Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

A hybrid model for the patent citation network structure

Konstantinos Angelou [a,b], Michael Maragakis [a,b,c], Kosmas Kosmidis [a,b], Panos Argyrakis [a,b,*]

[a] *Department of Physics, University of Thessaloniki, Greece*
[b] *Center of Complex Systems, University of Thessaloniki, Greece*
[c] *Department of Physics, International Hellenic University, Kavala, Greece*

ARTICLE INFO

Article history:
Received 29 March 2019
Received in revised form 18 October 2019
Available online 4 November 2019

Keywords:
Patent citation networks
Percolation
Preferential attachment
Hybrid networks

ABSTRACT

Percolation theory on the patent citation network is studied and the percolation threshold points are identified. The results show that there is a significant change of the threshold throughout our dataset years, implying changes in the formation process of the patent citation network. There is a first shift at around 2001, and a very delayed transition point after 2008. Giant component formation in such networks is an indication of the existence of inter-disciplinary patents. In order to explain the changes observed, a hybrid model for creating networks is suggested here. The model is based on a combination of random networks and preferential attachment. It is also compared with results from the well-known configuration model. The hybrid model fits better the data of the patent citation network, rather than a single scale-free or a single Erdős–Rényi network, and explains the increase in preferential attachment in later years. Both the degree distribution and the results of the analysis through percolation theory agree well with real data. This enables the formation of a plausible explanation for the structural changes of the patent citation network's evolution.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Percolation is a prototype model that reveals a phase transition. Broadbent and Hammersley first introduced the term in order to model the flow of fluid in a porous medium with randomly blocked channels [1]. Its simplest application is on a two dimensional square lattice. Some of the lattice sites are open and some are closed, which means that they cannot be accessed. We start with a lattice of no open sites and increase their concentration progressively, by adding open sites randomly. When a path of open sites from one end of the lattice to the other is formed for the first time, we say that the percolating giant component has just been emerged, and that we have reached the percolation threshold.

Percolation has been studied intensely [2–5], as it has application on numerous real-world systems, in a wide variety of scientific fields. For example, it is applied in chemistry [6], electromagnetism [7,8], geology [9] and many more. In addition to lattice system the idea of percolation can be used in networks. Here if an extended cluster of network sites exists, then this corresponds to a system above its percolation threshold, whereas if only small isolated network clusters are present, then this corresponds to the system being below its percolation threshold. Thus, percolation has been extensively used on networks to study their resilience against the random removal of nodes [10,11] and the spread of diseases [12,13]. In addition, it has been applied to social systems to estimate a community's lifetime [14], and to collaborative networks to identify how scientists are connected with other scientists of their field [15–17].

[*] Corresponding author at: Department of Physics, University of Thessaloniki, Greece.
E-mail address: panos@auth.gr (P. Argyrakis).

https://doi.org/10.1016/j.physa.2019.123363
0378-4371/© 2019 Elsevier B.V. All rights reserved.

In the current study we apply percolation on a patent citation network. Networks of patents and patent citations have been attracting interest for a long period of time [18–20] as they prove to be useful for the identification of the highly cited patents that reflect on the importance of technological needs of times.

The main goal is to study the minimum number of citations that are required for the formation of the giant component, the way this network evolves during a span of 38 years of data, and to propose a plausible mechanism for this evolution. In addition, we introduce a simulation model to replicate the equivalent behaviour of the actual patent citation network, and explain differences that occur in its formation throughout these years.

The data used, and the main methodology followed, are described in further detail in Section 2, Data and Methodology. The Results and Discussion derived from the percolation analysis are given in Section 3. Conclusions of this work are summarized in Section 4.

2. Data and methodology

2.1. Data

The data used for the current analysis are provided by the Organization for Economic Co-Operation and Development (OECD). They contain patents published from 1978 onwards, and up to 2016, along with their citations to other patents (prior art). These are recorded in the European Patent Office (EPO) and the Patent Co-operation Treaty (PCT). These two databases have been merged, the duplicates have been removed and a unified database has been created, resulting to a network whose nodes are the patents, and the citations are the directed links between them. The total number of patents in the network is 12,126,928 and the number of outgoing citations is 21,873,933. It should be noted that, since the primary data source is a European based organization, these data may have an under-representation of Asian and North American patents, while having an over-representation of European ones.

2.2. Methodology

In this study, we use our time resolved data in order to construct a collection of patent citation networks. We study the percolation properties of these networks and the way these properties evolve with time. Next, we compare our patent citation networks to simulated networks generated using the standard configuration model [21] with which networks can be created using a specific degree sequence. The sequence then results to a specific degree distribution for the given exponent, γ.

We observe that when we use simulated scale-free networks with any value of the γ exponent that we do not get good agreement with the real data, as far as mean node degree is concerned. We, thus, propose to use a hybrid method of creating networks by combining the methods of Erdős–Rényi [22] for random networks and the preferential attachment Barabási–Albert [23] model. This hybrid mechanism offers qualitative similarities with the real network and, thus, a plausible explanation for the formation of the patent citation network. Preferential attachment means that highly cited patents are cited by newer patents, thus, creating a credible scenario where the importance of the field is demonstrated.

More specifically, the algorithm for creating such hybrid networks is the following:

- Create an Erdős–Rényi network with the desired conditions (number of nodes, mean degree value).
- Initiate a new, hybrid, network with a small number of nodes (required for the Barabási–Albert method to work), taken randomly from the previous network.
- Insert new links in the network, depending on the probability of Erdős–Rényi as follows:
  - Create a random number in the range [0, 1).
  - If the random number is smaller than the Erdős–Rényi probability, insert an Erdős–Rényi link chosen randomly from the first constructed network into the new network, otherwise create a new link with the Barabási–Albert method.
  - Repeat the last two steps until the desirable number of links and nodes has been added.

The result of this methodology is the creation of networks with degree distribution $P(k) \sim k^{-s}$ with various slopes s, depending on the randomness inserted in the network, namely, the percentage of the links that are created by the Erdős–Rényi methodology.

In order to better understand the formation process we consider the size of the second largest cluster (SLC) [24,25]. This constitutes a measure to identify the point where a phase transition (percolation) takes place in a system. The point where the SLC reaches its maximum value is considered to be the percolation threshold. The analysis is performed by inserting links into an empty network and monitoring the size of the SLC, until it reaches its maximum value followed by a sharp drop.

Some examples of percolation in networks using this method are shown in Fig. 1. Simulated scale-free networks of size 500,000 nodes and $\gamma = 3$ have been created by the configuration model and percolation method is applied upon them. Due to the random creation of these networks we observe varying sizes and shapes of the SLC, as shown in Fig. 1. One can observe sharp peaks (Fig. 1a), peaks with noise (Fig. 1b) or multiple peaks (Fig. 1c), like the ones acquired in individual 2D lattice runs. In order to acquire smoother curves with reduced noise, the average of 50 iterations (by creating many random networks) is taken, Fig. 1d.

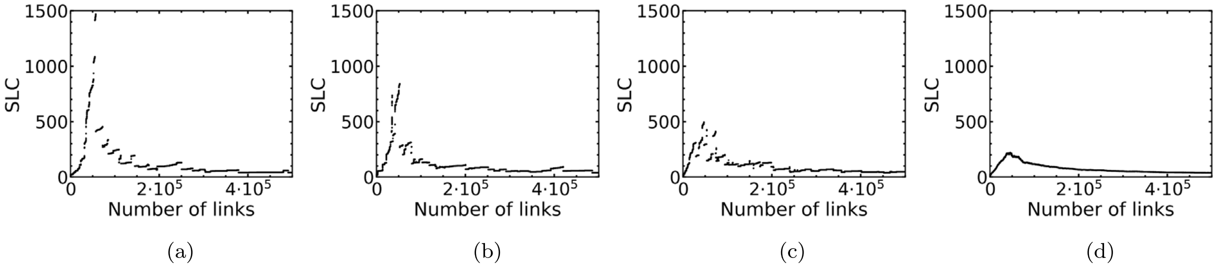

Fig. 1. Plot of the size of the SLC versus the number of links. The networks have been created using the configuration model. Their size is 500,000 and $\gamma = 3$. (a)–(c) plots of individual networks are indicative cases of types of behaviour that can be found. (d) Corresponds to the average of 50 such cases. More specifically (a) is a sharp peak of the SLC, (b) is a peak with noise, (c) are multiple peaks and.

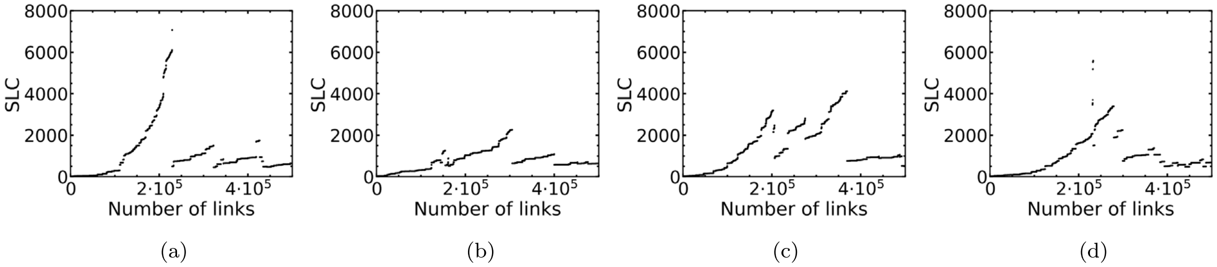

Fig. 2. Plot of the size of the SLC versus the number of links, that occur when percolation is applied to random starting dates of the patent citation network. (a) A sharp peak of the SLC indicates a clear transition point. (b) There is no sharp peak (c)(d) Unclear cases where two or more peaks of the SLC occur.

3. Results and discussion

The data show that the registration of patents takes place 1 to 3 days per week. On each such day, a different amount of patents, along with their outgoing citations, is recorded in the Patent Offices (in the current study patents are recorded in the European Patent Office and the Patent Cooperation Treaty). For the period of years that can be found in the database (1978–2016) the total number of these days is 3300.

For our analysis, each one of these days is considered to be the starting point of a sub-network, while citations and patents prior to that day are omitted. Specifically, 3300 different new sub-networks are formed, each one consisting of $\approx$ 500,000 outgoing citations (at which point we stop following the evolution of the sub-network). Networks are grown by adding citations one at a time, and examined by calculating the size of the second largest cluster at that point.

The results derived with this approach showed three different cases. The first case is where sharp peaks of the SLC exist and indicate the exact time (the specific citation) of the system's phase transition and the formation of the giant component (Fig. 2a). After that point (the specific citation inserted into the network which causes the sharp drop in the SLC), the giant component keeps growing at a much higher rate than the rest of the clusters. Its size gets extremely larger than that of the second, in size, cluster.

There are often cases where there is no sharp peak or any significant change of the SLC (Fig. 2b). They indicate that there is either no giant component, or, more likely, that the giant component is formed at the very early stages of evolution, when most of the patents are attached to it directly.

Finally, there are other cases where multiple peaks or peaks with noise are observed, resulting in unclear estimations of the exact time (exact citation) of the giant component's formation (Figs. 2c, 2d). This is, however, observed in the simulated networks too.

Next, the moving average process was applied to smooth out fluctuations and highlight longer-term trends [26] on the 3300 aforementioned days. The first window begins from day one and includes the results of the first 500 days. The second window begins from day two and it includes the following 500 days, and so on.

The results showed that the noise is indeed reduced and the data are smoothed out. By applying this method we get only one peak indicating a single phase transition or a transition from one peak to another. Specifically, for data concerning the period 1978–2000, it takes $\approx$ 200,000 citations for the formation of the giant component (Fig. 3a). During 2000–2001 a peak displacement of the SLC takes place to higher values (Fig. 3b), and after that year $\approx$ 300,000 citations are required for percolation (Fig. 3c). Finally, after 2008 a peak displacement of the SLC is also noticed to $\approx$ 500,000 links. At this point the SLC is only 2.5% (on average) of the largest cluster and, thus, we consider that the giant component is effectively formed at the very early formation stages of the network (Fig. 3d). The same results were derived by using as a metric

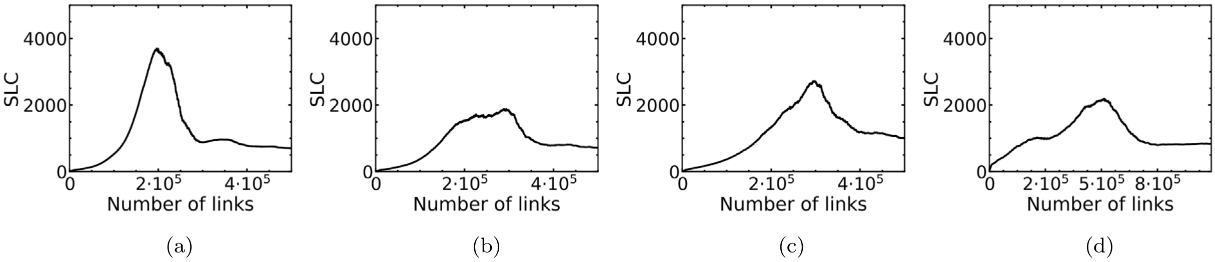

Fig. 3. Plot of the SLC versus the number of links, that occur from averaging all the various cases of patent citation network percolation. (a) From 1978 and up until 2001, ≈ 200,000 links are required for the formation of the giant component, where the SLC is on average 25% of the largest cluster. (b) The period around 2001 when the SLC peak changes from 200,000 links to 300,000. (c) After 2001, ≈ 300,000 links are required for the formation of the giant component, where the SLC is on average 12.5% of the largest cluster. (d) At around 2008, a displacement of the peak is noticed, similar to that of (b), to ≈ 500,000 links, after which the SLC is on average 2.5% of the largest cluster. Please note that the x-axis is different in figure d.

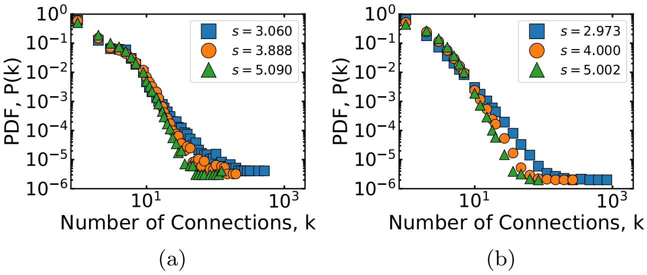

Fig. 4. Degree distributions for $3 \leq s \leq 5$ for (a) the patent citation network and (b) simulated networks created using the method in Section 2.2. The patent citation network has been split into networks of ≈ 500,000 patents and links, and for each network the degree distribution has been found. Respectively, simulated networks have been created with various values of s depending on the analogy of Erdős–Rényi and Barabási–Albert links. It should be noted that each point corresponds to a different analogy of the two types of networks. $s = 2.973$ corresponds to 1% Erdős–Rényi and 99% Barabási–Albert, $s = 4.000$ to 25% Erdős–Rényi and 75% Barabási–Albert and $s = 5.002$ to 40% Erdős–Rényi and 60% Barabási–Albert.

the reduced average cluster size ($I'_{average}$), which is commonly used in lattice percolation [27]. Both the thresholds and the Figs. 2, 3, 5 were found to be similar to the ones of the second largest cluster.

It is evident that something has changed throughout these years in the formation and structure of the patent citation network. The changes seen around 2001 can perhaps be attributed to the increase of inter-disciplinarity. This would explain the delay in the formation of the giant component, in the sense that many previously loosely related patent fields are now linked to begin with, while several others cannot coalesce with the largest one until much later on. A second and larger change is observed at around 2008. One plausible explanation could be the emergence of many new technological sub-fields. We have noticed that the number of citations per sub-class is an increasing function of time (results not shown). Several patents in recent years cite a much larger number of different technological fields and this number shows an increase in 2001 and a much larger one in 2008. After that year, clusters coalesce almost at the beginning of the network formation, suggesting that inter-disciplinary fields are now well-established.

Indeed, the size of the largest cluster becomes many times larger when compared to the second largest in a very short amount of time, and the second largest remains relatively small even after 500,000 links have been added.

We further studied the degree distributions throughout the entire data duration, to better understand the obtained results. More specifically, we have created a moving window of size equal to 500,000 patents, and the degree distribution for each window was calculated. We find that at the early years the slope of the degree distribution is $s \approx 5$, and decreases with time until it reaches $s \approx 3$ at later years. In order to explain the variation in the degree distributions we wanted to find a way to create networks with various values of s that could also explain why this happens in the patent citation network, as the configuration model could not.

The method described in 2.2 was used to create networks with ≈ 500,000 nodes and links (size of simulated networks was purposefully kept similar to that of the patent citation networks), for $\langle k \rangle = 2$. The results showed that the more Erdős–Rényi links that exist in the network, the larger is the slope of the degree distribution, s (Fig. 4b) .

Finally, to further examine the resemblance of simulated networks to patent citation networks, a percolation study, similar to the one above, has been performed. The patent citation network has been divided into 44 sub-networks of size 500,000. Percolation methodology has been applied to each one of these networks and the exact citation resulting in phase transition has been found. At the point where this critical link appears, the fraction of links per nodes is calculated and plotted versus the slope, s, that occurs from the degree distribution of the network.

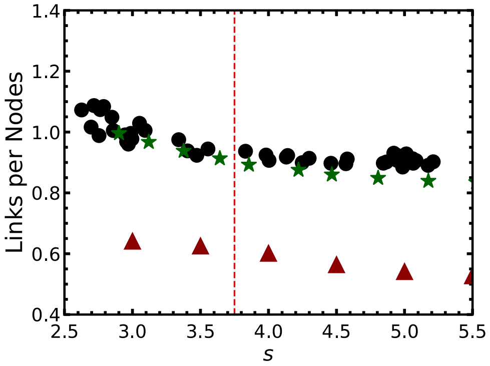

Fig. 5. Plot of the average number of links over nodes versus degree distribution slope, s, on the percolation threshold. Dots represent the patent citation network, triangles the configuration model and stars the hybrid model. To the right side of the dashed line are patent citation networks formed prior to 2008, and to the left side are after 2008. The dashed line coincides with a simulated network having 80% preferential attachment. To the left side each star has 5% more preferential links per star in the simulated network, to the right 5% fewer.

Similarly, we studied the simulated networks mentioned above (both the configuration and the hybrid model). The resulting 44 patent citation networks and the simulated networks are shown in Fig. 5. It is obvious that the three curves have a very similar behaviour. However, the hybrid model has a behaviour that is much closer to the patent citation network than the configuration model. This provides a plausible explanation for the behaviour of the patent citation network. This similarity supports the claim that the simulated networks created by the hybrid model fit much better the data of the actual patent citation network.

By examining the dates these networks were formed, we find that all networks to the right side of the dashed line were formed prior to 2008, see Fig. 5. This means that all networks prior to 2008 resemble a simulated network containing a higher number of Erdős–Rényi links, than those after 2008 (located in the left side of the line).

A plausible explanation might be that at the very beginning of the patent citation data (1978), the internet and the patent databases are not easily accessible from all the potential inventors. As a result, it is reasonable to assume that there is great difficulty to properly cite all the related and influencing patents. This task could only be partially done by either large corporations that have the means, or patent reviewers that were knowledgeable and added citations during the review. Thus, it is likely that the network resembles a mostly random one and shows smaller amount of preferential attachment (hence the values for $s \approx 4$–5). To add to that, that era is also an era where hard sciences had well defined boundaries between them.

With advancing time (later than 2008), the communication means and databases evolved, which may have resulted to an increase of well targeted (preferential) citations. Thus, the need for randomness decreases and the model that fits quite well the behaviour of the real data is that of almost pure preferential attachment. This is also indicated by the dashed line, which means that more than 80% of the links in the corresponding simulated network are chosen with the Barabási–Albert method.

4. Conclusions

Summarizing, the main focus of the current study is the existence of a giant component, and the change in the percolation threshold point where the component forms, on patent citation networks. The results showed that there are periods in time where we obtain a clear phase transition with distinct points. At the same time, there are other periods where the transition point changes. More specifically, from 1978 and up until 2001, $\approx$ 200,000 citations were required for a giant component to form throughout the entire period. This means that the mechanism behind the structure of the patent citation network was more or less the same for almost 23 years. During the next two years there is a noticeable displacement of the number of patent citations required for percolation to $\approx$ 300,000. This, again, remains a good approximation for about 5 years, up to year 2008. Following 2008, a displacement also takes place to $\approx$ 500,000 patent citations.

These results show that there is a significant change in the formation process of the patent citation network throughout the time period examined. At recent times, and ever since around 2008, and possibly due to the increase in inter-disciplinary research being performed worldwide, almost all patent citations are connected at the very creation of the network. Indeed, the size of the SLC is significantly smaller compared to the largest cluster ($\approx$ 2.5%) when they join. Inter-disciplinarity also explains why patents inserted in the network have higher probability to be attached to the largest cluster. Of course, checking the actual plausibility of this speculation is a matter for future work. It can be tested by performing community detection at different times and investigating the resulting community dynamics.

To this purpose, a model that provides a plausible explanation for the patent citation network behaviour is now suggested here. Specifically, a hybrid network that derives from Erdős–Rényi and Barabási–Albert methodologies has been presented. This model can relate to changes found in the patent citation network.

At around 1978 when the first patents are recorded in our data, the internet and the patent databases are not easily accessible from the inventors. Thus, a reasonable assumption would be that the related patents might not have been properly cited. As a result, the network's behaviour resembles that of an Erdős–Rényi random graph and less that of one with preferential attachment.

As time progresses, technology is advancing and means of communication and databases are evolving, which may have led to more citations to prior art. Thus, the Erdős–Rényi probability decreases (randomness) and the preferential attachment model is now dominant.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge the support provided by the IT Center of the Aristotle University of Thessaloniki (AUTh), Greece throughout the progress of this research work.

References

[1] S.R. Broadbent, J.M. Hammersley, Percolation processes, I. Crystals and mazes, Proc. Cambr. Philos. Soc. 53 (1957) 629–641, http://dx.doi.org/10.1017/S0305004100032680.

[2] V.K. Shante, S. Kirkpatrick, An introduction to percolation theory, Adv. Phys. 20 (85) (1971) 325–357, http://dx.doi.org/10.1080/00018737100101261.

[3] D. Stauffer, Scaling theory of percolation clusters, Phys. Rep. 54 (1) (1979) 1–74, http://dx.doi.org/10.1016/0370-1573(79)90060-7.

[4] D. Stauffer, A. Aharony, Introduction to Percolation Theory, Taylor and Francis, 1994.

[5] M. Newman, The structure and function of complex networks, SIAM Rev. 45 (2) (2003) 167–256, http://dx.doi.org/10.1137/S003614450342480.

[6] R.D. Groot, T.J. Madden, Dynamic simulation of diblock copolymer microphase separation, J. Chem. Phys. 108 (20) (1998) 8713–8724, http://dx.doi.org/10.1063/1.476300.

[7] L. Bergqvist, O. Eriksson, J. Kudrnovský, V. Drchal, P. Korzhavyi, I. Turek, Magnetic percolation in diluted magnetic semiconductors, Phys. Rev. Lett. 93 (2004) 137202, http://dx.doi.org/10.1103/PhysRevLett.93.137202.

[8] L. Hu, D.S. Hecht, G. Grüner, Percolation in transparent and conducting carbon nanotube networks, Nano Lett. 4 (12) (2004) 2513–2517, http://dx.doi.org/10.1021/nl048435y.

[9] D. McKenzie, R.K. O'Nions, Partial melt distributions from inversion of rare earth element concentrations, J. Petrol. 32 (5) (1991) 1021–1091, http://dx.doi.org/10.1093/petrology/32.5.1021.

[10] R. Cohen, K. Erez, D. ben Avraham, S. Havlin, Breakdown of the internet under intentional attack, Phys. Rev. Lett. 86 (2001) 3682–3685, http://dx.doi.org/10.1103/PhysRevLett.86.3682.

[11] R. Cohen, K. Erez, D. ben Avraham, S. Havlin, Resilience of the internet to random breakdowns, Phys. Rev. Lett. 85 (2000) 4626–4628, http://dx.doi.org/10.1103/PhysRevLett.85.4626.

[12] M.E.J. Newman, Spread of epidemic disease on networks, Phys. Rev. E 66 (2002) 016128, http://dx.doi.org/10.1103/PhysRevE.66.016128.

[13] L.K. Gallos, F. Liljeros, P. Argyrakis, A. Bunde, S. Havlin, Improving immunization strategies, Phys. Rev. E 75 (2007) 045104, http://dx.doi.org/10.1103/PhysRevE.75.045104.

[14] G. Palla, A.-L. Barabasi, T. Vicsek, Quantifying social group evolution, Nature 446 (2007) 664–667, http://dx.doi.org/10.1038/nature05670.

[15] M. Newman, Scientific collaboration networks i. network construction and fundamental results, Phys. Rev. E 64 (2001) 016131, http://dx.doi.org/10.1103/PhysRevE.64.016131.

[16] A. Garas, P. Argyrakis, A network approach for the scientific collaboration in the european framework programs, Europhys. Lett. 84 (6) (2008) 68005.

[17] M. Tsouchnika, P. Argyrakis, Network of participants in european research: accepted versus rejected proposals, Eur. Phys. J. B 87 (12) (2014) 292, http://dx.doi.org/10.1140/epjb/e2014-50450-4.

[18] P. Criscuolo, B. Verspagen, Does it matter where patent citations come from? inventor vs. examiner citations in european patents, Res. Policy 37 (10) (2008) 1892–1908, http://dx.doi.org/10.1016/j.respol.2008.07.011, special Section Knowledge Dynamics out of Balance: Knowledge Biased, Skewed and Unmatched.

[19] L. Wu, D. Wang, J.A. Evans, Large teams develop and small teams disrupt science and technology, Nature 566 (7744) (2019) 378–382, http://dx.doi.org/10.1038/s41586-019-0941-9.

[20] K. Angelou, M. Maragakis, P. Argyrakis, A structural analysis of the patent citation network by the k-shell decomposition method, Physica A 521 (2019) 476–483, http://dx.doi.org/10.1016/j.physa.2019.01.063.

[21] M. Newman, Networks: An Introduction, Oxford University Press, Inc., New York, NY, USA, 2010.

[22] P. Erdős, A. Rényi, On random graphs i., Publ. Math. (Debrecen) 6 (1959) 290–297.

[23] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512, http://dx.doi.org/10.1126/science.286.5439.509.

[24] A. Bunde, S. Havlin, Fractals and Disordered Systems, second ed., Springer, 1996, http://dx.doi.org/10.1007/978-3-642-84868-1.

[25] A. Margolina, H. Herrmann, D. Stauffer, Size of largest and second largest cluster in random percolation, Phys. Lett. A 93 (2) (1982) 73–75, http://dx.doi.org/10.1016/0375-9601(82)90219-5, URL https://www.sciencedirect.com/science/article/abs/pii/0375960182902195.

[26] Y.-L. Chou, Statistical Analysis: With Business and Economic Applications, Holt, Rinehart and Winston, 1969.

[27] J. Hoshen, R. Kopelman, Percolation and cluster distribution. i. cluster multiple labeling technique and critical concentration algorithm, Phys. Rev. B 14 (1976) 3438–3445, http://dx.doi.org/10.1103/PhysRevB.14.3438.