost of negotiating and concluding a separate contract for each exchange transaction which takes place on a market" (Coase 1937, 390-1). The passage is interpreted as meaning that, in the absence of a firm, each input owner would possibly have to contract with every other owner of the inputs whose cooperation is required in the production process, while within the firm each input owner negotiates a single contract with the entrepreneur. In case $N$ inputs must cooperate to obtain the final product, a total of $N(N-1)/2$ bilateral contracts would be required in the free market, while only $(N-1)$ contracts would suffice within a firm. If, as argued by Coase, each contract has a cost, the latter is the most efficient solution to produce the output. Firms thus emerge as the efficient response to the costly process of contracting.

This well-known explanation of the origin of firms relies on the implicit assumption that efficiency in production is tantamount to cost minimization. However, there may be production processes where an alternative notion of efficiency, based on the effectiveness with which knowledge, or information, or any other kind of valuable input, is transmitted from one agent to the other, sounds more appropriate. In such a case, the existence of a direct connection, in the form of a contractual relation, between any two agents participating in a common production process is a more effective way for exchanging knowledge between them than the alternative solution of having knowledge flow through the firm's hierarchy from, say, an agent up to the entrepreneur, and then from there down to another agent. Moreover, a network of direct relationships between input owners makes the production process more resilient to possible breakdowns in

communication flows. In the case of a firm, each agent is connected to the others only through its contractual relationship with the entrepreneur, so much so that any interruption of this relationship (due to, say, adverse weather conditions or a contractual breakdown) determines the agent's inability to contribute to the production process. This is not so when the agent is connected to many others via a number of contracts: even if one, or few, of its relationships break down, she will still be able to contribute to the joint endeavor (e.g., by providing essential information) through the remaining ones. An alternative definition of efficiency should ideally encompass these phenomena, while retaining a key role for cost minimization.

The paper's underlying intuition is that the heuristic potential of such an alternative efficiency notion would go much beyond the "market vs firm" dichotomy. For example, assume the production process under scrutiny is the provision of liquidity to end users (households or businesses) via the banking system. Coase's dichotomy may thus be interpreted as two possible ways to "produce" the liquidity required by the real sector of the economy. First, think of a centralized banking system, where the liquidity collected from depositors is channeled by individual branches towards a central institution – a money center (as in Eboli 2001) – which then oversees its redistribution to borrowers, either directly or through its branches. The main advantage of such a system would be similar to that of the Coasean firm: the minimization of the cost required to build the network of financial channels. Each contract to transmit liquidity entails a cost, if only in terms of the information about the borrower that the lender has to gather before making the loan. Hence, the centralized system is the most cost-effective way to provide liquidity. In addition, there is a further advantage in terms of the network's protection from systemic crisis: the failure of one branch to repay its debts (i.e., return the liquidity to its depositors) would have little if any impact at system level because the ensuing financial imbalance would not propagate beyond the money center that, by definition, has the potential to offset it in its sizeable balance sheets.

Think now of a decentralized financial system where each agent is an independent bank that exchanges liquidity with the other banks. Liquidity may now circulate much more effectively and always be promptly transmitted from agents in surplus (be they banks or depositors) to those in deficit. Moreover, an

interbank network for exchanging liquidity represents a safety net for each bank, which may now get liquidity from several channels, as well as for the system as a whole, because it warrants that, even in case of default of a single bank, liquidity will still continue to flow to and from every local area of the economy via alternative channels. On the minus side, such a system is very costly because now each bank has to gather information about every other bank with which it exchanges liquidity, as well as about each of its partners' partners (because financial troubles of any of the latter may impact on the bank's direct partners' ability to repay debts). Moreover, the very same safety net provided by the interbank network has the potential to create a contagion from the failure of a single bank to repay its debts to the balance sheets of its lenders that, when roughly of the same size, may in turn become unable to honor their own debts, thereby triggering a (possibly explosive) propagation of the liquidity crisis.

In Coasean fashion, we may ask ourselves which of the two systems is the most efficient one for liquidity circulation. Is it preferable to have a cost-minimizing, vertically integrated system, which is also less exposed to the risk of systemic crisis, or a decentralized, interlocked system that may quickly transfer liquidity everywhere in the market and is locally robust to failures, but that may more easily suffer from global collapse? A trade-off seems to exist between mutual insurance and systemic risk in a banking system.[1]

What limits the centralization of transactions? The standard answer, once again inspired by Coase 1937, is that as firms grow bigger, it becomes ever more costly to manage them. A threshold of managerial complexity exists beyond which it is less expensive for the firm to leave certain kinds of transactions or activities to the market, rather than integrate them. This intuition can be generalized by observing that every agent has only a bounded ability to handle the – potentially very large amount of – inputs (or knowledge, or information) received. A crucial insight is that an agent's *position* in the network determines the seriousness of the troubles caused by her limited capacity. For example, the cost-minimizing centralized system may breakdown when the amount of inputs received by the central agent trespasses the latter's maximum capacity to effectively manage

[1] The trade-off is well-known in the literature. See Allen & Gale 2000; Eboli 2001; Iori, Jafarey & Padilla 2006; Nier et al. 2007; Rotemberg 2008. As opposite real-world examples, think of a banking system that is closely supervised and coordinated by a central bank and a shadow banking system that is entirely decentralized.

them. The problem may be solved by either paying the cost of building a decentralized network that may spread the inputs' load over more agents or by bearing a new kind of cost addressed at augmenting the central agent's capacity.

Think again of the banking system. A bank exceeds its capacity when the demand for its liquidity is higher than its ability to collect it. The unsatisfied demand may then turn to another bank, but this may cause further troubles if the latter exceeds its own capacity to provide liquidity; the new troubles may then spread to a third bank, and so on. The crisis may be avoided by allowing a bank extra sources of liquidity – for example, by letting the bank sell its long period assets or by establishing a special "liquidity creator" (a.k.a. the central bank) with the duty to provide liquidity on demand. Yet, none of these solutions is zero-cost.

Summing up, we have four elements to take into account when evaluating the "market vs firm" dichotomy in its general form of a network of interconnected agents: 1) the cost of establishing a communication channel (say, a contract) between two agents; 2) the effectiveness in the transmission of the inputs/information among the agents; 3) the robustness of the entire system, or of parts of it, to the breakdown of an agent or a communication channel; 4) the cost of providing each agent with the capacity to manage the inputs/information received. The four features find easy translation in the banking system case. Standard economics only emphasizes the first and, sometimes, the fourth element. However, it is just through the interplay of all four that we may fully characterize the working of every system where either information or inputs or liquidity are transmitted from one agent to another.

For a complete solution we have to look outside economics and towards the theory of complex networks, a relatively young research area in physics and mathematics that aims at investigating the topological properties of non-trivial networks. In a series of recent papers physicists Vito Latora and Massimo Marchiori (LM thereafter) have proposed some notions of network cost and network efficiency that happen to nicely overlap with the four above-mentioned features. The rest of the paper aims at presenting these notions, showing that they can provide a richer toolbox for the analysis of interconnected systems, such as the banking one.[2]

[2] For an authoritative call to employ in macroeconomics the tools of network analysis, see Caballero 2011. See the previous footnote for some references in this literature, where however no mention is made of LM's formalism.

The paper is organized as follows. The next section covers the basic material of complex network theory. In §§ 3-4 we introduce LM's notions of, respectively, network efficiency and network cost and collapse. The fifth § examines how our approach may be applied to the analysis of banking systems and liquidity crises. §6 concludes.

## §2. Basic definitions and the notion of small worlds

Consider a generic network $\mathbf{G}$ with $N$ members and $K$ links connecting the members (with $K \ll N(N-1)/2$, to ensure that the network is sparse, i.e., that only a few of the total possible number of edges exist). If we represent $\mathbf{G}$ as a graph, $N$ is the number of vertices and $K$ that of the edges. The polar cases are those of *regular* and *random* networks. The former can be represented by a graph where each vertex is connected with the same number of other vertices, the latter are represented by the kind of irregular graphs that obtain when any two vertices may be independently connected with a given probability.

In a 1998 paper Duncan Watts and Steven Strogatz (WS thereafter) have shown that the connection topology of some real world networks is neither completely regular, nor completely random (Watts & Strogatz 1998). These networks, named *small worlds* in analogy with the notion developed by social psychologist Milgram (1967), enjoy both a property typical of regular networks, namely, a *high clustering*, and one typical of random networks, namely, a *small characteristic path length*. The first is a local property and refers to the high probability that the existence of a link between network members *i* and *j* and between members *j* and *k* entails the existence of a link also between members *i* and *k* – that is to say, there is a high probability that "the friends of my friends are also my friends". The second is a global property and refers to the shortest distance, in terms of edges to be traveled, between any two network members:[3] the average value of these shortest distances calculated over all network pairs of vertices is called the network's characteristic path length. In random networks this average value is

[3] This shortest distance is also popularly known as the number of degrees of separation.

relatively small, indicating the likely existence of direct connections even between faraway vertices.

WS use the mentioned properties to characterize small worlds. However, their definition can only be applied to a special class of networks, namely, those represented by graphs which are, at the same time, simple, connected and unweighted. A graph is *simple* when no multiple links connecting the same pair of vertices exist. A graph is *connected* when there exists at least one path connecting any couple of vertices (i.e., it must always be possible to reach a given vertex starting anywhere in the graph). Finally, a graph is *unweighted* when all the edges between vertices are equal: in such a case, the term topological networks is used, to indicate that what distinguishes edges is just their position and relation to other edges, but not their individual properties (like, say, length).

Latora and Marchiori (LM) aim at generalizing WS's definition of small worlds, in order to extend it to a more general class of networks. But before moving to LM's work, a little extra terminology is needed from complex network theory.[4]

In the case of WS's simple, connected and unweighted graphs, all the information necessary to describe a given graph **G** is summarized by the $N \times N$ symmetric matrix $\{a_{ij}\}$ - called the *adjacency matrix* – where each $a_{ij}$ is 1 if an edge exists joining vertex *i* to vertex *j* and 0 otherwise. Define the *degree* of vertex *i* as the number $k_i$ of its edges, i.e., the number of other vertices directly connected with *i* – the *neighbors* of *i*. Each $k_i$ can be determined from the adjacency matrix as $k_i = \sum_j a_{ij}$, i.e., as the sum of 1s which can be found in row *i*. The average value of $k_i$ captures an important characteristic of any graph – that is, its *connectivity* – and is calculated as $k = 1/N \sum_i k_i = 2K/N$.

Now define $d_{ij}$ as the smallest number of edges to go from *i* to *j* – what we call the *shortest path* between *i* and *j*. In general, it is $d_{ij} \geq 1$, with the equality when a direct link exists between *i* and *j*. Note that all entries in the shortest paths matrix $\{d_{ij}\}$ can also be determined from the information in the adjacency matrix.

[4] See Newman 2005; Caldarelli & Vespignani 2007, Ch.1; Caldarelli 2007, Ch.1.

Both notions used by WS to define a small world stem from the previous definitions. The *characteristic path length* $L$ of graph $\mathbf{G}$ is defined as the average of the shortest paths between two generic vertices:

$$L(\mathbf{G}) = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathbf{G}} d_{ij} .^{5}$$

The clustering of a network is measured by the *clustering coefficient*. Take $\mathbf{G_i}$ to be the subgraph of the neighbors of vertex *i*. If the graph is simple, there are $k_i$ such neighbors, hence $\mathbf{G_i}$ may have at most $k_i(k_i-1)/2$ edges in case it is completely connected (meaning that every neighbor of *i* is connected to every other neighbor of *i*). The clustering coefficient $C_i$ of subgraph $\mathbf{G_i}$ is defined as the fraction of possible edges of $\mathbf{G_i}$ that actually exist: $C_i = \frac{K_{\mathbf{G_i}}}{k_i(k_i-1)/2}$, where $K_{\mathbf{G_i}}$ indicates the actual number of edges in sub-graph $\mathbf{G_i}$. The clustering coefficient *C* of graph $\mathbf{G}$ is the average of the $C_i$, calculated over all possible *i*s:

$$C(\mathbf{G}) = \frac{1}{N} \sum_{i \in \mathbf{G}} C_i .$$

WS's small worlds are networks with a high clustering coefficient *C* and a small characteristic path length *L*.[6]

## §3. Network efficiency measures

In a series of works LM (with their co-authors) have proposed a new property for complex networks, called *efficiency*, and a couple of measures of the cost required to build a network's edges and vertices.[7] LM efficiency measures the effectiveness of communication between network members, i.e., how "fast" the information may

[5] The formula clarifies why connectedness is crucial in WS's analysis. If graph $\mathbf{G}$ is disconnected, there are at least two vertices with no finite length path connecting them, i.e., which are at infinite distance one from the other. In such a case $L(\mathbf{G})$ becomes an ill-defined quantity.

[6] Note that the existence, both theoretically and in the real world, of networks endowed with these two properties could not be taken for granted before WS, given that regular networks have both high *C* and high *L*, while random networks have both low *C* and low *L*. Small worlds are networks that exhibit a high level of clustering, but also preserve the possibility to directly (or, in any case, swiftly) travel between any pair of vertices, regardless of their position in the graph.

[7] See Latora & Marchiori 2001; 2002; Crucitti et al. 2003; 2004; Crucitti, Latora & Marchiori 2004.

travel between any two vertices of a graph. The notion is more general than WS's clustering and characteristic path length because it can be applied to every kind of graph, be it non-simple and/or disconnected and/or weighted. The latter circumstance is especially crucial. Real networks are almost never just topological, i.e., the links connecting network members almost always differ one from the other in terms of, say, physical length (think of a transportation network) or volume of "transmission" (think of computer networks, which transmit data, or financial networks, which transmit liquidity). Moreover, LM efficiency can be calculated both globally and locally, i.e., for the network as a whole and for any of its sub-networks.

LM's main use of their new notion is to prove that small worlds are characterized by high global and local efficiency. The former property means that in a small world communication between any two members is very effective, the latter that communication among neighbors of a given member is not disrupted even in case that member abandons the network. To give a concrete example,[8] a transport network is a small world if: *i*) it is usually possible to go quickly – i.e., with a relatively small number of intermediate stops – from one station to another (global efficiency), and, *ii*) if closing a station usually does not disrupt the possibility to travel from the station immediately preceding it to that immediately following (local efficiency).

Let us now define LM efficiency. Dealing with weighted graphs requires that we define a new matrix, i.e., the matrix of the weights associated to each link. Let $w_{ij}$ be the weight associated to the link between vertices *i* and *j*. Generally speaking, a weight is just a real number attached to an edge,[9] although its most intuitive interpretation – and the one followed here – is that of the physical distance between *i* and *j*.[10] The *weight matrix* $\{w_{ij}\}$ contains all the weights attached to graph **G**'s edges. Note that even non-simple graphs may be encompassed by the concept of weight. Assuming that multiple edges connect vertices *i* and *j* , the

[8] See Latora & Marchiori 2001.
[9] Caldarelli & Vespignani 2007, Ch.1.
[10] In their works LM always start from this simple interpretation and then adapt it to the different applications. For example, in the case of a computer network, $w_{ij}$ is a number proportional to the time needed to transmit a unit of information through a direct link between two computers.

weight $w_{ij}$ of the *i-j* connection may in fact be taken as equal to the inverse of the number of edges between *i* and *j*. In simple, unweighted graphs it is $w_{ij} = 1, \forall i \neq j$.

In a weighted graph, the definitions of a vertex degree $k_i$ and of the shortest path $d_{ij}$ have to be modified. The *weighted degree* (also called the *strength*) of vertex *i* now becomes $k_i^w = \sum_{j \neq i} w_{ij}$, while the shortest path $d_{ij}^w$ is determined by taking into account that the length of the paths connecting any two vertices *i* and *j* cannot be determined anymore by simply counting the number of edges: each link's weight now matters too.

In the archetypal case where weights measure physical distances, $d_{ij}^w$ is the smallest total of physical distances calculated over all possible paths that connect *i* and *j*. Hence, all entries in the shortest paths matrix $\{d_{ij}^w\}$ are now determined by using the information contained in the adjacency matrix $\{a_{ij}\}$ and the weight matrix $\{w_{ij}\}$. In general, it is $d_{ij}^w \geq w_{ij}, \forall i, j$, with the equality only in case a direct link connects *i* and *j*.

Assume that every vertex sends information along the network through its edges (namely, it is a parallel network). LM define the *efficiency* in communication between *i* and *j* as a quantity inversely proportional to the shortest path between the same vertices: $\varepsilon_{ij} = 1/d_{ij}^w, \forall i, j$. This quantity is well defined even in the case of disconnected graphs, because when no path exists between *i* and *j* we have $d_{ij}^w = +\infty$ and so $\varepsilon_{ij} = 0$.[11] The average efficiency of graph $\mathbf{G}$ is defined by LM as:

$$E(\mathbf{G}) = \frac{\sum_{i \neq j \in \mathbf{G}} \varepsilon_{ij}}{N(N-1)} = \frac{\sum_{i \neq j \in \mathbf{G}} 1/d_{ij}^w}{N(N-1)}$$

and normalized by comparing it to the efficiency of the fully connected graph $\mathbf{G}^{\mathbf{ideal}}$, i.e., of the graph where all the $N(N-1)/2$ possible edges exist and thus where the information is transmitted in the most efficient way given that every vertex is directly linked to all others.

Efficiency reaches its maximum value in $\mathbf{G}^{\mathbf{ideal}}$, so much so that, considering that in such a graph $d_{ij}^w = w_{ij}, \forall i, j$, we have:

[11] This represents a clear advantage with respect to WS's notion of characteristic path length which, though clearly related to LM's efficiency, is an ill-defined quantity in the case of disconnected graphs.

$$E^{MAX} = \frac{\sum_{i \neq j \in \mathbf{G}} 1/w_{ij}}{N(N-1)}$$

This is the value used by LM to normalize $E(\mathbf{G})$. The *global efficiency* of graph $\mathbf{G}$ is then defined as: $E_{glob} = E(\mathbf{G})/E^{MAX}$. By construction, $E_{glob}$ takes values in the unit interval. It captures the network's (relative) ability to transmit information from any vertex to another.

LM also define the *local efficiency* of graph $\mathbf{G}$ by evaluating for each vertex *i* the efficiency $E(\mathbf{G_i})$ of the sub-graph $\mathbf{G_i}$ of the $k_i$ neighbors of *i*. Again, we normalize the efficiency of $\mathbf{G_i}$ by calculating the efficiency $E^{MAX}_{\mathbf{G_i}}$ of the ideal, fully connected subgraph $\mathbf{G}_i^{\mathbf{ideal}}$, where each of the $k_i(k_i-1)/2$ possible edges exists. The local efficiency of $\mathbf{G}$ is calculated as the average of the local efficiencies over all possible subgraphs:

$$E_{loc} = \frac{1}{N}\sum_{i \in \mathbf{G}} \frac{E(\mathbf{G_i})}{E^{MAX}_{\mathbf{G_i}}}$$

$E_{loc}$ too takes values in the unit interval. Since $i \notin \mathbf{G_i}$, LM's local efficiency captures a basic feature of real world networks. It tells how much a network is fault tolerant, that is to say, how efficient is the communication between the neighbors of every given *i* whenever *i* is removed from the network.[12]

## §4. Network building and collapse

It follows from the previous definitions that both the global and the local efficiency of a network grow the larger the number of its edges. A fully connected graph would exhibit $E_{glob} = E_{loc} = 1$. However, it is a matter of life that a cost must be paid whenever a link is established connecting two network members.[13] Moreover, the cost is intuitively higher the larger the weight of the connection: the

[12] LM show that WS's clustering coefficient is always a reasonable approximation of $E_{loc}$, the latter being a more general and encompassing notion than the former. As noted before, LM's main theoretical result is proving that WS's small worlds enjoy both a high $E_{glob}$ and a high $E_{loc}$.

[13] In Coasean terms (see §1), a cost arises whenever a market relationship is established between two agents.

longer the edge, or the bigger the amount of information to be transmitted, the more expensive the connection.

That each link comes with a cost is explicitly taken into account by LM,[14] who exploit this simple fact to define a subcategory of small worlds networks, called *economic small worlds*, which enjoy high global and local efficiency, as every small world, but also a relatively low cost. These networks thus exhibit three desirable properties: they are very effective in transmitting information, very resilient to fault and not excessively expensive to build. In many sense, they represent an ideal anyone called to build or manage a real world network should aim at.

In order to quantify the cost of a network, LM define the *cost evaluator* $\gamma(\cdot)$ as the function determining the cost required to build a connection between *i* and *j* of a given weight $w_{ij}$.[15] The simplest such function is $\gamma(w_{ij}) = hw_{ij}$, with $h > 0$, i.e., cost is assumed as directly proportional to weight.[16] For simplicity, we take $h = 1$.

Using the information in the adjacency matrix, and observing that in the fully connected graph this matrix has all 1s ($a_{ij} = 1, \forall i, j$), we can define the *normalized cost* of graph $\mathbf{G}$ as:

$$\Gamma(\mathbf{G}) = \frac{\sum_{i \neq j \in \mathbf{G}} a_{ij}\gamma(w_{ij})}{\sum_{i \neq j \in \mathbf{G}} \gamma(w_{ij})}$$

Even this measure takes values in the unit interval. Note that in the case of an unweighted graph, the formula becomes: $\Gamma(\mathbf{G}) = 2K/N(N-1)$, i.e., the normalized cost of building a topological network is simply the ratio between the number of its links and the maximum number of edges it may have. It follows that the cost of a fully connected, unweighted network is $\Gamma(\mathbf{G}^{\text{ideal}}) = 1$.

---

[14] See Latora & Marchiori 2002. As noted by a referee, a feature that neither LM nor the present paper consider is that connections may also have a strategic interpretation. This e.g. may happen when any pair of vertices (agents) have the option of establishing a connection between them (or with someone else) and this option entails a payoff in a game-theoretic sense. Every connection in the graph may then be interpreted as the outcome of the vertices' Nash-equilibrium play. Multiple equilibria – i.e., multiple network structures – would then be an obvious possibility and they could be Pareto-ordered also in terms of LM's efficiency notions.

[15] See Latora & Marchiori 2002. Again, they translate the general notion of weight into the intuitive one of length. Hence what the function $\gamma(\cdot)$ calculates is the cost of building a link of a given length.

[16] Recall that in non-simple graphs the weight may be set equal to the inverse of the number of connections between *i* and *j*. Given that the higher the number of connections, the smaller the weight, it follows that in such a case $\gamma(\cdot)$ must be inversely related to $w_{ij}$.

However, the cost of building a network is not exhausted by what is required to establish each link, because there is also a cost in endowing the network's vertices with the ability to "handle" the information. Call *capacity* a vertex's ability to effectively "manage" the information it receives from the other vertices and that it transmits (or re-transmits) to them.[17] This capacity too comes with a cost – the bigger, the larger is capacity itself.

In a couple of later papers, LM and their collaborators have used the notions of capacity and its cost to model the dynamic process leading to another real-world phenomenon, network collapse.[18] To this aim, they define the *load* $L_i(t)$ of a given vertex $i$ at time $t$ as the number of most efficient (i.e., shortest) paths passing through $i$ at time $t$. Once more, the intuition behind this concept is straightforward. Consider again a transportation network, where vertices represent stations, edges represent routes directly connecting stations and a journey from one station to another is constituted by the total length of the route(s) which must be traveled to do it. A most efficient journey is a journey requiring the shortest length; the load of a given station is the number of such journeys passing through that station. A properly working station is a station that can effectively handle its load without affecting the efficiency of the journeys passing through it.

The definition of $L_i(t)$ is quite flexible. Rather than in terms of the sheer number of most efficient paths passing through it, a vertex's load may be defined in terms of the total weight of those most efficient paths – for example, in terms of the number of passengers traveling through that station. In such a case *two* weights would be attached to each link/route: one representing that route's length, the other the number of passengers transported along that route. Another possibility would be to model the network as a non-simple graph, with multiple links between $i$ and $j$. For example, each link might represent a passenger travelling along that route, so much so that the route's total weight would simply equal the number of links. Also note that the load is not identical to the notion of weighted degree $k_i^w$: the latter is a static measure that captures the weight of all

[17] See below for a more exact meaning of the loose words "handle" and "manage".
[18] Crucitti, Latora & Marchiori 2004; Crucitti et al. 2004. Also see Motter & Lai 2002.

the links of vertex *i*, while the former is a dynamic notion that considers only those links belonging to the set of the network's shortest paths.

The capacity of vertex *i* may then be defined as the maximum load the vertex can carry without starting to work less effectively. Formally, $C_i = \alpha_i L_i(0)$ – where $\alpha_i \geq 1$, called the *tolerance parameter*, indicates that the vertex may carry a bigger, though still bounded, load than the initial (at $t = 0$) one. The limit on capacity is due to cost $\Omega(\alpha_i)$: the larger a vertex's capacity, the higher the cost.[19]

We can now define more precisely what it means for a vertex to properly "handle" or "manage" its load. Following Crucitti, Latora & Marchiori 2004, we model the dynamics of the network's most efficient paths as an outcome of the relationship between a vertex's capacity and its load. Consider an *unweighted* network, where the length of a path between two vertices is simply the number of links to be traveled. Define $\lambda_{ij}$ as the length of a generic path between *i* and *j* and call $e_{ij} = \frac{1}{\lambda_{ij}} \in (0,1]$ the efficiency of that path.[20] With a slight modification of LM's formalism, let's denote by $e_{ij}^x$ the efficiency of an *i*-*j* path that goes through vertex *x* (where *x* may possibly coincide with either *i* or *j*). As before, $d_{ij}$ represents the length of the most efficient path between *i* and *j*, and $\varepsilon_{ij} \in (0,1]$ measures its efficiency. If the most efficient way to connect *i* and j is through vertex *x*, then it is $e_{ij}^x = \varepsilon_{ij}$. Hence, load $L_x(t)$ denotes the number of most efficient paths between any generic pair of vertices that go through vertex *x* at time t.

At time $t > 0$, the efficiency of any path between vertex *i* and *j* that goes through *x* then is:

$$e_{ij}^x(t) = \begin{cases} e_{ij}^x(0)\dfrac{C_x}{L_x(t)} & \text{if } L_x(t) > C_x \\ e_{ij}^x(0) & \text{if } L_x(t) \leq C_x \end{cases}$$

[19] Capacity may thus be raised by increasing the tolerance parameter $\alpha_i$.

[20] Intuitively, a very long path is highly inefficient (in the limit case of disconnected vertices, it is $e_{ij} \to 0$), while a direct path has maximum efficiency ($e_{ij} = 1$).

In words: if for any reason a vertex's load exceeds it capacity, the vertex begins to "work less properly": the efficiency of every path passing through it falls proportionally to the ratio between capacity and load.[21]

Here comes the crucial observation by LM and their collaborators. Among the *i-j* paths through *x* that witness a (possibly dramatic) fall in their efficiency, also feature a number of most efficient paths (i.e., of *i-j* paths for which at time $t = 0$ it is $e_{ij}^{x} = \varepsilon_{ij}$). This means that at least some of them may *not* qualify anymore as *most* efficient: the shortest path between two vertices in the network may not be anymore the one passing through vertex *x*. Formally, at time $t = 1$, it is $e_{ij}^{x} < \varepsilon_{ij}$. If this is so, a redistribution of the most efficient paths across the network is bound to happen: the new most efficient connection between *i* and *j* at time $t = 1$ becomes that passing through, say, vertex *z*, i.e., $e_{ij}^{z} = \widehat{\varepsilon}_{ij}$. But this in turn changes the load of other vertices, some of which will witness an increase in their load. This may bring these vertices' load to exceed their respective capacity, triggering a new reduction in the paths' efficiency and a possible new redistribution of most efficient paths.

LM's model may thus explain what in complex network literature are called *cascading failures* (Watts 2002), and account for why these failures may affect even seemingly efficient networks. The cascade may be triggered by the removal of vertex *x* (say, the closure of a station). This means that all most efficient paths previously passing through *x* must now be redistributed across different vertices. A redistribution of loads among vertices takes place too, but this may create overload in some vertices. As a consequence, the efficiency of some other connections may decrease – or, worse, a newly overloaded vertex may crash down. This in turn creates a new redistribution of most efficient paths across the network and may cause new overloads, and so on, until the whole network collapses.[22]

[21] In the transportation example, the overload starts causing delays to all journeys passing through that station.

[22] Obviously, the chain of overloads must not necessarily lead to the destruction of the whole system. It is indeed possible that the initial removal of *x*, and the related redistribution of most efficient paths, is absorbed by the other vertices without any further effect, or that the effect stops at the secondary level without any tertiary effect, and so on. LM note that this apparent robustness to local failures may itself be dangerous because it may lead the network managers to neglect the possibility that the breakdown of some crucial vertices may trigger a global collapse. Given that most of the time most troubles with most vertices are harmless at system level, a belief may well arise that the network is safe – a belief which in fact is false in the case of the breakdown of crucial vertices.

How to avoid a systemic collapse triggered by vertex *x*'s breakdown? A straightforward answer is to augment vertex *x*'s capacity, but this of course entails a cost $\Omega(\alpha_x)$. Alternatively, extra links may be added to the network, with the goal of spreading the load more evenly across vertices thanks to the creation of new most efficient paths. But, again, adding links entails a cost: in the general case, the cost is $\gamma\left(w_{ij}\right)$ for each connection of weight $w_{ij}$. Hence, making a network collapse-proof requires two kinds of cost: one is the cost of building the network itself and it is proportional to the number and weight of edges; the other is the cost of endowing the network's nodes with a capacity large enough to tolerate a large – potentially, very large – load. The network efficiency and resilience to shocks depend on the two costs; generally speaking, the higher the costs, the more efficient and robust the network. However, the decision of where exactly to spend the resources to increase the network efficiency and resilience may give rise to a trade-off.

It is immediate to apply LM's notions and analysis to the economic system. Modern economies can be depicted as consisting of a myriad of complex networks. In particular, as the next § shows, it is straightforward to extend LM's ideas to the banking (or financial) system.

## §5. Network efficiency in the banking system.

Assume each vertex in a complex network is a bank and each edge a financial relation between two banks – say, an interbank loan. An edge's weight is now the amount of liquidity transferred between two banks. Creating a new edge entails a cost for the lending bank in terms of, say, acquiring information about the borrowing bank. Load and capacity are, respectively, the amount of money that passes through a given bank and the maximum amount that bank may effectively handle. Expanding the latter has a cost in terms of, say, increasing the bank's mandatory reserves and/or its personnel. LM's global efficiency measures the effectiveness with which liquidity travels throughout the system, i.e., how easy and fast it is for any bank in the system to lend money to any other. Local

efficiency captures the system's ability to guarantee the transmission of liquidity despite the collapse of one or more of its members.

The dynamics of the network collapse described in the previous § finds an obvious counterpart in the systemic effects of a liquidity crisis. Therefore, we may apply our previous analysis – in particular, LM's efficiency notions – to one of the most popular models of liquidity crisis, Allen & Gale 2000.[23] In that paper, the authors (AG thereafter) show that a *complete* – i.e., fully connected – network can achieve the first best in the interbank deposit market where banks exchange liquidity. This result is consistent with our analysis: the more interbank links exist through which money can be transferred, the more efficient the network is in reallocating liquidity – where efficiency is, as we said, the measure of how effectively liquidity can travel along the network and the notion of most efficient path translates into that of the lowest number of steps required to transfer liquidity from a bank in excess to another lacking it. A complete network, where every bank can exchange liquidity with every other, is globally efficient ($E_{glob} = 1$) because all transfers can be made with just one step.[24]

AG's main result is perhaps the thesis that, as the number of banks in the network grows, the contagion risk of financial crisis decreases, *provided the network is complete.*[25] A clear policy implication follows: a fully decentralized system where a multitude of banks exist, each one free to directly exchange liquidity with any other, is a very robust system with respect to global financial contagion. Regulators should therefore be sympathetic to systems of this kind and, ideally, favor their development. Again, AG's rationale is consistent with network efficiency notions. The initial impact of a liquidity shortage somewhere in the network (AG's "regions") becomes ever more negligible at system level (i.e., in terms of LM's global efficiency) the larger the number of nodes in the network. At the same time, and consistently with LM's local efficiency, a complete network is

[23] See Rotemberg 2008 for an alternative, thoroughly analytic, application of network theory to financially interconnected systems. Eboli 2001 is one of the seminal papers in this literature.

[24] AG also claim that the first best allocation of liquidity can be achieved even under an incomplete network structure. However, while complete structures have several equilibria, i.e., there are several ways to efficiently allocate liquidity between banks when all links exist (this again is obviously consistent with our formalism), the incomplete structure analyzed in their paper has only one equilibrium.

[25] In the AG model the unique equilibrium in the case of an incomplete network (see previous footnote) corresponds to the pattern of liquidity allocation that, in case of a liquidity shock, may trigger a contagion and lead to the collapse of the whole system. On the contrary, in the complete network setup there are equilibria with zero risk of contagion, as well as others with a positive contagion risk.

also very efficient in each of its subparts. The intuition is that in such a case liquidity will always find a way to travel through the whole network – or a part of it – even if a link or a node is wiped off, i.e., even in the case of a bank's default or bankruptcy.

On the contrary, when the network is *incomplete* AG argue that small shocks in a region can trigger large effects, the bigger the higher the number of nodes. In other words, a banking system where a multitude of agents operate, but some agents may for any reason be unable to directly exchange liquidity with some others, is especially subject to systemic risk. Our analysis shows that this claim is not entirely correct, because the global outcome of a local shock depends on the local efficiency of the sub-network where the shock has taken place.

As we argued in §4, when local efficiency is high a sub-network may still work properly even if a node or a link is deleted. That is to say, in a locally efficient region liquidity may still find a way to circulate even if a "regional" bank breaks down. This would prevent a crisis to spread from that region to the whole system, despite the latter's incompleteness. The result goes beyond the limits of AG's analysis. As we argued in §1, a trade-off exists between the typical opposing effects of higher interconnectedness in banking or financial systems, namely, the increasing contagion risk and the larger cushioning potential. By providing an exact measure of a sub-network's resilience to local shocks, LM's notion of local efficiency may allow a specific determination of which of the two effects prevails in any given network.

Network analysis also improves with respect to AG model on the side of network costs. As AG themselves recognize, every link has a cost: "The banking sector is interconnected in a variety of ways, but transaction and information costs may prevent banks from acquiring claims on banks in remote regions. To the extent that banks specialize in particular areas of business or have closer connections with banks that operate in the same geographical or political unit, deposits may tend to be concentrated in 'neighboring' banks" (Allen & Gale 2000, p.13). This is a specific instance of the Coasean "cost of using the price mechanism" – and a reasonable one as well, because a bank would not trust another bank with which it has no other financial relations and so might prefer not to exchange liquidity with it rather than pay the price of acquiring information on its financial situation and business reliability. The formalism presented in §4 improves with respect to

this basic intuition by highlighting the trade-off between the cost of building a liquidity network and the efficiency with which liquidity may travel in the system. An immediate implication of the trade-off is that, quite obviously, the network will always be incomplete.[26]

Moreover, we may now also take into account the capacity of each bank to manage the flow of liquidity passing through it. LM's notion of load – which measures the number of most efficient paths passing through a node – can be applied to measure the amount of money intermediated by a single bank. Each node has a bounded capacity, i.e., a maximum amount of liquidity it can intermediate. Capacity is bounded because it is costly to build. This feature captures, as in AG model, the opportunity cost of liquidity in terms of the higher return that would accrue to the bank investing in long assets rather than in short, liquid ones. Thus, increasing the capacity to transmit liquidity through the system is costly – indeed, *very* costly if a bank is forced to undersell its long assets at a very low price.

Finally, our dynamics of the capacity-to-load ratio explains in very general terms how a liquidity crisis can spread through the system.[27] The fall of a node's efficiency depicts the collapse in a bank's ability to promptly provide liquidity to everyone entitled to it. Hence, liquidity will find new, most efficient paths to circulate in the system, but this may trigger a crisis in another bank that, in turn, cannot handle (i.e., lacks the capacity to sustain) an increase in its liquidity requirements, and, at the same time, cannot expand its capacity without losing a big part of its long assets' value. A cascade effect may follow.

Yet, LM's formalism shows that the breakdown may be avoided if either, as in AG, the network is fully connected (maximum global efficiency) or, at least, the neighbors of the node where the crisis has begun are themselves fully connected (high local efficiency). The latter is the most interesting case because it shows that it is possible to insulate a potentially systemic crisis by creating locally efficient sub-networks of banks that can guarantee the excess liquidity a single

[26] Iori et al. 2008 show that the Italian overnight money market is indeed incomplete, with a few big banks trading with (actually, borrowing from) many, usually small, counterparties, while the majority of banks trade with a few partners.
[27] Iori, Jafarey & Padilla 2006 and Nier et al. 2007 offer several simulations of default and contagion risks in banking systems modeled as complex networks.

network member is not able anymore to provide.[28] Alternatively, a crisis can be stopped when there exist, in well-defined areas of the network, one or more nodes with a large excess capacity. This means one or more banks that must accept to keep idle an extra amount of liquidity – bearing the relative opportunity cost – only to be able to use it whenever a crisis erupts, i.e., whenever a failure somewhere in the network attracts towards that node a number of new most efficient paths (viz., new liquidity flows). Ideally, the presence of one such bank, endowed with the proper amount of excess liquidity, in every not-fully-connected sub-network could guarantee that a financial crisis would always remain confined within the sub-network itself. In short, our formalism provides a micro-structural rationale for the role of "money centers" (see §1) as backup providers of liquidity to the interbank market.

## §6. Conclusion

As we argued in the Introduction, the classic "market vs firm" dichotomy of Coase 1937 can be properly analyzed only when all its elements are taken into account, namely, contracting costs, effectiveness of inter-agent communication, systemic resilience and managerial ability. To this aim, and given the crucial theoretical role played by the Coasean dichotomy, the paper suggests that attention should be paid to complex network analysis, and in particular to the new notions of network efficiency and network costs introduced by physicists Latora, Marchiori and their collaborators.

The new tools' potential is not just theoretical, though. Cost and efficiency play a crucial role in the design and management of real networks. It is claimed that the same kind of analysis that is commonly performed in the case of physical systems, such as transport or computer ones, should also be undertaken for economic networks, such as, for instance, the interbank market. In the latter case, the new efficiency notions allow a straightforward improvement with respect to the well-known analysis of financial crisis in Allen & Gale 2000. In particular,

[28] The cost of establishing all the necessary links in a financial *sub*-network should be generally low. Intuitively, it should be less costly for a bank (in terms of, say, information gathering) to exchange liquidity with a neighboring partner than with a very distant one.

either locally efficient sub-networks of banks or the establishment of local "money centers" emerge as possible solutions to minimize the risk of systemic meltdown.

Still, the paper's most general implication goes beyond the banking system. Thanks to complex network theory, the "market vs firm" dichotomy appears even more fundamental than commonly understood. It actually lies behind the most efficient planning and implementation of every system of interconnected agents – that is to say, almost everywhere in modern, network-based economies and production processes.

## References


ALLEN F. & GALE D. 2000, "Financial contagion", *Journal of Political Economy*, 108:1, 1-33.

CABALLERO R.J. 2010, "Macroeconomics after the crisis: time to deal with the pretense-of-knowledge syndrome", *Journal of Economic Perspectives*, 24:4, 85-102.

CALDARELLI G. 2007, *Scale-Free Networks. Complex Webs in Nature and Technology*, Oxford: Oxford University Press.

CALDARELLI G. & VESPIGNANI A. 2007, *Large Scale Structure and Dynamics of Complex Networks*, World Scientific Press.

COASE R.H. 1937, "The nature of the firm", *Economica*, 4:16, 386-405.

CRUCITTI P., LATORA V. & MARCHIORI M. 2004, "A model for cascading failures in complex networks", arXiv:cond-math/0309141v2 [cond-mat.other]

CRUCITTI P., LATORA V., MARCHIORI M. & RAPISARDA A. 2003, "Efficiency of scale-free networks: error and attack tolerance", *Physica A*, 320, 622-642.

CRUCITTI P., LATORA V., MARCHIORI M. & RAPISARDA A. 2004, "Error and attack tolerance of complex networks", *Physica A*, 340, 388-394.

EBOLI M. 2001, "Systemic risk in financial networks: a graph-theoretic approach", unpublished paper.

IORI G., JAFAREY S. & PADILLA F.G. 2006, "Systemic risk on the interbank market", *Journal of Economic Behavior and Organization*, 61, 525-542.

IORI G., DE MASI G., PRECUP O.V., GABBI G., CALDARELLI G. 2008, "A network analysis of the Italian overnight money market", *Journal of Economic Dynamics & Control*, 32, 259-278.

LATORA V. & MARCHIORI M. 2001, "Efficient behaviour of small-world networks", Physical Review Letters, 87:19, 198701(4).

LATORA V. & MARCHIORI M. 2002, "Economic small-world behavior in weighted networks", arXiv:cond-math/0204089v2 [cond-math.stat-mech].
MOTTER A.E. & LAI Y.-C. 2002, "Cascade-based attacks on complex networks", *Physical Review E*, 66, 065102(R).
NEWMAN M.E.J. 2005, "Random graphs as models of networks", in: Bornholdt S. & Schuster H.G. (eds), *Handbook of Graphs and Networks: From the Genome to the Internet*, Weinheim: Wiley-VCH Verlag.
NIER E., YANG J., YORULMAZER T. & ALENTORN A. 2007, "Network models in financial stability", *Journal of Economic Dynamics & Control*, 31, 2033-2060.
ROTEMBERG J.J. 2008, "Liquidity needs in economies with interconnected financial obligations", *NBER Working Paper* #14222.
WATTS D.J. & STROGATZ S.H. 1998, "Collective dynamics of 'small-world' networks", *Nature*, 393, 440-442.