EMERGING PROPERTIES OF THE CITATIONS NETWORK

Adrià Hernandez Morell
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Dimitri Marinelli
(Dated: June 14, 2024)

Abstract: This essay presents an analysis of the American Physical Society (APS) dataset, comprising metadata and citations of APS articles. We focus on network structures behind the dataset. We build three different networks: citation network, author citation networks, and co-author network. We compute key network metrics such as degree distribution, clustering coefficients, and centrality measures. We expect the degree distribution of the different networks to follow a power-law distribution. Through simulations, we compare the citation network structure to randomly generated directed graphs built by the configuration model with the same degree distribution, obtaining substantially different structural configurations. The in-degree distribution of the citation network exhibited scale-free properties, compatible with the preferential attachment network development. This study provides a comprehensive understanding of the APS citation network and emphasizes the importance of temporal and preferential attachment mechanisms in shaping real-world networks.

I. INTRODUCTION

In recent years, the study of complex networks has become a crucial field to understanding systems in various scientific fields. Networks can represent a multitude of systems and it's relations, from social interactions to biological systems, and eventually, scientific collaborations and citations [1].

Citation networks, in particular, are directed graphs where nodes represent scientific papers and directed edges represent citations from one paper to another. These networks are invaluable for analyzing the dissemination of knowledge, the influence of researches, and the evolution of scientific fields [2].

The American Physical Society (APS) is one of the foremost organizations in the field of physics, publishing a vast array of research papers across its numerous journals. By analyzing this network, we can uncover the structural characteristics of scientific collaborations, the impact of individual papers, and the overarching trends within the field of physics.

A key concept in the study of complex networks is the power-law distribution, which is often observed in the degree distributions of these networks. In a power-law network, a small number of nodes (papers) have a very high degree (many citations), while most nodes have a low degree (few citations). Preferential attachment (PA) is a mechanism that allow us to obtain a power law distribution from a simple rule where new nodes are more likely to connect to highly connected nodes. Therefore, PA mechanism plays a significant role in the development of scale-free networks. In the context of citation networks, this means that new papers are more likely to cite well-known, highly cited papers, leading to a "rich-get-richer" phenomenon.

Furthermore, time it's a critical factor in the evolution of citation networks, a property that makes these networks unique. The temporal aspect influences on how citations accumulate and how the network grows over time. Older papers have had more time to accumulate citations, leading to temporal biases in the network structure.

For this study, we utilize a unique dataset from the APS, specifically the dataset of citations since 1893 up to 2022. This dataset provides a comprehensive view of the citation dynamics within the APS. It is composed by 720.535 files corresponding to 2.76 GB of data.

The primary objective of this study is to investigate the statistical properties and structural characteristics of the APS citation network. Key metrics such as degree distribution, clustering coefficients, and centrality measures will be studied to provide a comprehensive understanding of the network. Degree distribution reveals the spread of citations among papers, clustering coefficients indicate the tendency of papers to form tightly-knit groups, and centrality measures identify influential papers within the network.

II. DATASET AND NETWORKS

The APS citations dataset consists of a CSV file that relates the DOI of citing papers to the DOI of cited papers, this will be used as our edge list for the graphs creation since it forms the basis for constructing the citation network, allowing us to identify the directional relationships between papers based on their citations. Additionally, we utilize a metadata dataset that provides detailed information about each paper. Key fields in these JSON files include the paper's title, publication date, journal information, and authorship details.

Using these datasets, and with the help of libraries such as NetworkX and graphtools, we decided to model the system through three different networks represented by three different graphs. In Table I we can see the configuration of these graphs, detailing the number of nodes and edges in each network:

In first place we have the *citation network (CN)*, where each node represents a paper and each directed edge represents a citation between two papers. Given that the citations can be considered with a direction, this graph is a directed graph (DG).

Graph	# Nodes	# Edges
Citations Network (CN)	709.782	9.832.517
Author Citations Network (ACN)	504.282	74.775.017
Coauthors Social Network (COAN)	500.083	3.578.387

TABLE I: Amount of nodes and edges related to each one of the networks. It's relevant to see that the ACN it's a much more dense graph than the CN and the COAN. Also the number of authors in the COAN is less than the number of authors in the ACN, this difference can be associated to the number of authors that have never collaborated since they won't be part of the COAN.

The second network is the *author citations network (ACN)*. In this graph, nodes represent the authors of the papers, and directed edges represent the citations between authors. If an author cites another author multiple times, the edge will be weighted according to the number of citations, therefore consolidating multiple parallel edges into a single weighted one.

The last network is the *coauthors social network (COAN)*, which focuses on the collaborative relationships between authors. Here, nodes represent authors, and undirected edges represent the existence of a paper co-authored by two authors. Unlike the authors' citation network, the coauthors network highlights collaboration rather than citation relationships, providing insights into the co-authorship patterns within the APS community.

III. NETWORK COMPONENTS

We look at the size of the components of the graph. A Connected Component (CC) is defined as the maximal subgraph in which any two nodes are connected by paths. For directed graphs, the paths have a direction, and two nodes must have two paths to be connected. This is called strongly connected component (SCC). If the directionality is ignored, we have Weakly connected component (WCC). Note that a node is always connected with itself.

During the initial analysis of the citation network's structure, we computed the number of SCC within the graph. Due to temporal coherence, it should be impossible for a paper to cite another paper that cites it. On this basis, when computing the SCC of the citation network, we expect no cycles or, at most, only small exceptions [11], which means that mainly we should find SCC composed by just one node. However, the citation network exhibited a SCC comprising 67.876 nodes, which constitutes approximately 10% of all the nodes.

Upon closer examination of this large SCC, several inconsistencies were discovered in the database (see Appendix A for further explanation). The most critical error identified was that some papers were erroneously citing other papers that were published much later in the future (see FIG. 1).

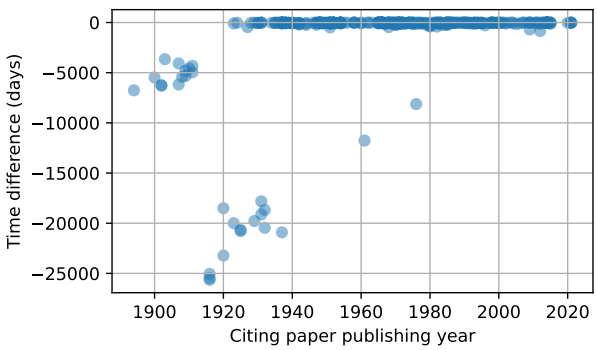

FIG. 1: Representation of the erratic citations in the dataset. The figure shows the year of publication of the paper vs. the number of days passed to the publication of the cited paper.

To address these inconsistencies, we implemented two corrective measures. In first place the removal of self-citations, all instances where papers cited themselves were removed from the database. In second place, we used the time gap as means to remove incorrect citations. In fact, we can see in FIG. 1 that these inconsistencies are majorly found in old papers, suggesting an OCR problem. Given this, we decided to neglect all the edges pointing to papers that were published more than 1.000 days later.

Altogether this accounts for only 29 references. Although the number of edges neglected is small, the SCC comprising the 10% of the nodes disappears. The tables for SCC and WCC are presented in the Appendix B.

After the correction, the CN comprises the approximated acyclicity expected. The results are coherent with the expected properties of a citation network. The absence of a large SCC aligns with the temporal constraints of citation practices. Instead, we observe many single-node and some small SCC, likely representing groups of papers published simultaneously and citing each other. The largest WCC encompasses almost all citations, highlighting the extensive connectivity of the APS citation network. Given the absence of big SCC and the uniqueness of a huge WCC indicates that the CN mainly has the form of a unique global tree [3].

On the contrary, the ACN presents a big SCC and therefore an even bigger WCC. In the COAN case, which is not directed and only presents connected components (CC), we have the presence of a large principal CC, altogether with the ACN indicating a vast collaborative hub within the scientific community. This extensive interconnectedness illustrates the openness and collaborative spirit prevalent among physicists. Smaller WCC/CC likely represent isolated research groups or specialized subfields with limited interaction outside their niche.

IV. POWER-LAW DISTRIBUTION AND SCALE FREE NETWORKS

Upon analyzing complex networks often is indicated that many real-world networks exhibit a power-law degree distribution. This characteristic can be indicative of a scale-free network, where a few nodes have a very high

degree, while most nodes have a low degree. The term "scale-free" is rooted in a branch of statistical physics called the theory of phase transitions [4]. The degree distribution $p(k)$ of a scale-free network follows the form:

$$p(k) = C \cdot k^{-\alpha} \tag{1}$$

where α is the exponent of the power law, and C is a normalization constant. A network is considered scale-free if $2 \leq \alpha \leq 3$. This range of the exponent has significant implications for the network's properties. To understand this, we can look at the moments of the degree distribution [5]. The mean degree for a power-law distribution $\langle k \rangle$ is given by:

$$\langle k \rangle = \sum_k k \cdot p(k) = C \sum_{k=k_{\min}}^{\infty} k^{1-\alpha} \tag{2}$$

where $k_{\min}$ is the minimum degree. Given this, on one hand for $\alpha \geq 2$, this series converges, and the mean degree is finite. On the other hand, the variance is given by $\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$ where

$$\langle k^2 \rangle = \sum_{k=k_{\min}}^{\infty} k^2 \cdot p(k) = C \sum_{k=k_{\min}}^{\infty} k^{2-\alpha}, \tag{3}$$

this makes the series diverge when $\alpha \leq 3$ since the term $k^{2-\alpha}$ decreases slowly enough. Thus, $\langle k^2 \rangle$ becomes infinite, leading to an infinite variance.

An infinite variance implies that the distribution has a heavy tail, characteristic of scale-free networks. This heavy tail means that while most nodes have a low degree, there are a few nodes with very high degrees, forming hubs. The mechanism of PA, is an example that explains the emergence of scale-free networks. In a PA model, new nodes are more likely to connect to already highly connected nodes. This "rich-get-richer" mechanism leads to the formation of hubs, or nodes with a very high degree, which are characteristic of scale-free networks [4].

V. FITTING OF A POWER LAW DISTRIBUTION

Given the power-law distribution observed in the degree distributions of the three networks, we need to rigorously test whether these distributions can indeed be classified as scale-free networks. To achieve this, we must fit a power-law distribution to the empirical data. Conventional methods such as least-squares fitting are often inadequate for this purpose as they can produce inaccurate parameter estimates and fail to confirm whether the data truly follow a power-law distribution [6].

A. Estimation of the Scaling Parameter

To estimate the scaling parameter α, we need to identify the lower bound $k_{\min}$ of the power-law behavior in the data. Initially, we assume this value is the minimum degree, but it will be determined more accurately from the data later. The method for fitting parameterized models like power-law distributions is the method of maximum likelihood estimator (MLE).

a. MLE Assuming that the data follows a power law for $k \geq k_{\min}$, we can find α with the MLE, an approximation for the result is given by [6]:

$$\alpha \approx 1 + N \left(\sum_{i=1}^{N} \ln \frac{k_i}{k_{\min} - \frac{1}{2}} \right)^{-1} \qquad \sigma_\alpha = \frac{\alpha - 1}{\sqrt{N}}, \tag{4}$$

where N is the number of nodes with a degree greater than or equal to $k_{\min}$, and the sum is over all nodes with $k \geq k_{\min}$, and σ_α is the statistical error on α. The normalization constant in (1) is computed as $C = 1/\zeta(\alpha, k_{\min})$ where ζ is the Riemann zeta function.

b. Estimation of k_{min} To refine the value of $k_{\min}$, we proceed as follows [6]:

1. Choose an initial range for $k_{\min}$ from the minimal degree to a given $k_{\max}$.
2. For each $k_{\min}$ in this range, estimate the degree exponent α using 4.
3. With the obtained $(\alpha, k_{\min})$ pair, assume the degree distribution has the given power-law form and so the associated cumulative distribution function (CDF) is:

$$P(k) = 1 - \frac{\zeta(\gamma, k)}{\zeta(\gamma, k_{\min})} \tag{5}$$

4. Use the Kolmogorov-Smirnov test to determine the maximum distance D between the CDF of the data $S(k)$ and the fitted model provided by the CDF, with the selected $(\gamma, k_{\min})$ parameter pair:

$$D = \max_{k \geq k_{\min}} |S(k) - P(k)| \tag{6}$$

5. Finally, repeat steps 1-4 by scanning the whole $k_{\min}$ range. We end up obtaining the $k_{\min}^{opt}$ value for which D is minimal.

c. Goodness of Fit To validate the power-law model, we employ a goodness-of-fit test using synthetic data [4]. This involves:

1. Using the CDF to estimate the KS distance between the empirical data and the best fit, denoted as D_{real}.
2. Generating synthetic degree sequences consistent with the estimated degree distribution and calculating the corresponding KS distances, $D_{\text{synthetic}}$.
3. Repeating step 2 multiple times to obtain the distribution $p(D_{\text{synthetic}})$ and comparing D_{real} against this distribution.

If D_{real} is within the distribution of $D_{\text{synthetic}}$, the power-law model is considered a reasonable fit for the data. When the value of p is close to 1, the difference between the empirical data and the model can be attributed to statistical fluctuations alone. By following this method, we ensure that the degree distributions of the CN, ACN, and COAN are thoroughly tested for power-law distribution, and in the case of having $2 \leq \alpha \leq 3$ determining their scale-free behaviour.

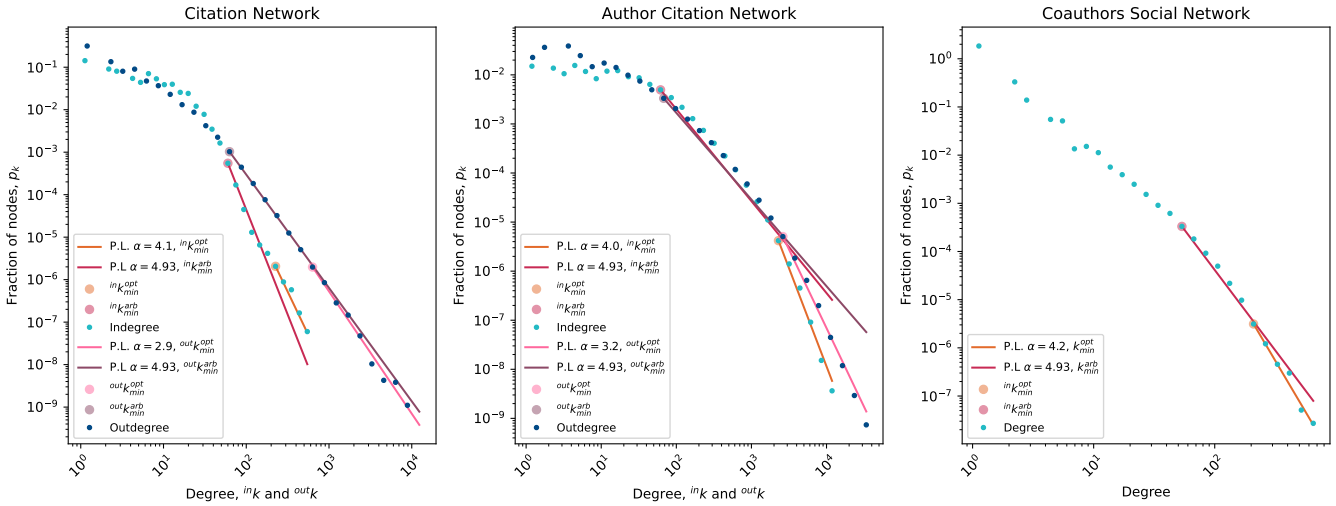

FIG. 2: Representation of the power law (P.L.) fittings done for each of the networks in our model and their respective degree distributions. The k_{min}^{opt} is the optimum value for the minimum degree to make the P.L. fitting whereas k_{min}^{arb} is chosen arbitrarily, both can be for the indegree or the outdegree. The α value is the scaling parameter.

B. Results

The detailed values of the different parameters used in the fitting process are summarized in a table included in the Appendix C, along with the goodness-of-fit p-value for each distribution.

Both the in-degree and out-degree [12] distributions for the CN and ACN follow a power law distribution. However, only the in-degree distribution for the CN exhibits the characteristics of a scale-free network [7], as indicated by the value of the power-law exponent α. The degree distribution for the COAN was also fitted to a power law successfully but, as the ACN, it does not exhibit scale-free properties. Notice that this distributions might not be entirely following a power law but a similar, even though not the same, distribution as it could be the log-normal [2].

VI. CLUSTERING MEASURES

Apart from analyzing the degree distribution of the graphs, additional measures such as the number of triangles and transitivity have been computed to better understand the structure of these networks. These measures provide insights into the clustering tendencies and local interconnectedness of the nodes. A triangle is a set of three nodes that are all connected to each other. The presence of triangles in a network indicates a tendency for nodes to form tightly-knit groups or clusters. Transitivity, also known as the global clustering coefficient, is a measure of the overall tendency of nodes to cluster together. It is calculated as the ratio of the number of triangles in the graph to the number of possible triangles:

$$T = 3 \cdot \frac{\#\,Triangles}{\#\,Triads} \tag{7}$$

where triad is the identifier for a possible triangle, this can be computed as the amount of pair edges that share a same vertex. Therefore, transitivity can be perceived as a normalized measure of clustering that is independent of the size of the network.

Graph	Triangles	Transitivity
CN	22.931.859	0.05
ACN	3.688.142.228	0.10
COAN	13.264.171	0.15

TABLE II: Table with the amount of triangles and transitivity for all the graphs representing the three different networks. In the case of the ACN, the computation for the triangles and transitivity has been done neglecting the repeated citations in-between two authors and the self citations.

The number of triangles in the CN is substantial, but this value alone may not be representative due to the differences in the number of nodes and edges between the different graphs. The transitivity for the CN is relatively small, indicating that it is uncommon for two cited papers to be cited between them. Both the the ACN and the COAN, have a higher number of triangles and transitivity compared to the CN. This higher transitivity in both implies a greater tendency for authors to form collaborative clusters.

VII. SIMULATION OF THE CITATION NETWORK

To investigate whether the structural properties of the citation network can be easily replicated, we performed a simulation by generating a random directed graph based on the number of nodes and the degree sequence of the citation network.

A. Configuration Model Algorithm

The configuration model algorithm is a well-established method for generating random graphs with

a specified degree sequence [8, 9]. The algorithm works as follows:

1. A prerequisite is a given degree sequence, which specifies the number of edges (in-degree and out-degree) for each node.
2. Create a list of "stubs" (half-edges), where each node contributes a number of stubs equal to its degree.
3. Randomly pair the stubs to form edges, ensuring that each pair of stubs connects two nodes to form a directed edge and assemble these edges into a directed graph.

This process preserves the degree sequence of the original network while randomizing the connections between nodes.

B. Results of the simulation

We used the configuration model algorithm together with the degree sequence from the CN. Once generated the random directed graph, we can compare its structural properties with those of the actual CN, focusing on the number of triangles, transitivity, and connected components.

The number of triangles obtained 270.259 in the randomly generated graph is significantly lower than in the CN. This indicates that the CN has a higher tendency for local clustering that can't be accounted for by the degree sequence alone.

The transitivity of the randomly generated graph is $6 \cdot 10^{-4}$, again much lower than the one obtained for the CN. This suggests that the citation network has a more pronounced clustering structure, likely influenced by factors such as the temporal evolution of citations and the PA mechanism.

The amount of SCC in the random graph differed markedly from the CN. The random graph exhibited a huge SCC, making it a big cycle, therefore having no time coherence in the citations network.

These big differences underscore the fact that the degree sequence, while an important factor, does not fully capture the structural complexity of the citation network, the use of more complex algorithms might overcome this difficulties[10].

VIII. CONCLUSIONS

In this study, we analyzed the structural properties of the American Physical Society (APS) citation network to understand its statistical characteristics and emergent behaviors. Using the APS citation network database, we calculated key network metrics such as degree distribution, clustering coefficients, and centrality measures. Our initial analysis revealed significant SCC, which were identified as systematic errors such as self-citations and anachronistic citations. The investigation confirmed that the degree distributions of the citation network (CN), the authors citation network (ACN), and the coauthors network (COAN) follow a power-law distribution. However, only the in-degree distribution of the CN exhibited the characteristics of a scale-free network, suggesting the significant role of PA in the citation network. Furthermore, additional metrics such as the number of triangles and transitivity provided deeper insights into the clustering tendencies of these networks. The results indicated lower transitivity in the CN compared to the ACN and COAN, reflecting less clustering in the citation network. A simulation using the configuration model algorithm highlighted the unique properties of the CN that cannot be replicated by degree sequence alone, emphasizing the importance of incorporating temporal dynamics and PA mechanisms. Future research could explore more sophisticated models to further replicate and understand the unique structural characteristics of citation networks.

[1] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, *et al.*, "Science of science," *Science*, vol. 359, no. 6379, p. eaao0185, 2018.

[2] F. Radicchi, S. Fortunato, and A. Vespignani, "Citation networks," *Models of science dynamics: Encounters between complexity theory and information sciences*, pp. 233–257, 2011.

[3] D. B. West *et al.*, *Introduction to graph theory*, vol. 2. Prentice hall Upper Saddle River, 2001.

[4] A.-L. Barabási, "Network science," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1987, p. 20120375, 2013.

[5] M. Newman, *Networks*. Oxford university press, 2018.

[6] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[7] A. D. Broido and A. Clauset, "Scale-free networks are rare," *Nature communications*, vol. 10, no. 1, p. 1017, 2019.

[8] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random structures & algorithms*, vol. 6, no. 2-3, pp. 161–180, 1995.

[9] B. Bollobás, "A probabilistic proof of an asymptotic formula for the number of labelled regular graphs," *European Journal of Combinatorics*, vol. 1, no. 4, pp. 311–316, 1980.

[10] Z.-X. Wu and P. Holme, "Modeling scientific-citation patterns and other triangle-rich acyclic networks," *Physical review E*, vol. 80, no. 3, p. 037101, 2009.

[11] For example, when a given journal publishes more than one paper at the same time, it might lead to mutual citations.

[12] The in-degree of a node in a citation network represents the number of citations a paper receives from other papers. Conversely, the out-degree represents the number of citations a paper makes to other papers.

Appendix A: Inconsistencies in the database

Firstly, there were instances of papers citing themselves, which is logically inconsistent. Additionally, it was observed that some journals publish their papers in batches when they are related, and as they cite each other this lead to bidirectional citations that should not exist. Furthermore, there are citations between a published paper and another one that is about to be published in the near future.

Appendix B: SCC and WCC

1. Citations Network

Size of the SCC	Amount of components
1	700.937
2	4.041
3	212
4	22
5	4
6	2
7	1

TABLE III: Amount of SCC given it's size for the citation network graph.

Size of the WCC	Amount of components
708702	1
19	2
18	1
13	1
12	2
8	1
7	3
6	4
5	9
4	22
3	75
2	288

TABLE IV: Amount of WCC given it's size for the citation network graph.

2. Author Citations Network

Size of the SCC	Amount of components
1	68260
2	172
3	54
4	17
5	4
6	4
7	2
9	1
13	2
14	1
435.357	1

TABLE V: Amount of SCC given it's size for the author citation network graph.

Size of the WCC	Amount of components
1	16
2	28
3	11
4	6
5	3
504.154	1

TABLE VI: Amount of WCC given it's size for the author citation network graph.

3. Coauthors Social Network

Size of the CC	Amount of components
476.479	1
29	1
22	1
21	2
20	1
19	1
18	1
17	3
16	3
15	4
14	8
13	7
12	11
11	22
10	29
9	43
8	75
7	141
6	226
5	412
4	897
3	1908
2	3863

TABLE VII: Amount of CC given it's size for the coauthors social network graph.

Appendix C: Values for the fittings

Graph		**Scaling parameter (α)**	**Normalization constant (C)**	**Minimum degree (k_{min})**	**Goodness of fit (p-value)**
Citations Network (CN)	Indegree	$4,1 \pm 0,2$	37.562.628,37	203	0,01
	Outdegree	$2,88 \pm 0,07$	214.619,62	492	0,74
Author Citations Network (ACN)	Indegree	$4,00 \pm 0,07$	26,93	2259	0,01
	Outdegree	$3,21 \pm 0,04$	15,47	2551	0,12
Coauthors Social Network (COAN)	Degree	$4,2 \pm 0,3$	22.334,10	193	0,02

TABLE VIII: Table with the data for each of the networks in our model and their respective distributions. It has the data for the scaling parameter α, the normalization constant C, the minimum degree k_{min}, and the goodness of fit p value.