Fractality in Empirical Social Networks

Pouria Mirelmi*
Chair for Computational Social Sciences and Humanities,
RWTH Aachen University, Aachen, Germany

(Dated: August 26, 2023)

As social networks emerge and grow in size, they go through a transition phase before they get too large and dense and become small-world. In this transition phase, they develop fractal characteristics until a certain peak, and then gradually lose fractality as it is in contradiction with small-world characteristics. The fractal state could be an optimal point between order and disorder, at which networks show interesting features. The problem with identifying this point in time is that the previous algorithms that can assess fractal property are too slow for large-scale networks. With the invention of the new Sketch-Based Box-Covering algorithm, this has been made possible to assess the fractality of large-scale networks. In this work, I apply the algorithm to the APS dataset of co-authorships in physics journals. I show that susceptibility of a network could be a good predictor for the fractal property, and discuss the patterns in sizes of the critical aggregation windows that form fractal networks.

I. INTRODUCTION

Empirical social networks might develop multiple different patterns as they emerge and grow in size.

In scale-free networks, vertex connectivities follow a power-law distribution as more and more vertices are added to the network, resulting in a few nodes -known as hubs- having significantly higher numbers of connections compared to the majority of nodes, which have fewer connections [1].

In small-world networks, as the network evolves, nodes exhibit large average clustering coefficients and short average path lengths at the same time, allowing the network to have tightly interconnected clusters and efficient communication between nodes despite their distant connections [2].

Fractal networks, on the other hand, exhibit similar structures over different length scales; as the network grows, new vertices and links maintain and develop patterns that are similar under different magnification levels [3, 4]. At first sight, it appears that the concept of fractality is in contradiction with the small-world characteristic; it is because in a fractal network different length scales do not even exist. Therefore, these two qualities cannot exist in a network at the same time. [5]. Csányi and Szendrői concluded about the fractal–small-world dichotomy in 2004: "The status of human social networks in this dichotomy is far from clear" [6]. Hardly any progress has been made until today, to address this matter. In few studies, the co-acting networks of IMDB have been proven to be scale-free [1], small-world [2], and fractal [7] at the same time. This is unexpected, according to the dichotomy described above.

Addressing the relation between fractality and small-world characteristics of social networks could be useful in that it can give us valuable information about the interactions in the societies they represent, and to better understand these interactions, aggregation of the edges is the key [8]. As an example, take the network of social coordination among scientists; too little aggregation in such a network does not allow for coordination, as there are not enough links to spread useful information in the network. On the other hand, too much aggregation means too much coordination. This may contradict innovation, in that it leads to fewer isolated components that can stay unaffected by the mainstream trends in their field [9].

Lee and coleagues have shown that in these coordination networks, when the edge aggregations are too little or too much, the networks are not fractal; there is a certain aggregation window in which networks are fractal [9]. This window is an optimal state between order (in this example: stability or sticking to tradition) and disorder (innovation or progressive change), in which networks also have other characteristics like increased robustness against intentional attacks [5] and less vulnerability when removing hubs from the system [4].

The research gap here is that this window has not been identified for social systems. The problem with identifying it is that fractality analysis requires to decide if a power-law or an exponential distribution better fits the data when applying a box-covering algorithm [3], and making this decision is only possible if the networks are large enough. Box-covering algorithms have been too slow when running on large networks, making it difficult for researchers to tackle the problem. Thanks to a new sketch-based box-covering algorithm which works in near-linear time with a guarantee of accuracy [10], we are now able to study the fractality-to-small-world transition and identify the time window which creates the optimal fractal state.

In this work, I apply the algorithm to the APS dataset of co-authorships in physics papers. Using different time windows each growing in size by 3 months, I indicate that

* pouria.mirelmi@rwth-aachen.de

the optimal fractal states appear around the point where the networks show the most susceptibility. Restarting each new window at the point where the susceptibility peaks, I also indicate that the fractal states (and peaks in susceptibilities) tend to appear at points that are closer and closer to the beginning of the windows.

II. DATA

The corpus of Physical Review Letters, Physical Review, and Reviews of Modern Physics consists of more than 450,000 articles and has origins tracing back to 1893. Based on this corpus, two datasets have been made available on the APS website. The first one is citing article pairs; it consists of pairs of APS articles that cite each other. As an example, when article A references article B, there will be a record in the dataset containing the pair of DOIs for A and B. The second dataset is article metadata; the dataset contains fundamental metadata of of all APS journal articles. In general, the APS data are very large and dynamic, and they are a good choice for constructing co-authorship networks for our analyses.

The data I used for creating the networks was initially available in the form of a single dataframe with 18.2 million rows (Special thanks to Dr. H. Lietz for preliminary pre-processings). The columns consist of the time in which the paper was published, the name of the journal, the doi of the paper, the id of the first author used as the node label in networks, the id of the other author, and an edge weight (1 divided by the number of all co-authors who collaborated in the paper). In fact, each row corresponds to a directed link from one author to another, at a certain time and in a certain paper. One significant pre-processing step in creating this data-frame has been to disambiguate the names in the data; this makes us able to distinguish between the authors who have the same names but are in fact different people [11].

To further pre-process the data and make it ready for the algorithm, I first removed parallel edges and self-loops, as it does not make sense for the box-covering algorithm to be run on networks with such edges. I also ignored the direction of the links, as the algorithm is meant for undirected graphs. Other pre-processing steps consist of removing outliers and extracting the largest connected components for each network. The outliers are removed based on the number of authors in the papers; I removed the papers in which the numbers of authors are more than 2 standard deviations larger than the means. Extracting the edges according to the time windows was of course another step to take before constructing the networks from the data. The final pre-processed data to be fed to the algorithm is in the form of an edge list of the desired network.

The analyses are done on the five major journals in the APS dataset: PRA, PRB, PRC, PRD and PRE. Physical Review A covers topics like atomic, molecular, and optical physics and quantum information. PRB covers

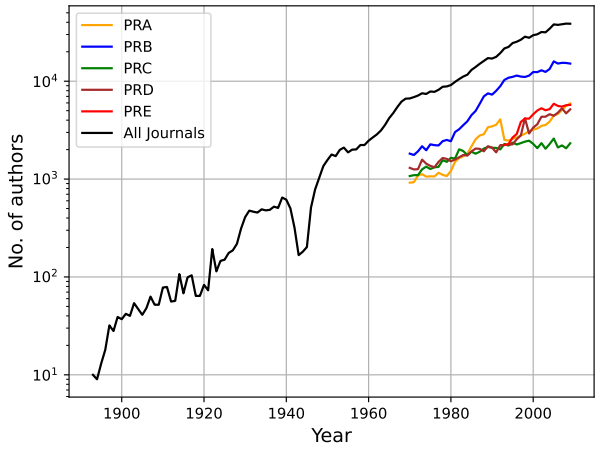

FIG. 1. The number of authors per year

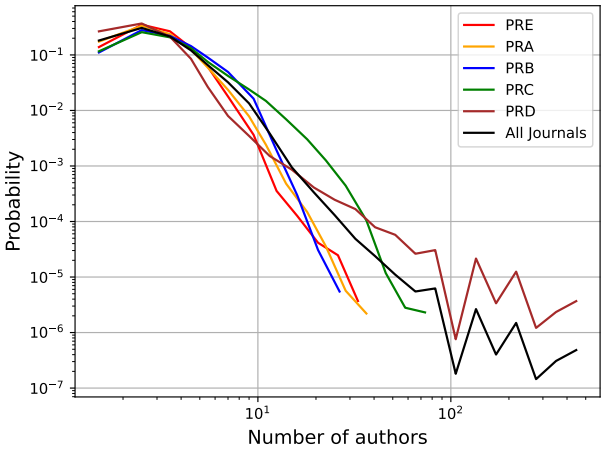

FIG. 2. The number of authors per paper

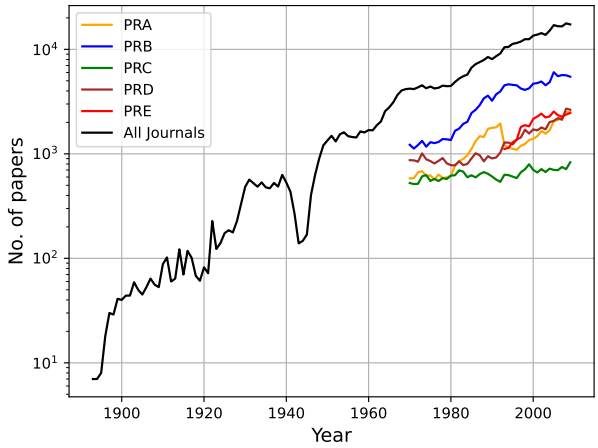

FIG. 3. The number of papers per year

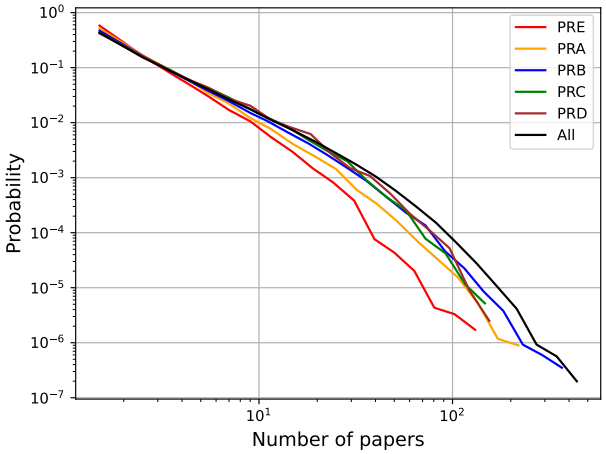

FIG. 4. The number of papers per author

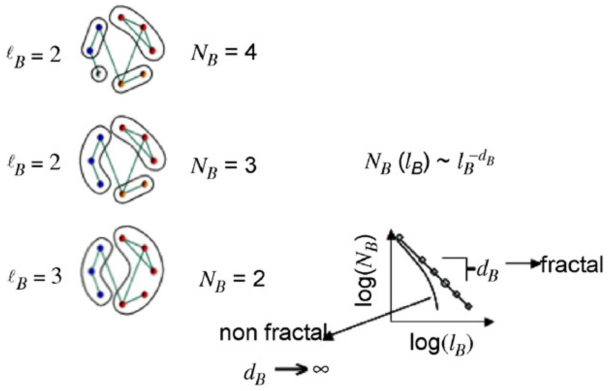

FIG. 5. The box covering algorithm applied to a network of eight nodes. The behavior of N_B as a function of l_B determines the presense of fractal characteristic. (Figure taken from Akiba et al. [10])

topics like condensed matter and materials physics. PRC is specifically aimed at nuclear physics, while PRD covers particles, fields, gravitation, and cosmology. The last one, Physical Review E, covers statistical, nonlinear, biological, and soft matter physics.

Some statistical infomation about these five journals together with all of the journals in the dataset could be found in figures 1 to 4. In figure 1, you can see the differences in the numbers of authors per year for the different journals. As shown in the figure, PRB has the biggest numbers of authors among all. You can also see that the papers in PRA to PRD begin from 1970, while the papers of PRE begin at 1993. The last papers in all of them are published in the end of 2009.

The number of authors per paper is shown in figure 2. PRD has the biggest teams of authors among all. While PRC is somewhere in the middle, PRA, PRB and PRE have the smallest teams.

The number of papers per year and number of papers per author are shown in figures 3 and 4 respectively. PRB has the most papers in each year, and authors of PRE have published the least papers among all.

III. METHODS

A. Box-covering a network

The process of recognizing fractal property within a complex network is similar to the approach employed for regular fractals (See figure 5) [5, 12]. The algorithm begins with covering the nodes with boxes, in a way that the distance l_{ij} between any two nodes i and j in a box is less than l_B, where l_B is the maximum box diameter. If we need N_B number of boxes to cover the whole network, then l_b and N_B scale as $N_B \sim {l_B}^{-d_B}$, where d_B is the fractal dimension of the network. This relation is valid only when N_B is the minimum possible for a given l_B [5, 12].

As we apply the algorithm to different networks, we find that there are two main groups of networks; the first one is the fractal networks group, where the power-law form of $N_B \sim {l_B}^{-d_B}$ is verified, and d_B has a finite value. The second group is the group of non-fractal networks, in which there is a sharp decay of N_B with l_B, which could be better described by an exponential form, or in fact an infinite fractal dimension d_B (See also figure 5) [5, 12].

B. Sketch-based box covering

The first step in older box-covering algorithms involves creating all boxes explicitly, followed by transforming the box cover problem into the famous set cover problem [10]. This method demands a substantial $\theta(n^2)$ space to represent neighbor sets, which becomes impractical for extensive networks with millions of vertices [10]. In contrast, the fundamental concept driving the sketch-based approach is to address the issue within the sketch space [10]. This signifies that neighbor sets are not generated explicitly; instead, the *bottom-k min-hash sketch representation* of boxes are formed and utilized [10, 13, 14].

Akiba and colleagues introduce a range of innovative concepts and algorithms. Initially, to make the approach based on sketches viable, a more relaxed problem termed as (1 - ϵ)-BOXCOVER is presented [10]. Additionally, a key subproblem known as the (1 - ϵ)-SETCOVER problem is defined [10]. The proposed box-cover algorithm consists of two segments; first, by generating min-hash sketches of all boxes, the (1 - ϵ)-BOXCOVER problem is reduced to the (1 - ϵ)-SETCOVER problem [10]. The sketch generation algorithm does not necessitate the direct instantiation of actual boxes and demonstrates effectiveness in both time and space aspects [10]. Second, the efficient sketch-space set-cover algorithm is applied to get the final result. The sketch-space set-cover algo-

rithm is based on a greedy approach, yet it is carefully designed with event-driven data structure operations ensuring nearly linear time complexity [10]. In theory, the scalability and solution quality of the proposed algorithm are guaranteed[10].

The authors have published the implementation of their algorithm -together with previous box-covering algorithms- at https://git.io/fractality. I wrote a python wrapper for the C++ code in order to employ this implementation in my work.

C. Aggregating co-authorships and network metrics

As previously mentioned in the introduction section, this work is typically based on analyzing networks made by edge aggregations of time windows each growing in size by 3 months. The corresponding network of the first time window consists of the co-authorship links from the papers that have been published in the first 3 months of the respective journal. The network of the second time window consists of the links from the papers in the first 6 months, and so on. I aggregate edges for 30 different time windows (making it 7.5 years of edge aggregation for the 30th time window), and visualize the behavior of the networks in terms of the metrics shown in table I.

The error ratio of a network indicates the extent to which the network shows fractal property; it is the ratio of the weighted mean squared error of the exponential fit to the weighted mean squared error of the power-law fit, after fitting the two functions to N_B and l_B values.

In each netwrok, the susceptibility per node is defined as $S = \Sigma_s s^2 n_s / N$, with n_s being the number of clusters that contain s nodes (The summation includes all clusters but the largest connected cluster) [15]. Therefore, S characterizes the variance of fluctuations per node in cluster size for the system, reaching its peak during the emergence of a giant graph component at the critical point [15]. According to my primary experiments, the peak in susceptibility might be a good predictor for the peak in error ratio, thus the peak in fractal property.

The ADAGE method of Soundarajan et al. requires a criterion for closing the windows [16]. I dicided that the susceptibility could be the best criterion for partitioning the data stream and building structurally mature graphs [16]. When observing a peak in susceptibility, I take a "snapshot" from the window up to that point [16], and continue by considering 30 new time windows re-starting right after the peak, looking for the next critical point where the susceptibility is maximized.

The complete pseudo-algorithm to identify the critical times is described in table II. I have considered a condition for collecting the error ratio results; if the network diameter is bigger than 3, then I fit the functions and calculate the error ratio. This leads to a degree of freedom of 2, making the results from fitting the power-law and exponential functions more reliable.

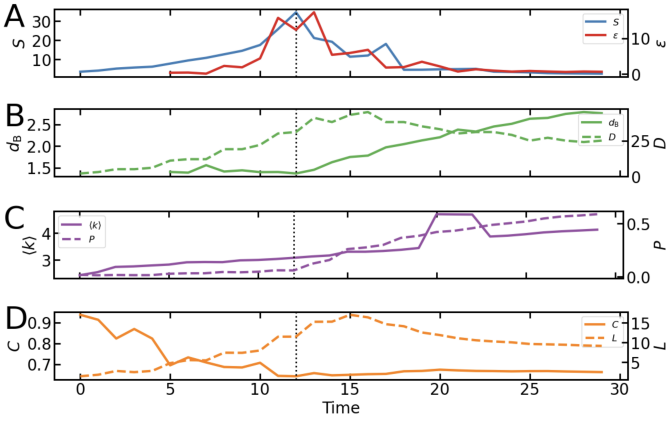

FIG. 6. First critical point in PRE

When the largest possible window size in the sanpshot-taking-loop becomes smaller than 3 years (when we almost reach the last papers published in the journal), the algorithm ends and will not take further sanpshots (see table II).

I have also used other metrics to describe the effect of variable window sizes on the networks. It is expected that the percolation probilities (the ratio of no. of nodes in LCC to the no. of nodes in the whole network) and average clustering coefficients rise at criticality, and keep going up until the end, as with more and more aggregations, the networks keep growing in size and becoming more dense.

The average shortest path lengths and network diamaters are expected to be large at the critical point, and as the network gets more dense over time, they are expected to fall.

The fractal dimension is expected to be smaller than or equal to 3 at criticality, so its variance would be infinite.

IV. RESULTS

A. Physical Review E

The detailed results from Physical Review E are presented in this section. The results from other journals are more briefly discussed in section B.

PRE has papers published from the beginning of 1993 to the end of 2009. By running the algorithm on the first 30 aggregations of the data, one can see that the first critiality in PRE appears at the window number 12 (See figure 6.A). This means that the network resulting from aggregating edges since the beginning of 1993 until the end of March 1996 is the network with highest susceptibility among all networks of the first 30 aggregations. As shown in figure 6.A, the peak in fractality appears just a window after that, at the 13th aggregation window. As expected, they both begin to fall afterwards.

Figure 6.B illustrates the changes in network diame-

TABLE I. Parameters to describe network snapshots

	Parameter	Description	Expectation at criticality	Expectation above criticality
P	Percolation probability [15]	Fraction of nodes in LCC[a]	Rise from zero	Large
S	Susceptibility [15]	$\sum_s s^2 n_s / N$[b]	Fall from large values	Small
d_B	Fractal dimension[c]	Inferred with linear model	≤ 3	–
ϵ	Error ratio[c]	$WMSE_{\text{Exp}}/WMSE_{\text{PL}}$	Maximum	Small
$\langle k \rangle$	Average degree	Average number of neighboring nodes	–	–
D	Network diameter[c]	Largest pairwise shortest path length	Large	Fall
C	Average clustering coefficient[c]	[2]	Rise	Large
L	Characteristic path length[c]	[2]	Large	Fall

[a] The LCC is the largest connected component in a snapshot.
[b] n_s is the number of components that contain s nodes, N is the number of nodes in the snapshot, and the sum extends over all components except the LCC.
[c] Computed for the LCC.

TABLE II. Pseudo-algorithm to identify critical time

```
date_begin = ...  # store journal begin date (e.g., 1993-01-01)
date_window = []
dates_end = [date_begin + 3months, date_begin + 6months, ..., date_begin + 90months] # prepare 30 increasing-
                                                                                      # ly large windows
while date_begin + 36months < 2009-12-31:  # while a window of 3 years is still possible
    for date_end in dates_end:  # for each window
        construct network for window [date_begin, date_end]
        compute for the whole network:  P, S, and <k>
        compute for the largest connected component:  D, C, and L
        apply box-covering algorithm to largest connected component to compute:  l_B and N_B
        if D > 3:
            fit power law and exponential to N_B = f(l_B) and compute:  d_B and error ratio
    date_end_max = ...  # date_end where S has global maximum
    date_begin = date_end_max
    append date_begin to date_window
```

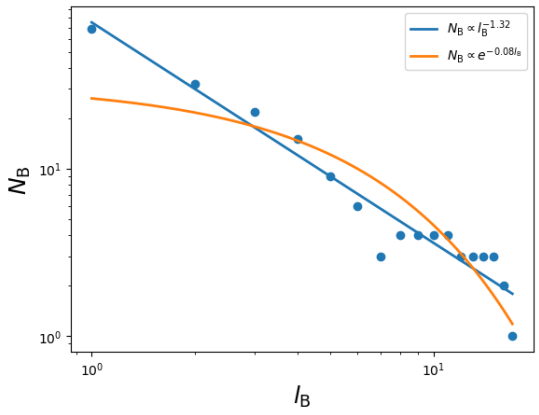

FIG. 7. PRE, aggregated from 1993-01 to 1996-03

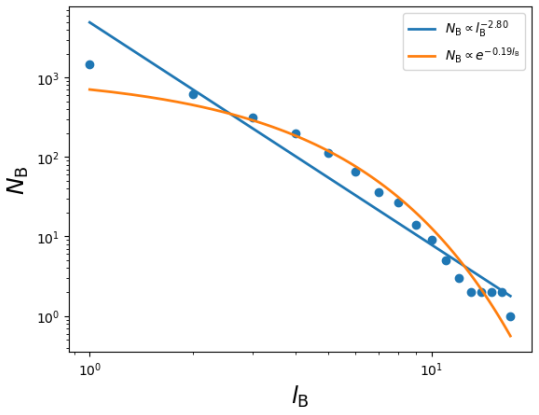

FIG. 8. PRE, aggregated from 1993-01 to 2000-06

ter and fractal dimension. The fractal dimension has its lowest value (around 1.4) in the critical point, and as the networks grow, it keeps ascending until the end. The network diameter also rises a little bit after the critical point, but as the networks grow more in size and become more dense, the network diameter eventually falls.

In 6.C, changes in average degree and percolation probability are illustrated. As expected, percolation probability rises after criticality, but it doesn't rise from zero, which is unexpected. Figure 6.D shows small world-characteristics. Average clustering coefficient slightly rises, which is expected. Average shortest path length

rises after criticality, which is unexpected, but it begins to fall after a short while and continues falling, identifying the small-world property in more-aggregated networks.

As previously mentioned in the methods section, the snapshot is taken at window 12 since the peak in susceptibility is observed, and aggregations are restarted from the timestamp after that, in order to take the next snapshot at the next peak. The fractal property is significant at the critical point in susceptibility, peaking just one window later (in window 13). In figure 7, you can see the function fittings at the critical point; at a highly fractal state, the values are better fitted to the power-law function compared to the exponential function. This is the point where the small-world characteristics are in direct contradiction to the fractal property (See also figure 6).

The function fittings of the 30th network (the network aggregated from the beginning of 1993 to the end of June 2000) are illustrated in figure 8. As you can see, the values of N_b as a function of l_B are almost perfectly fitted to the exponential function. This is an example of a poorly fractal network, in which the small-world characteristics are rather significant. As you can see in the plots in figure 6, the characteristic path length in the network is short, the average clustering coefficient is higher compared to the critical point, and the average degree and percolation probability are high. The fractal dimension is also higher than those of all the other 29 networks.

Based on the pseudo-algorithm in table II, we can take 8 more snapshots following the first one. In all of them, the peak in susceptibility appears only one or two windows before or after the peak in the error ratio; the only exception is the last snapshot, where the peak in error ratio appears 4 windows ahead of the critical point in susceptibily. Given the fact that the last snapshot has been taken analysing only 13 possible last windows in the data, we can consider its result as an outlier.

In the end, susceptibility works well as a predictor of fractality in PRE, and the median of the differences between the peaks in susceptibility and the peaks in error ratio is -1.

B. Physical Review A to D

Based on my experiments, the results obtained for Physical Review E are mostly verified for Physical Review A to C as well; the peak in susceptibility is a good predictor for the fractal property. This is not true about PRD, as there are significant structural differences in the networks made from PRD papers compared to the others. The medians of the differences between the peaks in susceptibility and the peaks in error ratios are -2, -1, -1 and -5 for PRA, PRB, PRC and PRD, respectively. The difference in results from PRD are mainly due to the differences in sizes of the author teams, as already mentioned in the data section (See figure 2). I will discuss that more in details in the discussion section.

Figure 9 illustrates the histograms of some of the met-

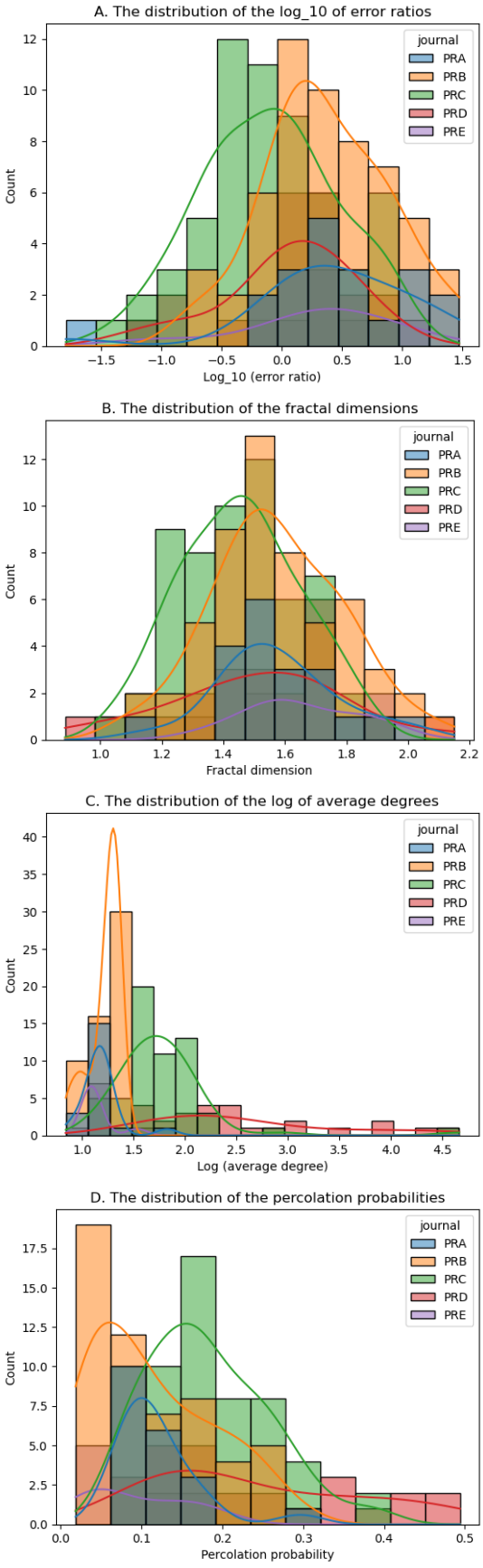

FIG. 9. Histograms of A. Error ratios, B. Fractal dimensions, C. Average degrees, D. Percolation probabilities in all snapshots.

rics of the 5 journals at the critical points of all snapshots. In 9.A, the distributions of $\log_{10}$ of the error ratios in all snapshots of the journals are illustrated. This indicates that, for example, PRC shows less fractality in general compared to the other 4 journals, as its error ratios are centered below zero, implyting that the networks are mostly non-fractal. The histogram also shows how the magnitude of the error ratios differ for the journals; PRE is in general the most fractal jounal according to that.

In 9.B, The distributions of the fractal dimensions at critical points are illustrated. As you can see, almost all of the fractal dimensions have values between 1 and 2. They are in fact centered around 1.5, which is expected by theory [17].

In 9.C, the log of average degrees of the networks at all critical points are shown. Unlike in random graphs, the average degrees in all of the journals are much larger than 1. PRD has the highest average degrees, and PRE, as the most fractal journal, has the lowest average degrees compared to others.

Figure 9.D shows the distributions of the percolation probabilities at critical points. In general, P is smallest for PRE and largest for PRD. This means that in PRD, a large share of aggregated links affect the fractality analysis, whereas in PRE much fewer links are taken into account when applyting the box-covering algorithm, as the algorithm is only applied to the largest connected components of the networks.

The number of snapshots taken in PRA to PRD are 20, 56, 59 and 26, respectively. This is the reason why the kernel density heights vary in all of the histograms in figure 9.

C. Decreasing patterns in the sizes of the critical aggregation windows

One insteresting pattern which exists in all of the journals except PRD is that as we take more and more snapshots from the data, smaller aggregation windows are needed to reach the critical point. In figure 10, The first 5 snapshots of PRE are used as an example to indicate that. As you can see, the vertical-dotted lines showing the critical points tend to move towards the beginning of the x-axis as further snapshots are taken.

In plots A and B in figure 11, you can see the sizes of the critical aggregation windows in all of the snapshots of PRE and PRB, respectively. As you can see in figure 11.C, PRD is an outlier here as well; the pattern cannot be seen in the changes of the critical window sizes.

V. DISCUSSION

Based on the extensive experiments, the critical point in susceptibility is proved to be a good predictor for the fractal property in the co-authorship networks of PRA, PRB, PRC and PRE. In PRD, susceptibility is not as reliable, as the networks in the journal appear to be different in various structural characteristics compared to the other journals. According to figure 2, the co-authorship teams in PRD are significantly larger than those of other journals; this leads to more dense netowrks as we aggregate the edges, and as the networks get more dense, we expect less fractal property in them.

As you can also see from figure 2, PRC is the second

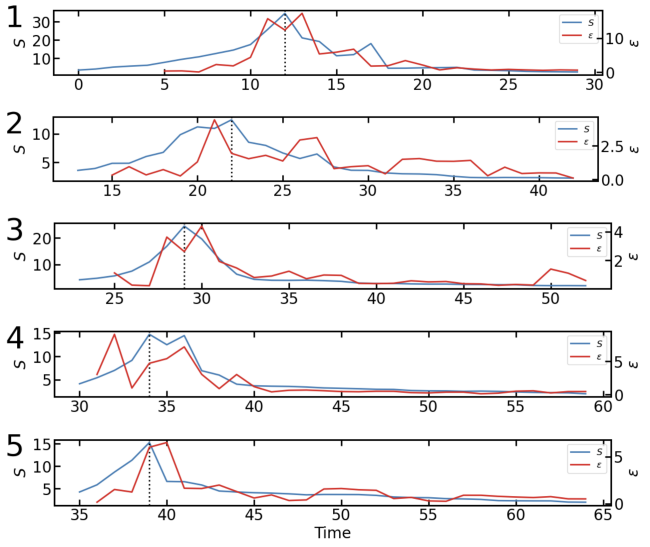

FIG. 10. Susceptibility and error ratio plots in the first 5 snapshots of PRE.

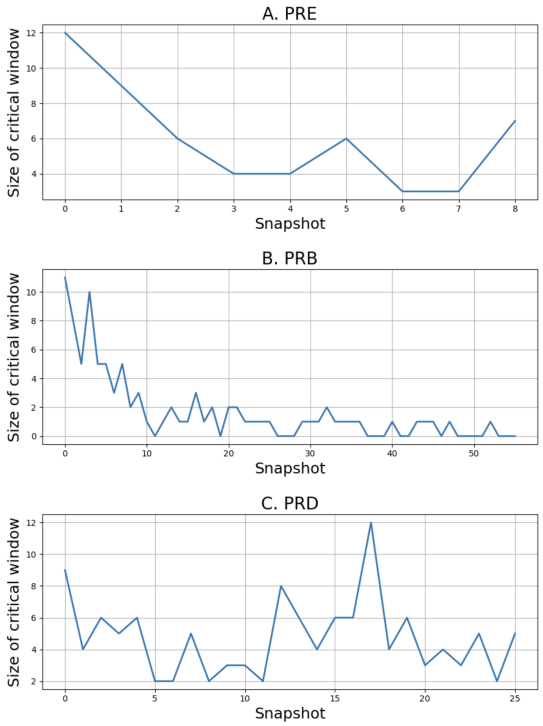

FIG. 11. Critical window sizes in PRE, PRB and PRD.

journal in terms of having large co-authorship teams. Although we expect more fractality in it compared to PRD, we can see in figure 9.A, that PRD is relatively more fractal in general. The difference in the sizes of teams can be more related to the goodness of susceptibility in predicting fractality; as already mentioned in section III.B, peaks in susceptibility work well in predicting the fractal properties of PRC, but they are not as reliable for PRD. This could be further investigated in future works. Extending the study to other datasets could also be interesting for future work.

VI. CODE IMPLEMENTATIONS

The code implementations together with the results could be found at the RWTH Aachen University gitlab: git.rwth-aachen.de/pouria.mirelmi/fractality.

ACKNOWLEDGMENTS

I would like to thank Dr. Haiko Lietz for extensive discussions on fractality and box-covering methods, as well as input on which papers to read. The expert insights greatly enriched my understanding and provided crucial direction for this research.

[1] A.-L. Barabási and R. Albert, Emergence of Scaling in Random Networks, Science **286**, 509 (1999).
[2] D. J. Watts and S. H. Strogatz, Collective dynamics of "small-world" networks, Nature **393**, 440 (1998).
[3] C. Song, S. Havlin, and H. A. Makse, Self-similarity of complex networks, Nature **433**, 392 (2005).
[4] C. Song, S. Havlin, and H. A. Makse, Origins of fractality in the growth of complex networks, Nature Physics **2**, 275 (2006).
[5] L. K. Gallos, C. Song, and H. A. Makse, A review of fractality and self-similarity in complex networks, Physica A: Statistical Mechanics and its Applications **386**, 686 (2007), disorder and Complexity.
[6] G. Csányi and B. Szendrői, Fractal–small-world dichotomy in real-world networks, Physical Review E **70**, 016122 (2004).
[7] H. D. Rozenfeld, C. Song, and H. A. Makse, Small-World to Fractal Transition in Complex Networks: A Renormalization Group Approach, Physical Review Letters **104**, 025701 (2010).
[8] L. K. Gallos, F. Q. Potiguar, J. S. Andrade, and H. A. Makse, Imdb network revisited: Unveiling fractal and modular properties from a typical small-world network, PLoS ONE **8** (2013).
[9] D. Lee, K.-I. Goh, B. Kahng, and D. Kim, Complete trails of coauthorship network evolution, Physical Review E **82**, 026112 (2010).
[10] T. Akiba, K. Nakamura, and T. Takaguchi, Fractality of Massive Graphs: Scalable Analysis with Sketch-Based Box-Covering Algorithm, in *2016 IEEE 16th International Conference on Data Mining (ICDM)* (IEEE, Barcelona, Spain, 2016) pp. 769–774.
[11] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási, Quantifying the evolution of individual scientific impact, Science **354**, aaf5239 (2016).
[12] C. Song, L. K. Gallos, S. Havlin, and H. A. Makse, How to calculate the fractal dimension of a complex network: the box covering algorithm, Journal of Statistical Mechanics: Theory and Experiment **2007**, P03006 (2007).
[13] E. Cohen, Size-estimation framework with applications to transitive closure and reachability, Journal of Computer and System Sciences **55**, 441 (1997).
[14] E. Cohen, All-distances sketches, revisited: Hip estimators for massive graphs analysis, IEEE Transactions on Knowledge and Data Engineering **27**, 2320 (2015).
[15] L. M. A. Bettencourt and D. I. Kaiser, Formation of Scientific Fields as a Universal Topological Transition (2015), published: SFI Working Paper 2015-03-009.
[16] S. Soundarajan, A. Tamersoy, E. B. Khalil, T. Eliassi-Rad, D. H. Chau, B. Gallagher, and K. Roundy, Generating Graph Snapshots from Streaming Edge Data, in *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion* (ACM Press, Montréal, Québec, Canada, 2016) pp. 109–110.
[17] H. D. Rozenfeld, S. Havlin, and D. ben Avraham, Fractal and transfractal recursive scale-free nets, New Journal of Physics **9**, 175 (2007).