The application of network theory in official statistics

Áron Kincses

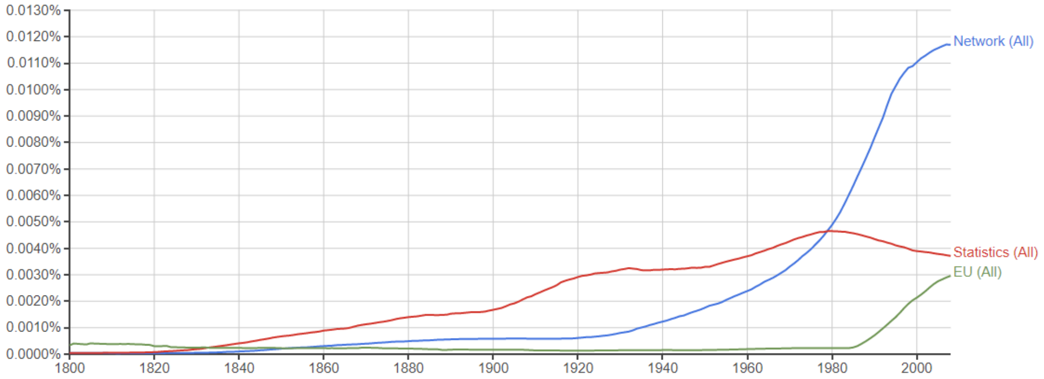
Deputy President, Hungarian Central Statistical Office, Hungary

Abstract. The challenges faced by official statistics in the 21st century are manifold. We are surrounded by systems that are becoming substantially more and more complex. The emergence of new phenomena, namely, globalisation, digitalisation, global demographic trends and sustainable development, added to the complex realities that need to be meaningfully and timely captured by official statistics, have resulted in the development of new patterns, routes and types of data, offering us with the opportunity to further improve the relevance of statistics. In response to these trends we need to find new, usable tools and methods for the measurement of such changing phenomena. Network theory is an innovative tool and approach in our changing world that can help us handle the complexity of the 21st century. However, so far it has not featured in mainstream official statistics.

1. Networks

The following plot indicates the frequency of use of the words *EU*, *statistics*, and *network* in published books since 1800, which reveals a rise in public awareness of networks during the past decades.

1. Figure: The Rise of Networks[1]



Network analysis has come to the foreground of interest for a very simple reason: the 21st century requires new, timely and useable tools and methods, capable of capturing the essence of new phenomena and complex realities in a *simple fashion*. Network theory can fulfil this role [13].

In order to gain understanding of a complex system [10], we first need to know the ways in which its components interact with one another. A network is a catalogue of a system's components, often

[1] The plots were generated by Google's Ngram platform (https://en.wikipedia.org/wiki/Google_Ngram_Viewer).

called *nodes*, and the direct interactions between them, or *links*. This network representation offers a common language for the study of systems that may differ greatly in nature, appearance, or scope [1]. The way in which we define the links between two individuals will determine the nature of the questions we can explore [12]. For instance, social networks reveal the spread of knowledge, news or behaviours. Communication networks, which describe the interactions between communication devices, are at the centre of the modern communication system. Business networks are owned by business enterprises, where the aim of the network is to support the informational and operational requirements of the business, such as in sales or manufacturing issues [11]. The variety of relationships within multinational enterprises, or MNEs, between parent companies and affiliations underpin the importance of dynamic capabilities in the global market. These systems are collectively called complex systems, and they capture the fact that it is difficult to derive collective behaviour only by knowing the system's components [3].

1. Table: ***Simplified Network Maps***

Network	**Nodes**	**Links**	**Directed / Undirected**[2]	**Average degree <k>**
Internet	Routers	Internet connections	Undirected	6.34
WWW	Webpages	Links	Directed	4.60
Power Grid	Power plants, transformers	Cables	Undirected	2.67
Mobile-Phone Calls	Subscribers	Calls	Directed	2.51
Email	Email addresses	Emails	Directed	1.81
Science Collaboration	Scientists	Co-authorships	Undirected	8.08
Actor Network	Actors	Co-acting	Undirected	83.71
Protein Interactions	Proteins	Binding interactions	Undirected	2.90

Source: http://networksciencebook.com/

2. The nature of networks

The degree of nodes represent the number of links a given node has to other nodes. The degree distribution (p_k) has a central role in network theory. The reason is that the precise functional form of p_k determines many network phenomena, from network robustness to the ability to evolve. The average degree of a network can be expressed as:

$\langle k \rangle = \sum_{k=1}^{\infty} k * p_k$, where $\sum_{p=1}^{\infty} p_k = 1$ and $p_k = \frac{N_k}{N}$ (N_k is the number of degree-k nodes[3]).[4]

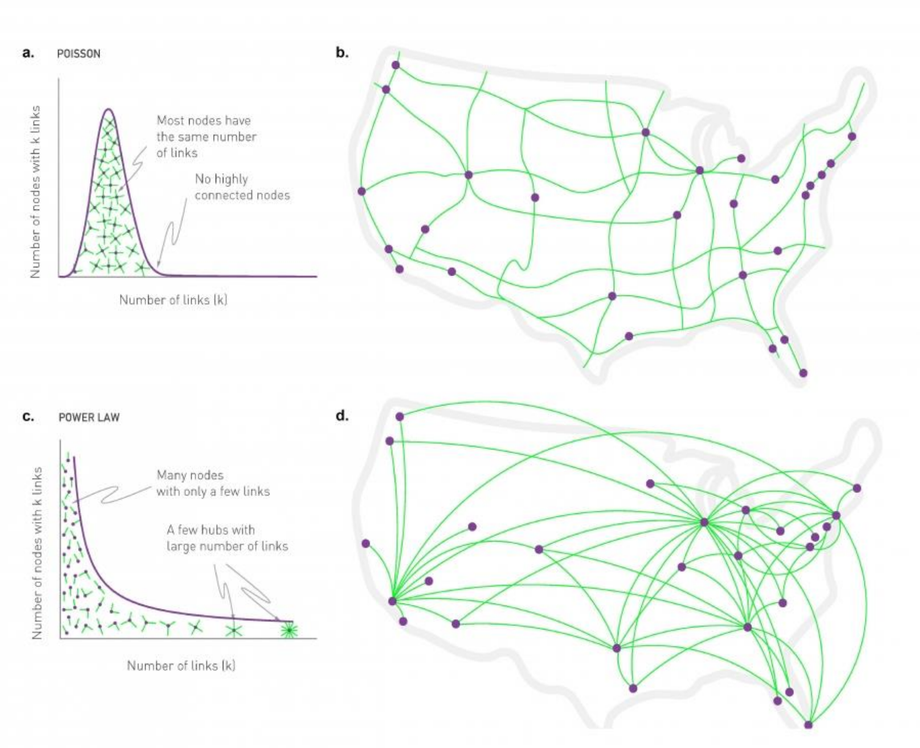
[2] Some systems have directed links, such as phone calls, where one person calls the other. Other systems have undirected links, such as transmission lines in the power grid, where the electric current can flow in both directions.

[3] $N_k = N * p_k$

[4] Real Networks are supercritical: Once the average degree exceeds ‹*k*› = 1, a giant component should emerge that contains a finite fraction of all nodes. Hence only for ‹*k*› › 1, the nodes organize themselves into a recognizable network. For ‹*k*› › ln*N* all components are absorbed by the giant component, resulting in a single connected network.

Based on degree distribution, we can theoretically differentiate between two types of networks: random and scale-free networks [2].

2. Figure: Random versus scale-free networks[5]


a.
POISSON
Number of nodes with k links
Most nodes have the same number of links
No highly connected nodes
Number of links (k)
b.
c.
POWER LAW
Number of nodes with k links
Many nodes with only a few links
A few hubs with large number of links
Number of links (k)
d.

Source: http://networksciencebook.com/

The degrees of a random network follow the Poisson distribution, similar to a bell curve. Therefore, most nodes have similar number of degrees and nodes with a large number of links do not exist. The model suggests that they should be described as purely random. A random network looks somewhat like the national highway network, in which nodes are cities and links are major highways. There are no cities connected to hundreds of highways, and no city is disconnected from the highway system [3].

In a network with power-law degree distribution, most nodes have only a few links. These numerous small nodes are held together by a few highly connected hubs [7]. A scale-free network looks like the air-traffic network, where nodes are airports and links are the direct flights between them. Most airports are tiny, with only a few flights. In this network, however, we can reach most destinations via single hubs, like Chicago. Airlines deliberately build hubs to decrease the number

[5] Poisson-distribution: $p_k = e^{-\langle k \rangle} * \frac{\langle k \rangle^k}{k!}$

Power law distribution: $p_k = \frac{k^{-\gamma}}{\zeta(\gamma)}$, where $\zeta(\gamma)$ is the Riemann-zeta function: $\zeta(\gamma) = \sum_{k=1}^{\infty} k^{-\gamma}$ [8].

(more about this function see: http://mathworld.wolfram.com/RiemannZetaFunction.html) of transfers between two airports. Hubs affect the “small world nature”. The distances in a scale-free network seem smaller than the distances observed in a similar, but randomly arranged network. These networks are characterized by the small world phenomenon, thus the distance between two randomly chosen nodes in a network is short. So, we are always close to hubs. Once hubs are present, they fundamentally change the system’s behaviour [3; 4].

The key difference between random and scale-free networks is rooted in the different shapes of the Poisson and that of the power-law function: random networks have a scale. In other words, nodes in a random network have comparable degrees, and the average degree ⟨k⟩ serves as the “scale” of a random network. Scale-free networks lack a scale; thus, the average degree does not orient us so much, and this means that when we randomly choose a node, we do not know what to expect. The selected node’s degree could be tiny or arbitrarily large. Hence, networks do not have a meaningful internal scale, but are “scale-free” [3]. The presence of hubs and the small world phenomenon are *universal* characteristics of the scale-free network.

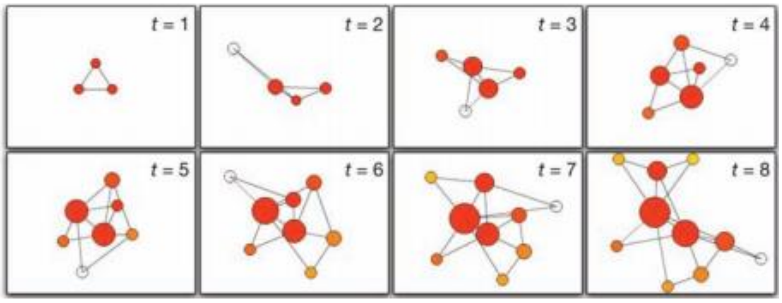
But *not all networks are scale-free***.** On the contrary, several important networks do not share these features. Networks appear in materials science, and describe the bonds between the atoms in crystalline or amorphous materials. In these networks, each node has exactly the same degree, determined by the laws of chemistry. A carbon atom can share only four electrons with other atoms; regardless of how we arrange these atoms relative to each other, in the resulting network a node can never have more than four links. So the diamond and the graphite networks do not constitute scale-free phenomenon [3].

The main question is: what causes the development of scale-free networks? The growth and preferential[6] attachment are jointly responsible for the scale-free feature [1; 6; 9].

The simplest process that can produce a scale-free topology is the following: starting from three connected nodes (top left), in each image a new node (shown as an empty circle) is added to the network. When deciding where to link, new nodes prefer to attach to more connected nodes, a process known as preferential attachment. Thanks to the growth and preferential attachment, a rich-gets-richer process is observed, meaning that highly-connected nodes acquire more links than those less-connected, leading to the natural emergence of a few highly-connected hubs. Node size, which is presented proportional to the node’s degree, illustrates the natural emergence of hubs as the largest nodes. The degree distribution of the resulting network follows the power law [3].

[6] The likelihood of connecting to a node depends on that node’s degree *k*.

3. Figure: The birth of a scale-free network.



Source: http://science.sciencemag.org/content/325/5939/412/F1

3. Networks in statistics and their usability

Official statistics offer a new field to harvest the results of network theory. Network analysis offers us the opportunity to better understand the processes of globalization beyond the figures and improve the relevance and quality of official statistics. Through examples I provide some of the most important tangible outcomes of network analysis in official statistics (including usability, degree distribution and consequence). The following networks disseminations and calculations were made with UCINET's NetDraw software [5].

2. Table: Examined networks overview

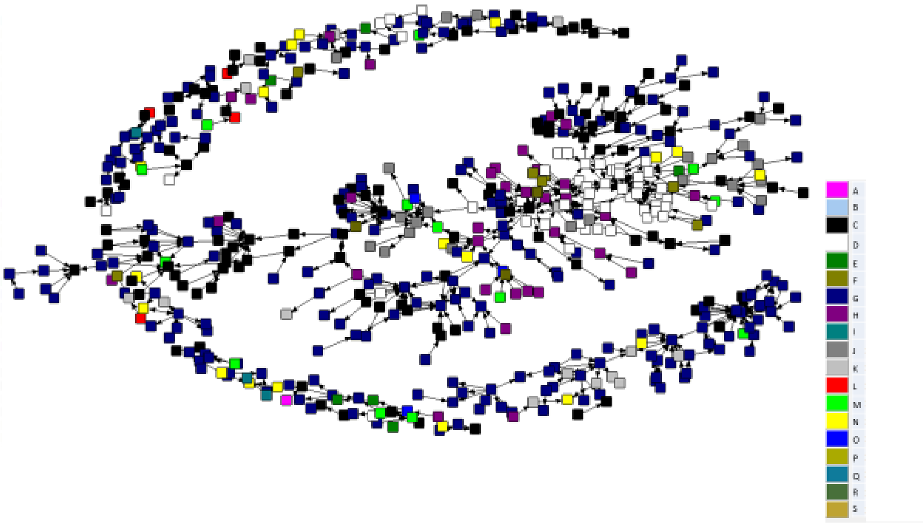
Network	**Nodes**	**Links**	**Nodes**
Companies’ sales	Top 1000 companies in Hungary (according to domestic sales)	Sale	Data referring to yearly total sales in 2016 (more than 1 million forint/transaction)
International migration	Counties	migration	UN migration database, total migrant stock at mid-year by origin and by country of destination, 2015

3. Table: The characteristics of the examined networks

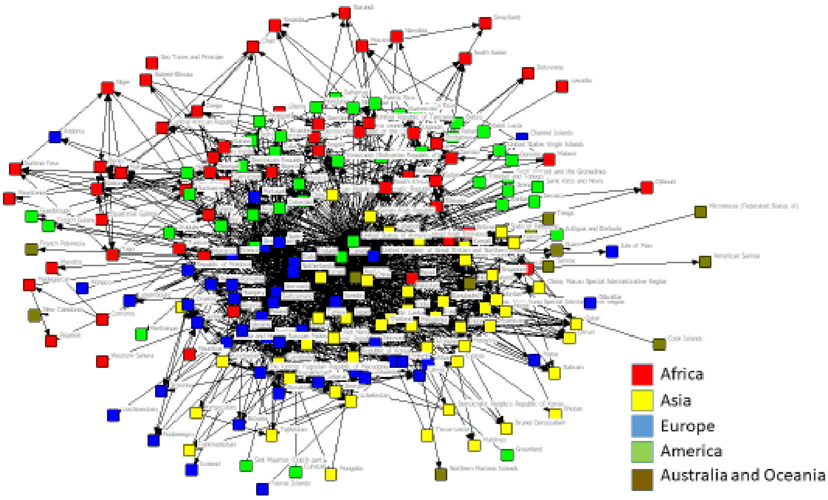
Description	Companies’ sales	International migration
Average geodesic distances within the network	3.8	1.9
Std Dev	1.1	0.6
Density (matrix average)	0.0093	0.2054
Std Dev	0.0962	0.4040

The length of a path in a network is the number of links it contains. The geodesic distance between two nodes is the length of the shortest path. The *distances in these networks are relatively small.* The density of a network is the total number of existing ties divided by the total number of possible ties. The density of the examined networks are increasing over time, showing that connectivity is strengthening. In these networks the nodes have only a few links. These numerous small nodes are held together by a few highly-connected hubs.

4. Figure: Companies' sales network in Hungary by NACE level1[7]



5. Figure: International migration network[8]

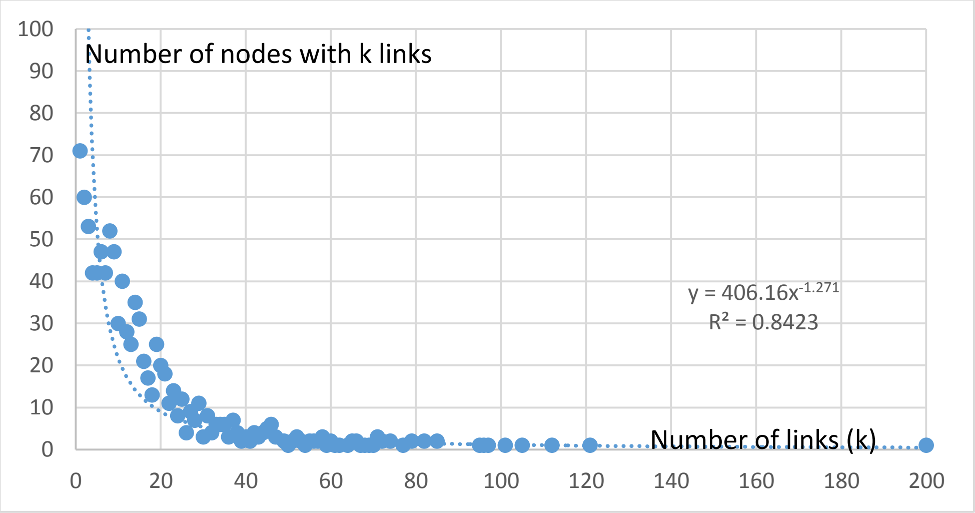


[7] Due to visibility, only trades above EUR 10 million are shown.

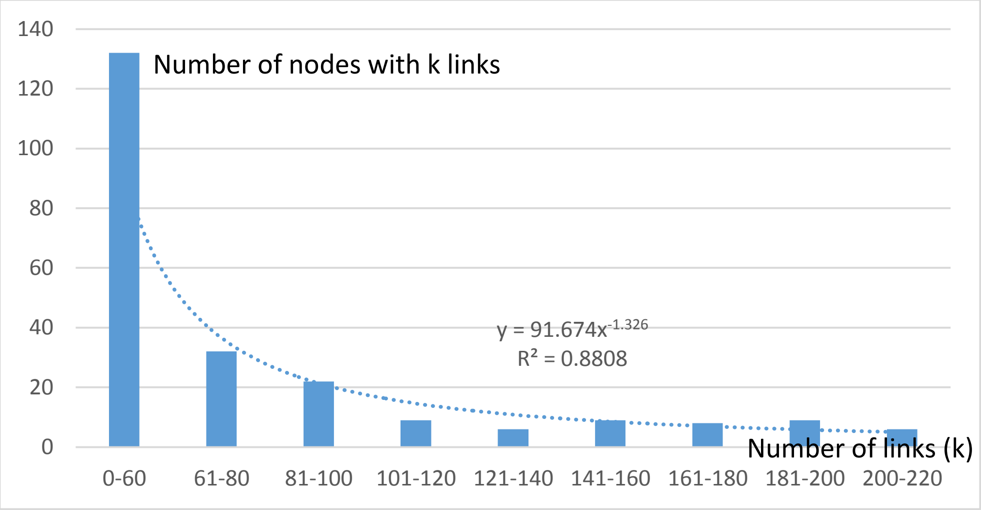
A - Agriculture, forestry and fishing; B - Mining and quarrying; C - Manufacturing; D - Electricity, gas, steam and air conditioning supply; E - Water supply; sewerage; waste management and remediation activities; F - Construction; G - Wholesale and retail trade; repair of motor vehicles and motorcycles; H- Transporting and storage; I - Accommodation and food service activities; J - Information and communication; K - Financial and insurance activities; L - Real estate activities; M - Professional, scientific and technical activities; N - Administrative and support service activities; O - Public administration and defence; compulsory social security; P - Education; Q - Human health and social work activities; R - Arts, entertainment and recreation; S - Other services activities

[8] Due to visibility, only migration above 10.000 migrants are shown

6. Figure: Degree distribution of company network

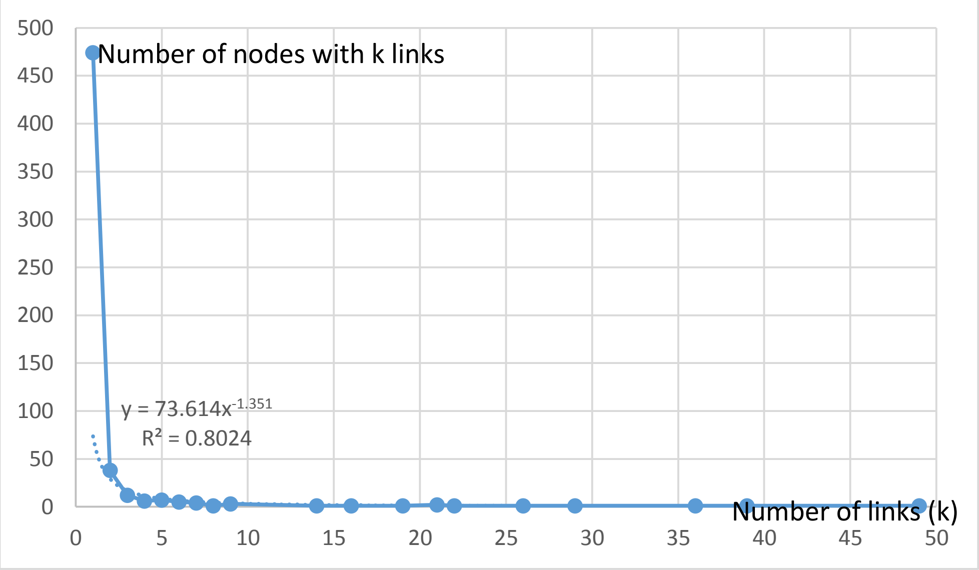


7. Figure: Degree distribution of migration network



The networks follow the Power-law distribution. The above facts highlight that the examined networks of official statistics are scale-free nature. In these cases we also may assume, that the reason is the growth and preferential attachment.

8. Figure: Degree distribution of management network



The presence of hubs and small world phenomenon are valid in the analysed statistics. This means that we can use the consequences and advantages of the *universal* characteristics of scale-free models.

In the globalised world, various activities (business, migration, etc.) arrange into networks with scale-free topology, and through these skeletons we can observe with official statistics the different phenomena that take place.

The complex systems and their collective behaviour cannot be recognized soundly just from the knowledge of the system's components. The networks do not stop at the borders of the countries, nor can they be effectively examined at country level: they require collaboration at EU and global levels. The global perspective is crucial to gain understanding of the full picture.

Networks with power law distribution do not have a meaningful internal scale. Observed units (the data providers) are not equally relevant. There are few vital and many trivial nodes. The hubs and the small world phenomenon allow us to better understand the processes of globalisation beyond the figures, and improve relevance and quality of official statistics. Hence:

- The presence of Large cases units (LCU) and the European networks of LCU are essential. These units should focus on global networks, hubs, key enterprises, MNEs, core phenomena and the global supply chain.

- The exchange of microdata is important in general, but the largest enterprises, their activity and connectivity should be the centre of interest.

- Official statistics need to reorganize their data collection work (few vital and many trivial units):

- It is necessary to rethink sample selection methods;
- It is important to foster differentiated checking aspects: more connected nodes are more important than others; we need to prioritize them in our control system;
- These actions allow us to reduce the data providers' burden parallel to allocating more efforts to hubs.

4. Conclusions

Network theory is an innovative tool that reflects a new type of thinking in our changing world, which can help us handle the challenges of the 21st century.

The scale-free nature of networks has played an important role in the development of networks as a whole, as can be seen in many scientific networks and practical interest networks. This scale-free property an unavoidable issue in many disciplines. Once the hubs are present, they fundamentally change a system's behaviour. The statistics of the 21st century have had scale-free features. This means that in the globalised world, various activities (business, migration, etc.) fall into networks with scale-free topology, and through these skeletons we can observe with official statistics the different phenomena that take place.

We should move forward from the traditional thinking and traditional distributions. The meaning of average has gradually lost its importance, there are no averagely-sized companies (just tiny or arbitrarily large). If we want to increase the quality and relevance of statistics, we should focus on the hubs and networks behind the numbers.

Networks do not stop at the borders of the countries, nor can they be examined effectively at country level: they require collaboration at EU level. Given the important roles that complex systems play in our daily lives and in our economy, understanding and eventually controlling them is one of the major intellectual and scientific challenges of the 21st century. It is a challenge that European statistics cannot afford not to tackle. We see many potential to develop official statistics in line with network analysis and further research may need to be conducted at the European level.

5. Acknowledgements

This paper was supported by the János Bólyai Scholarship of the Hungarian Academy of Science.

6. References

[1] Albert-László Barabási (2009): Scale-Free Networks: A Decade and Beyond; Science, Vol 325, pp.412-413.

[2] Albert-László Barabási (2010): Bursts: The Hidden Pattern Behind Everything We Do; Button, New York; p. 310

[3] Albert-László Barabási (2016): Network Science, Cambridge University Press, p.453

[4] Battiston, Federico; Nicosia, Vincenzo; Latora, Vito (2017): The new challenges of multiplex networks: Measures and models. The European Physical Journal Special Topics. 226 (3): 401–416. doi:10.1140/epjst/e2016-60274-8. ISSN 1951-6355.

[5] Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). Ucinet 6 for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

[6] Cardillo, A.; et al. (2013): Emergence of network features from multiplexity. Scientific Reports. 3:1344.

[7] D. Shah and T. Zaman (2011): Rumours in a Network: Who's the Culprit? IEEE Trans. Inform. Theory 57:pp. 5163-5187

[8] Bombieri Enrico (1992): Problems of the Millennium: the Riemann Hypothesis, Institute for Advanced Study, Princeton, NJ 08540: https://www.claymath.org/sites/default/files/official_problem_description.pdf

[9] Kryven, Ivan (2016-07-27): Emergence of the giant weak component in directed random graphs with arbitrary degree distributions". Physical Review E. 94 (1): 012315. doi:10.1103/PhysRevE.94.012315.

[10] Lawyer, Glenn (March 2015): Understanding the spreading power of all nodes in a network. Scientific Reports. 5 (O8665):8665.

[11] Lundy Lewis (2001) Managing Business and Service Networks. p. 138

[12] Newman, M.E.J. (2010): Networks: An Introduction. Oxford University Press. ISBN 978-0199206650

[13] T. W. Vante (1995): Network models of the diffusion of innovations. Hampton Press, Cresskill, NJ; p.18