Exploration into the network structure of the Economy

Frank Verschaeren[1]

Paper prepared for the European Establishment Statistics Workshop,
Poznań, Poland, September 7-9 2015

Abstract: *A statistical approach to the network structure of the economy has received little attention in the world of official statistics. The core of business statistics aims at revealing macro phenomena by extrapolating the characteristics of a "representative agent" to the entire population rather than paying attention to the interaction within populations of heterogeneous actors.*

In the European Statistical System ownership relations between units (the link between units rather than the units themselves) are recorded as relevant characteristics to be measured in Foreign Direct Investment Statistics (FDI), Foreign Affiliates Statistics (FATS), in Business Registers in building enterprise group structures and in Structural Business Statistics (SBS, annex 9) in the determination of "demographic events". Ownership relations also play a role in the delineation of "enterprise" units.

The academic world has already been looking at the economy as a complex network. Chaos, complexity and entropy as mathematical concepts in the realm of statistical mechanics have been applied to the network structure of production, with inter alia special attention to the buyer-supplier relationship between base units.

With the increase of computational power available to the statistical institutes, it becomes possible to test different network approaches and to compare different attempts to describe connected businesses as "evolving networks". Progress made in the field of statistical mechanics of complex networks could be helpful in getting a better understanding of e.g. the occurrence of power law distributions. One useful application of mapping a network structure seems to be that it facilitates assessment of a system's vulnerability to shocks.

Power Laws and Establishment Statistics

In the 1830s the Belgian statistician Adolphe Quetelet developed a theory of "l'homme moyen" (the average man), applying the Gaussian distribution to the study of all kinds of human characteristics. In 1890 Alfred Marshall, one of the most influential economists of his time, introduced the concept of "representative firm". Nowadays the term "representative agent" refers to the typical decision-maker of a certain type (for example, the typical consumer, or the typical firm). If this agent is truly representative, then his behaviour will be reflected in the economy. A model contains representative agents while agents may differ individually but act in such a way that the sum of their choices is mathematically equivalent to the decision of a (random) subset of agents.

The assumption that every unit under study is alike is extremely convenient for relating individual to aggregate behaviour. In statistical and economic models normality is often assumed. In practice [1] the

[1] frank.verschaeren@economie.fgov.be The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Belgium.

distribution of data is found to be different more frequently than expected. Social and natural scientists have drawn the attention to the prevalence of power law distributions.

Zipf, Pareto and power law distributions are closely related:

The Zipfian distribution is one of a family of related discrete power law probability distributions. Zipf's law states that the size of the r'th largest occurrence of an event is inversely proportional to its rank (e.g. the frequency of any word is inversely proportional to its rank in a frequency table, "the" 7%, "of" 3.5%...). :

$$\mathbf{y \sim r^{-b}} \qquad \text{with b approximately equal to 1.}$$

Pareto's law is formulated in terms of the cumulative distribution function (e.g. how many people have an income higher than x):

$$\mathbf{P[X > x] \sim x^{-k}}$$

Most of the time a power law distribution refers to the probability distribution function associated with the cumulative distribution function given by Pareto's law (e.g. number of people with income x):

$$\mathbf{P[X = x] \sim x^{-(k+1)} = x^{-a}}$$

The probability that random variable X exceeds some level x is proportional to $1/x^k$. In other words the probability of X being large is much higher than in a normal distribution (in a normal distribution the probability of large events decays exponentially with their size, making large events increasingly rare at a rapid rate). When k takes a low value, the tail of the distribution is fat.

Power law-distributed data only have a well-defined mean when $a > 2$ and a finite variance only exists when $a > 3$. This makes it incorrect to apply traditional statistics that are based on variance and standard deviation (such as regression analysis), central limit theorem may not apply to power law-distributed variables, there is no "regression to the mean" if the mean is ill-defined and the variance unbounded.

Examples:

- Word frequency[2]
- Web hits[3] (cumulative distribution of the nr of hits received by web sites during a single day)
- Citations of scientific papers
- Email messages that people send and receive[4]
- Populations of cities
- Income and wealth distribution

…and many more. A more in-depth review can be found in Gabaix (2009), Power Laws in Economics and Finance.[5]

Scientific interest in power-law distributions is related to the possible mechanisms that underlie the phenomenon; they are often thought to be the result of specific stochastic processes. Clearly there is also an interest from the study of probability distributions, the fat upper tail having to do with the frequency of extremely rare events like stock market crashes and large natural disasters.

The linear models used to forecast economic conditions have great difficulties in coping with sudden changes. In January 2008, real GDP growth in 2009 was expected to be around 2% in the Eurozone and 2.5% in the U.S.. By January 2009 these forecasts were -1.5% and -2 %. Instability in the global credit markets prompted the worst economic crisis for 70 years. The credit network had become so highly interlinked that each participant faced high risks from the possible collapse of their partners. Traditional central bank controls had focused on the health of individual institutions rather than on networks of many institutions in interaction. In the opinion of network scientists[6] “The banking failure is, in many respects, a failure of economic science to have any well-developed understanding of the financial system as a complex dynamical network. (…) regulators will need to take a more holistic view, monitoring the nature of the links between institutions and the overall stability of the credit network.”

Describing large scale behaviour of a system is always done by simplifying the mathematical description of the system, so that there are only a very limited set of possible behaviours that can happen. This idea is used in traditional theory by using the normal distribution for many different biological and social systems. It is possible because when a system has independent parts, the way they aggregate is the same, and the result is the normal distribution as the largest scale behaviour of the system. When there are dependencies, the normal distribution no longer applies, but there are behaviours that are characteristic of other kinds of dependencies. The idea of universality recognises that systems map onto a small set of large scale models, each of which applies to a large set of possible systems with widely different micro details.[8]

With the advance of network science, interest has grown in approaches that go beyond traditional methods. The United States National Research Council defines network science as "the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena." Economic-, trade- and financial networks can also be seen as complex systems[7].

In complex systems, the units are acting neither totally independently nor totally coherently; rather, they are interdependent, both influencing each other and compelled by common causes. An example can be found in commodity markets. The traditional theory of markets assumes that people decide on investments independently and rationally, and therefore predicts a supply and demand equilibrium. Interestingly, it is not so much the assumption of rationality that does not hold up in complex systems analyses of markets today, but rather the independence. The breakdown of equilibrium due to trendfollowing has been well-established since 1990, but the theory at that point, subject to the constraints of the concepts and mathematics of traditional economics, was not able to represent the dynamics after the breakdown.[8]

The study of systems of interdependent elements implies making use of existing knowledge in mechanics and statistics at the same time, i.e. statistical mechanics (or thermodynamics). Advances in understanding the scaling properties of evolving networks have benefited from concepts like e.g. nucleation theory and gelation.[9]

Complex systems

Traditionally the study of complex networks has been the territory of graph theory. Since the 1950s large-scale networks have been described as random graphs. The most common type is the Erdős-Rényi random graph. In the model, we start with N nodes and connect every pair of nodes with probability p, creating a graph with approximately pN(N-1)/2 edges distributed randomly.

Many real-world networks display similar topological properties, strikingly different from those shown by random graphs. This motivated the search for theoretical models aimed at understanding the mechanisms at the basis of network organization.

Albert-László Barabási introduced[10] in 1999 the concept of scale-free networks and proposed the Barabási–Albert model to explain their widespread emergence. A common property of many observed networks is that the node connectivities follow a scale-free power-law distribution. His model is based on two generic mechanisms: (i) networks expand continuously by the addition of new vertices, and (ii) new vertices attach preferentially to sites that are already well connected. It reproduces the observed stationary scale-free distributions, which indicates that the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems.

A second class of models is based on the hidden variable hypothesis: vertices are assumed to be characterized by an intrinsic quality or fitness that determines their connection probability. The model has been applied[11] for instance on the world wide network formed by trade relationships, with the GDP of each country as fitness variable.

Network models are defined by a number of properties that may or may not agree with empirical results on real networks. The most common properties that are seen as robust measures of a network topology :

Size: usually the number of nodes (N).

Average degree: degree k of a node is the number of edges connected to it.

Let E be the number of edges.

The average degree $\langle k \rangle = 2E / N$

The average number of edges is not always meaningful. In networks with power-law degree distributions, most of the nodes are of low degree, but there are also highly-linked nodes (nodes of high degree) called “hubs.”

Average path length: average shortest distance between any two nodes in the network

Consider a network G with a set of nodes (vertices) V. Let $dist(v_1,v_2)$ denote the shortest distance between v_1 and v_2 ($v_1,v_2 \in V$). Assume $has\_path(v_1,v_2) = 0$ if $v_1 = v_2$ or when v_2 cannot be reached from v_1 and $has\_path(v_1,v_2) = 1$ in other cases. The average shortest-path length $ASPL_G$ is:

$$\text{ASPL}_G = \frac{\sum_{i,j}^{N} dist(v_i,v_j)}{\sum_{i,j}^{N} has\_path(v_i,v_j)}$$

Clustering coefficient:

Sometimes described as "all-my-friends-know-each-other". The clustering coefficient of node i is:

$$C_i = \frac{2e_i}{k_i(k_i-1)}$$

where k_i is the number of neighbours of node I and e_i the number of connections between these neighbours.

The clustering coefficient of the entire network is the average of the clustering coefficients of all the nodes.

These global properties give an overall view of the network, but might not be detailed enough to capture complex topological characteristics of large networks, they are weak predictors of network structure. (networks can have the same size and degree distribution but still have a very different graph structure)

Another type of properties has gathered much attention as useful concepts to uncover design principles of complex networks: local network properties. Research has shown the existence of recurring small graph structures in many types of real networks. These recurring subgraph patterns are called network motifs, graphlets or more simply subgraphs. It has also been shown that complex networks can be compared and classified into distinct functional families, based on their typical motifs

Motifs:

A network motif is a small over-represented partial subgraph in a real network. The “null-model” to identify motifs as over-represented can be a random graph model, e.g. Erdos-Renyi random graph. Research demonstrated for instance that the motifs shared by ecological food webs were distinct from the motifs shared by the genetic networks of Escherichia coli and Saccharomyces cerevisiae or from those found in the World Wide Web. Motifs may reflect functional properties within networks.

Graphlets:

Graphlets are small connected non-isomorphic sub-graphs of a graph G induced on $n \geq 3$ nodes of G (there are e.g. 21 graphlets for $n = 5$). They don’t have to be over-represented: graphlet degree vectors (signatures) and signature similarities are compared. This has been successfully applied to biological networks to identify groups (or clusters) of topologically similar nodes in a network and predict biological properties of yet uncharacterized nodes based on known biological properties of characterized nodes[12].

Economic networks

Robert Gibrat, a researcher in economics, published in 1931 “Inégalités Economiques”[13], a book that became the starting point of one of the most important strands in the literature on market structure, and in which he presented a formal model of the dynamics of firm size and industry structure (‘Gibrat’s law). He described a stochastic process in which[14] :

a. the expected increment to a firm’s size in each period is proportional to the current firm size
b. the increments have no temporal correlation (random walk)
c. there is no interaction between firms

With growing availability of data and processing capacity, it has become clear that these assumptions do not hold empirically[15] , but Gibrat’s model is still prominent in literature on the topic[16], and is often used as a benchmark. Over the last decades, various factors that could possibly influence firm growth have been considered as well as different definitions of what represents growth (e.g. employment, cost of goods sold, assets,…). At the same time the interaction between the elements in the economic network became an element of interest.

In game theory the “random growth” element is replaced by a process in which firms that differ in various attributes make different choices. As the number of choices grows very fast, the focus is on the simplest topologies(e.g. cycles of four agents)where agents decide to add or delete links between them. An Example is the interplay of cooperation and competition in the formation of innovation networks[17].

The network view pays more attention to the network as a whole and statistical properties as results of interaction. Buyer-seller networks, ownership networks, supply chain hierarchies,… The focus being on their “wiring diagram” and their functional properties like stability and information processing capacity. In this view, interaction means that it is not possible to define a representative agent because the dynamics of the system is originated just from the interaction among heterogeneous agents.

The notion of economic networks can be traced back to the 1930s[18], although the nodes used in the Leontief analysis at that time are economic sectors rather than individual firms. Within the European Statistical System, this is reflected in the production of input output tables. While these tables are based on products or industries, a number of events have shown that the analysis of aggregated data fails to take into account the economic dependencies between the entities at a lower level (e.g. firms, enterprises,…). With regard to supply chains for instance, an earthquake and tsunami in Japan caused disruptions that affected economic activity in many other countries. Microeconomic shocks (idiosyncratic shocks to individual firms) are not evened out by the law of large numbers as was believed, but can propagate to the rest of the economy through production chains, leading to fluctuations in production. On a larger scale, the economic crisis also highlighted the importance of interconnections between firms. Both the spread of the risks emanating from toxic assets on the balance sheets of several financial institutions to the rest of the financial sector, and the transmission of the economic problems to the rest of the economy have been linked to such interconnections.

Within the 7 th Framework Programme for Research and Technological Development (FP7) a project [19] called CRISIS ran from 2011 to 2014. It was conducted by a consortium of ten universities and a number of central banks collaborating to develop a next generation macroeconomic-financial system model using "agent-based" modelling to be able to model crises in a more realistic and bottom-up way than has been possible previously. The CRISIS model was intended to enable central bankers and other policymakers to explore policies for managing systemic risk to reduce the chances of future crises and minimise their impact. The consortium states on its website:

> *The goal is to build new models, one for the EU financial system and one for its macroeconomy. When they are completed, CRISIS will turn the models into software that can be used by central banks and governments. For that reason, CRISIS will work in collaboration with major central banks, government economic and finance ministries, and with multilateral institutions.*

The latest insights on how we should produce business and economic statistics are a game changer. What we have done until now is based on three assumptions:

a. Households, firms, and governments are perfectly rational and tend to behave in similar ways to each other
b. The economy settles into a balanced "equilibrium" state
c. The detailed institutional structures and interconnections of the system do not generally matter for macroeconomic policy

The fact that an important and growing part of the users of official statistics claim that all three assumptions are wrong can have far reaching consequences for the future. If we continue without taking notice of the changes, our work might become obsolete.

There are also opportunities that come along with the shift from aggregating data from "representative agents" to describing network properties that emerge from the interaction of elements in a network. A range of statistics and methods that have as yet no place within the European Statistical System are there to explore.

Some of the statistics in the European Statistical System incorporate to a certain extent a network view. For instance ownership relations between units (the link between units, or edges) are recorded as relevant characteristics to be measured in Foreign Direct Investment Statistics (FDI) and Foreign Affiliates Statistics (FATS). The enterprise group structures maintained in the Business Register are essentially graphs representing a control relationship.

In these examples cited in the previous paragraph, we can see that a specific network model is used: the pyramid model, with the top of the pyramid directly or indirectly controlling the units at a lower level of the group structure. A possible extension could be found for instance in analyzing a number of frequently observed complex ownership patterns to better understand their role in these ownership networks and to predict properties of certain units from the properties of units in similar sub-graphs.

[1] Haldane A.G.,Nelson B. (2012); Tails of the unexpected; www.bankofengland.co.uk/publications/Pages/speeches/default.aspx
[2] G. K. Zipf (1949); Human Behaviour and the Principle of Least Effort; Addison-Wesley; Reading, MA.

[3] L. A. Adamic and B. A. Huberman (2000); The nature of markets in the World Wide Web; Quarterly Journal of Electronic Commerce 1, 512.
[4] H. Ebel, L. Mielsch, and S. Bornholdt (2002);Scale-free topology of e-mail networks; Phys. Rev. E 66, 035103.
[5] X. Gabaix; (2009); Power Laws in Economics and Finance; Annu. Rev. Econ. 2009. 1:255–93
[6] M. Buchanan, G. Caldarelli (2011); Perspectives on Complex Networks; EOLSS, Developed under the Auspices of the UNESCO, Eolss Publishers, Paris, France
[7] Y. Bar-Yam (1997); Dynamics of Complex Systems; Perseus Press, Reading
[8] Y. Bar-Yam, M Bialik (2013); Beyond Big Data: Identifying important information for real world challenges; http://necsi.edu/projects/yaneer/information/infoimportance.pdf
[9] R. Albert, A.L.Barbarasi (2002); Statistical Mechanics of Complex Networks; Rev. Mod. Phys.
[10] A.L. Barabási,R. Albert (1999); Emergence of Scaling in Random Networks; October 15, 1999; Science, 509-512
[11] D. Garlaschelli et. al. (2004) ; Fitness-dependent topological properties of the World Trade Web; PHYSICAL REVIEW LETTERS 93, 188701
[12] T Milenkovic, N. Przulj (2007) ; Uncovering Biological Network Function via Graphlet Degree Signatures http://www3.nd.edu/~tmilenko/papers/CI_08_GDD_sign.pdf
[13] Gibrat, R. (1931) ; Les inégalités économiques ; Librairie du Recueil Sirey.
[14] J. Suttom (1997) ; Gibrat's Legacy ; March 1997 ;Journal of Economic Literature, 40-59
[15] M. Stanley,et al.(1996); "Scaling Behavior in the Growth of Companies," Nature 379, 804-806
[16] Riccaboni et al. (2008); The size variance relationship of business firm growth rates. Proc. Natl. Acad. Sci. USA 105:19595–600
[17] M. D. König, S Battiston, F Schweitzer (2008) Modeling Evolving Innovation Networks; In: Innovation Networks - New Approaches in Modeling and Analyzing; Springer
[18] W. Leontief (1936); Quantitative Input and Output Relations in the Economic System of the United States. Review of Economics and Statistics 18(3), 105-125.
[19] http://www.crisis-economics.eu/