CS224W: Social and Information Network Analysis

Project Proposal

Adithya Rao, Gautam Kumar Parai, Sandeep Sripada

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

1 Introduction

In this project, we plan to explore the property of self-similarity exhibited by real world networks, and the use of Kronecker graphs to model and analyze such networks. It has been observed that self-similarity is an emergent property of many real world networks such as WWW, e-mail and biological networks. These networks show properties such as heavy tails for the in- and out-degree distribution, heavy tails for the eigenvalues and eigenvectors, small diameters, and densification and shrinking diameters over time. Recently, Kronecker Graphs have been shown to elegantly model these networks, while being mathematically tractable [5]. We would like to show that this property also holds for a social network like Twitter.

2 Previous Related Work

In some of the early work in this area, Ravasz et al. [1] analyzed the metabolic networks of 43 distinct organisms, by calculating the average clustering coefficient for each organism. They proposed a simple heuristic model of metabolic organization, referred to as a "hierarchical" network, which showed both scale-free nature and modularity. Somewhat surprisingly, such self-similarity was also observed in human network interactions, and was studied by Guimera et al. [2]. They analyzed an email network and determined that the network self-organizes into a self-similar structure. The results on the email network clearly showed that the community structure and the branching resembled those of river networks.

The fact that this property of self-similarity was observed in various real world networks, led to further studies to understand the underlying mechanism in the formation and evolution of such networks. In [3], Song et al. analyzed several real-world networks and computed their 'fractal dimension' using box counting and cluster growing methods. The authors conjectured that the renormalized network generates a new probability distribution of links invariant under renormalization, and demonstrated its validity by showing a data collapse of all distributions for the WWW. Extending this work in [4], the authors showed that the architecture of fractal networks is mainly because of the strong repulsion (disassortativity) between hubs on all length scales, leading to a robust modular network with fractal topology. They also proposed that network growth dynamics could be modeled as the inverse of the renormalization procedure.

In [5], Leskovec et al. introduced Kronecker Graphs, a generative network model which obeyed not only all the main static network patterns that had been observed in real networks, but also temporal evolution patterns. In this paper, the authors also presented KRONFIT, a fast and scalable algorithm for fitting Kronecker graphs by using the maximum likelihood principle. The algorithm was then used to fit the stochastic Kronecker graph model to some real world graphs, such as Internet Autonomous Systems graph and Epinion trust graphs. 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 (also in [7]). In [6], Mahdian et al. studied the basic properties of stochastic Kronecker products. Through a series of theorems, the authors showed a phase transition for the emergence of the giant component and another phase transition for connectivity. They also showed that Kronecker Graphs do not admit short decentralized routing algorithms based on local information alone, unless the path is deterministic.

In this project, we aim to extend the work in [5], [6] and [7], and analyze the Kronecker Graph model as applied to other real world networks such as Twitter and/or Wikipedia. The next two sections describe the aspects that we plan to consider and the goals we aim to accomplish.

3 Tasks to be accomplished

In [5], it is shown that Kronecker Graph is constructed by repeatedly taking the Kronecker product of an initiator matrix. The Kronecker graph model is thus based on a recursive construction of self similar graphs. The major advantage of Kronecker graphs over other network models is that it is possible to prove analytical results about many properties related to the real-world network unlike previous models which target specific properties. The entries in the initiator matrix can take values between 0 and 1, to generate a stochastic model rather than a deterministic one. In this case, the stochastic adjacency matrix encodes the probability of the particular edge appearing in the graph. Two natural interpretations of the generative process are:

- Networks are hierarchically organized into communities (clusters), which grow recursively, creating miniature copies of themselves.
- Each node is described by a sequence of categorical attribute values or features, where the probability of two nodes linking depends on the product of individual attribute similarities.

In [5], the authors address the issue of automatically estimating the Kronecker initiator graph parameters. The approach to the problem of estimating Stochastic Kronecker initiator matrix is by defining the likelihood over the individual entries of the graph adjacency matrix. It was also shown that the KRONFIT algorithm efficiently performs this fitting in linear time, and the corresponding results for AS-Routeviews and Epinions were presented.

3.1 Task 1

The second interpretation mentioned above can be explored by changing the values of entries to observe their effect on the generated graphs. Homophily can be modeled by higher diagonal entries and heterophily is modeled by higher off-diagonal entries. We plan to explore this variety by creating models using various values for initial parameters and observing the properties of the generated graphs.

3.2 Task 2

Graph similarity: Once the parameters for the Twitter and/or Wikipedia have been estimated, we can compare them to that of networks of different structures and sizes, whose parameters have been found in [5]. The estimated parameters can be used as a similarity measure.

3.3 Task 3

Given a real network G, the aim is to discover the most likely parameters that ideally would generate a synthetic graph K having similar properties as real G. We would like to apply similar techniques to the Twitter and/or Wikipedia network datasets. Thus given the Twitter dataset G, we would run KRONFIT to obtain parameter estimates $\hat{\theta}$. Using the $\hat{\theta}$ we then generate a synthetic Kronecker graph K, and compare the properties of G and K as mentioned in the section 4.

3.4 Task 4

It was shown in [5] that parameters estimated from the historic data can extrapolate the graph structure in the future for a temporally evolving graph. We aim to examine this temporal evolution of the Twitter and/or Wikipedia networks, once the parameters have been estimated. Several dynamic properties of evolving graphs can then be examined and modeled, as they evolve over time. The metrics that we would use are enumerated in section 4.

3.5 Task 5

Influence Maximization:

- One of the aims of the project is to leverage the self-similar structure of the graph to choose a subset of nodes which would maximize the spread of influence. Because of the self similar structure, we conjecture that there could be an efficient way to generate the subset of initial nodes.
- Another question we would like to explore is whether performing influence maximization on a subset of the network would scale to the entire network. If so, are there any specific node characteristics that we could exploit while scaling to the entire network. We would like to compare the performance of such a scheme to the linear threshold and independent cascade models, as well as the approximation algorithm in [8].

The following tasks would be taken up if time permits.

3.6 Task 6

Sampling and Properties across scales: Since the graph is self-similar, we can observe various properties across different scales of the network. If such properties are verified to be true across scales, then it may be useful to run simulation experiments on smaller graphs of the same network as these algorithms may be too slow to run on the entire graph. Some such properties would be:

- Robustness: We would want to observe how vulnerable the network is by removing nodes in progressively larger subsets of the graph, and comparing the average path lengths. We expect that the graph would have similar behavior across all scales.
- Information cascades: We would also like to observe how information would cascade at different scales of the network.

3.7 Task 7

By using more than one initiator matrix, we can model multiple attributes of the network. In this case each initiator matrix is an attribute similarity matrix for a particular attribute.

3.8 Doubts

- We are unclear at this point, on how will we map a subset of nodes from the synthetic network to some subset of nodes in the real world network, so as to apply node properties found in the smaller subset to nodes in the entire graph.
- We might want to try the same on Wikipedia if the Twitter dataset does not work well.

4 Evaluation

In addition to the tasks above, we would also like to evaluate and compare properties such as degree distribution, scree plot (eigenvalues of graph adjacency matrix vs. rank) and hop plots for the static instances of the Twitter and/or Wikipedia datasets. We would also evaluate temporal patterns including the diameter over time, and the densification power law.

Thus broadly our evaluation would involve the following static properties:

- Network structure: Degree distribution, small world property.
- Hop-plot: Such a plot (number of reachable pairs g(h) within h hops vs hops h) gives us a sense of how quickly nodes neighborhoods expand with the number of hops.

- Scree plot: Plot of the eigenvalues (or singular values) of the graph adjacency matrix, versus their rank, using the logarithmic scale.
- Node triangle participation: Node triangle participation measures the transitivity in networks by counting the number of triangles a node participates in.

We would also verify that the Twitter and/or Wikipedia networks as well as their Kronecker models show the temporal properties that:

- Densification power law: The graphs obey the densification power law (DPL) as a function of time.
- Shrinking diameter: The diameter shrinks and then seems to stabilize as the network grows.

5 Data & Algorithms

Data: Twitter and/or Wikipedia
Source:

- http://an.kaist.ac.kr/traces/WWW2010.html
- Tweet dataset for the experiments on influence maximization

Algorithms: Use algorithms described above from SNAP, NetworkX packages

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).

[8] Kempe, D., Kleinberg, J., Tardos, E., Maximizing the Spread of Influence through a Social Network, SIGKDD '03 Washington, DC, USA