CS224W: Social and Information Network Analysis
Reaction Paper

Adithya Rao, Gautam Kumar Parai, Sandeep Sripada

Keywords: Self-similar networks, fractality, scale invariance, modularity, Kronecker graphs.

1 Introduction

In this reaction paper, we investigate the property of self-similarity exhibited by various real-life networks, and the models that have been proposed to describe such networks. We begin with the paper [1], in which the authors analyzed metabolic networks in organisms. We next look at the self-similar structure seen in email networks and their comparison to river networks in [2]. In [3] this property of complex networks is studied further and a mechanism for the growth of such networks is proposed in [4]. Finally we look at more recent work in [5], [6] and [7] related to Kronecker Graphs, which have been shown to accurately model many properties of real-life networks with very few parameters. This reaction paper thus relates to the following concepts of networks studied in class: (a) Network connectivity (b) network diameter (c) network topology (d) clustering coefficient (e) decentralized search and (f) robustness.

2 Hierarchical organization of modularity in metabolic networks

2.1 Summary

In Ravasz et al. [1] the authors analyzed the metabolic networks of 43 distinct organisms, by calculating the average clustering coefficient for each organism. They observed that the high, size independent clustering coefficient offers strong evidence for modularity, while the power law degree distribution of all metabolic networks strongly support the scale-free model, ruling out a modular topology. The paper proposes a simple heuristic model of metabolic organization, referred to as a "hierarchical" network, in order to resolve this conflict. In this model, a small cluster of four densely linked nodes is considered as a starting hypothetical module. Next, three replicas of this module are generated and the three external nodes of the replicated clusters are connected to the central node of the first cluster, obtaining a large 16-node module. These replication and connection steps are recursively applied to get a large network which integrates the properties of scale-free topology; high, system size independent clustering coefficient and the power-law scaling of $C(k) \sim k^{-1}$ in metabolic networks.

To verify the model, the authors subjected *E. colis* metabolic organization to a three step reduction process, by decreasing its complexity without altering the network topology. Next, the topological overlap matrix was investigated to check whether potential functional modules encoded in the network topology can be uncovered automatically. The results showed that the hierarchy contained nested topological modules of increasing sizes and decreasing interconnectedness. Also there are strong correlations between the global topological organization and the biochemical functional classification of the metabolites.

2.2 Critique

This paper was one of the early studies into networks having an inherent self-similar property, and elegantly resolved the conflicting observations of properties in metabolic networks. It showed that the system-level structure of cellular metabolism was best approximated by a seamlessly embedded modular structure in a hierarchical network organization. Some of the shortcomings of this paper were that the procedure of

choosing the starting module (4 nodes) was not delineated clearly. Also, some of the observations were said to be "visually apparent" without rigorous proofs.

3 Self-similar community structure in a network of human interactions

3.1 Summary

In Guimera et al. [2] the authors describe a procedure to analyze the hierarchical structure of nested communities based on a community detection algorithm due to Girvan and Newman(G&N). They analyze an email network and determine that the network self-organizes into a self-similar structure. They observe that the cumulative degree distribution is exponential unlike some real-world networks which incur a high cost of maintaining social connections unlike the email network. They also report the statistical properties of the giant connected component of the network and for comparison generate a random network with an exponential degree distribution(RE). The G&N algorithm repeatedly eliminates the edge with highest betweeness and recomputing the betweeness of the resulting network. It is also represented as a binary tree with the leaves representing the actual network nodes and the rest representing the community structure at each split. The results on the email network clearly show the community structure and the branching resembles those of river networks. In contrast, the branching of the RE network is trivial with no community structure. They compute the cumulative community size distribution from the binary tree and determine that it obeys power law and beyond size ∼100 shows a rapid decay followed by a cutoff corresponding to the size of the network like river networks. They consider another hierarchical model due to Ravasz et al. [1] which exhibits the same scaling behavior. In contrast, the RE network exhibits an entirely different distribution. They subsequently use the Horton-Strahler Index to compute a quantitative measure for topological self-similarity and find that other models including the RB models don't exhibit self-similarity except the community tree model. They conclude that the similarity with river networks may imply an underlying mechanism in the formation and evolution of social networks.

3.2 Critique

The authors use bidirectional emails and exclude bulk emails which is a good experimental design as it would represent a more accurate community structure. The comparison with the RE/RB network clearly explains the difference between the proposed model and other network models. However, the university email network they use is very small (∼1600 nodes). Additionally, their comparison and resemblance of properties with the Fella river network could be applied to a wider range of naturally occurring networks.

4 Self-similarity of complex networks

4.1 Summary

In Song et al. [3] the authors analyze several real-world networks and conclude that they have a self-similar structure by using a renormalization procedure that divides the network into boxes containing nodes within a given size and identify a power-law between the number of boxes required to cover the network and the size of the box. They describe:(a) small world (b) scale free properties of complex networks. They unravel the self-similar properties of such networks by computing the 'fractal dimension' using (a) box-counting (b) cluster growing methods. In box-counting, the network is covered with N_B boxes of size l_B. The fractal dimension $N_B \sim l_B^{-d_B}$. In cluster growing, a seed node is chosen and nodes separated by a minimum distance l are clustered around the seed node. This is repeated for many seed nodes and the average mass of the clusters is calculated as a function of l. The fractal dimension is then computed as $<M_c> \sim l^{d_f}$. For a homogeneous network the equations are equivalent because of same degree distribution whereas for a non-homogeneous network they behave differently. Thy run experiments on networks like the WWW etc. and confirm that all are scale-free. However, after renormalization, box-counting has a power law growth, whereas cluster growing has an exponential growth. They reason that the topology of these networks is dominated by several highly connected hubs, hence, the average obtained by the cluster growing method is biased, since almost

all nodes are connected to hubs in a very few steps. In the box-counting method (a)once a hub is covered it can't be covered again (b)each part of the network is covered with equal probability, thereby making it unbiased. The authors conjecture that the renormalized network generates a new probability distribution of links invariant under renormalization, and demonstrate its validity by showing a data collapse of all distributions for the WWW.

4.2 Critique

The authors have presented an intuitive box-counting method and shown that the renormalization does indeed unravel the self-similar nature of social networks backed up with experiments involving real-world networks. For demonstrating the validity of the invariance of link probability distribution, the authors show the collapse of distributions for WWW which has the following issues: (a) the sample WWW network is small and (b) the authors limit box sizes to only 2, 4, 6 which don't have a large difference to verify the invariance.

5 Origins of Fractality in the growth of complex networks

5.1 Summary

The authors look at the emergence of self-similarity in complex networks and investigate the process of evolution using the concept of renormalization for growth of fractal and non-fractal networks. In Song et al. [3], we talked about the fractal nature of organization in networks and this paper [4] builds on that and talks about the network evolution over time. They show that the architecture of fractal networks is mainly because of the strong repulsion (disassortativity) between hubs on all length scales thereby leading to a robust modular network with fractal topology. They propose a network growth dynamics as the inverse of the renormalization procedure i.e. if a network with $N(t)$ nodes is tiled with $N_B(l_B)$ boxes of size l_B, then each box represents a node in a previous time step. Using a probabilistic model, they show that Mode I (hub-hub connections) alone generates a network that shows small-world effect but is not fractal. Mode II (non-hub node in a box is connected to another non-hub node in a different box) alone gives rise to a fractal network. In general, the growth process is a stochastic combination of Mode I and II (this is consistent with the procedure described in [1]). Based on the above growth dynamics and by using the ratio of the number of links that are connected to hubs to the degree of the most connected node in the box, they show that fractal networks have strong hub repulsion at all length scales and non-fractal networks have weak hub repulsion. They also show that fractal networks are more robust as the hubs (core components) are dispersed over the network.

5.2 Critique

This simple model is very minimalistic and captures just one essential property of networks: relationship between anti correlation and fractality. Also, the models described are not generic and cannot be applied to all networks as they require modularity as a necessary network property.

6 Stochastic Kronecker graphs

6.1 Summary

Leskovec et al. [5] proposed the Kronecker Graph model which simultaneously captures several well-known properties of real-world networks; in particular, a heavy-tailed degree distribution, a low diameter, and the densification power law. This paper shows that Kronecker graphs naturally obey common network properties, and can accurately model global network structure using just four parameters. Leskovec et al. [5] fit the stochastic Kronecker graph model to some real world graphs, such as Internet Autonomous Systems graph and Epinion trust graphs, and found that with appropriate 22 initiator matrices many properties of the target graphs could be modeled very well. Properties of the Kronecker model such as connectivity and

diameter were rigorously analyzed in the deterministic case, and empirically shown in the general stochastic case.

In Mahdian et al. [6], the basic properties of stochastic Kronecker products are studied. Through a series of theorems, the authors show a phase transition for the emergence of the giant component and another phase transition for connectivity. The authors show that the necessary and sufficient condition for Kronecker graphs to be connected with high probability is $\beta + \gamma > 1$ or $\alpha = \beta = 1, \gamma = 0$, where α, β, γ are the entries of the 2x2 initiator matrix. The proof consists of using the min-cut size of the weighted graph. Further it is shown here that, the necessary and sufficient condition for Kronecker graphs to have a giant component of size $\theta(n)$ with high probability is $(\alpha + \beta)(\beta + \gamma) > 1$, or $(\alpha + \beta)(\beta + \gamma) = 1$ and $(\alpha + \beta) > (\beta + \gamma)$. They also prove that if $\beta + \gamma > 1$, the diameters of Kronecker graphs are constant with high probability, as per the observation of Leskovec et al. [7] that in many real-world graphs the effective diameters neither increase nor shrink as the sizes of the graphs increase. Finally it is shown in [6] that networks generated by Kronecker products are not searchable using a decentralized algorithm. First, a monotonicity result on general random graphs $G(n, P)$ is given, which is then used to prove that Kronecker graphs with $\alpha < 1$ is not poly-logarithmic searchable. That is, they do not admit short decentralized routing algorithms based on local information alone, unless the path is deterministic.

6.2 Critique

The positive aspects about this paper are that it gives clear mathematical arguments that justify the choice of stochastic Kronecker graphs to model real networks accurately. The results are also consistent with previously obtained results by Lescovec et al [5]. Some shortcomings of the paper are that it does not discuss the potential cause of the emergence of such properties in Kronecker Graphs. Also it is not very clear if the same results hold for graphs that are produced by initiator matrices of size greater than two.

7 Future Work

Further studies in [1] could involve using the mathematical framework for hierarchical modularity, and apply it to analyze other complex networks. [1] also conjectures that organization of metabolic networks is likely to result from an evolutionary mechanism to combine a capacity for rapid flux reorganization with a dynamic integration with cellular functions. This line of thought could be further investigated with respect to non-biological and other networks as well, by using a similar node-contraction or network-reduction process.

Using the quantitative framework in [2], we can compare: (a) flow of information, (b) robustness in the proposed network to other networks such as RB networks etc and observe whether it actually optimizes the metric in comparison to other models. We can also perform experiments on networks of large size and compare results with other naturally occurring networks.

This property of self-similarity as investigated in [3] can be extended by: (a) performing the analysis on large networks to demonstrate the validity of the invariance of degree distribution, (b) performing experiments to verify the property that most nodes are connected within one or few hops of hubs by computing the distribution of node distances from hubs on a variety of large social networks, (c) for the cluster growing method, we can try some variant as distributing the network into regions and then applying the method on a specific region, to overcome the bias in the cluster growing method and see how it compares to the box-counting method.

The analysis in [4] could be extended to look at other ways to prove robustness in a network. Also, applying the same techniques to non-modular networks to get a quick understanding of how the properties vary could be done. Other techniques to increase robustness in a network could be tried and then the same analysis could be applied for comparison.

The authors in [6] mention that the monotonicity theorem might lead to an independent line of thought, which can be further investigated. The theorems proved here could also perhaps be extended to initiator matrices of size greater than two. This paper considers three properties of stochastic Kronecker graphs, and the investigation of other properties of such a graph can also be studied in future work. Since it is

proved here that such graphs are not searchable using a decentralized algorithm, it may be of interest to find approximation algorithms to find such routes, or other techniques to enable faster decentralized routing algorithms.

8 Project Proposal

Our project proposal would include the study of Kronecker graphs and its properties along with some of the extensions mentioned above. We would analyze the model on the possible dataset(s) mentioned below. Some of the properties that we could investigate are: basic network structure, identifying fractality, robustness of the network, flow of information/influence (i.e. influence maximization). We can also experiment with different values for the parameters of the 2x2 initiator matrix for the Kronecker graph model to fit it to the below mentioned dataset(s) or generated variants of it. This could be done using the KronFit algorithm described in [5].

Possible datasets: Twitter link graph, Wikipedia, Blog links.

References

[1] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabasi, A.L., Hierarchical organization of modularity in metabolic networks, Science 297, 15511555 (2002).

[2] Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F. & Arenas, A., Self-similar community structure in a network of human interactions, Physical Review E 68, 065103(R) (2003).

[3] Song, C., Havlin, S. & Makse, H. A., Self-similarity of complex networks, Nature 433, 392395 (2005).

[4] Song, C., Havlin, S. & Makse, H. A., Origins of Fractality in the growth of complex networks, Nature, April 2006.

[5] Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., Ghahramani Z., Kronecker graphs: An Approach to Modeling Networks, Journal of Machine Learning Research (JMLR) 11(Feb):985-1042, 2010.

[6] Mahdian, M., Xu, Y., Stochastic Kronecker graphs, WAW 2007, LNCS 4863, pp. 179186, 2007.

[7] Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: Densification and shrinking diameters, ACM Transactions on Knowledge Discovery from Data 1(1) (2007).