Networks as vectors of their motif frequencies and 2-norm distance as a measure of similarity

CS322 Project Writeup

Semih Salihoglu
Stanford University
353 Serra Street
Stanford, CA
semih@stanford.edu

ABSTRACT

Previous studies have uncovered a list of interesting properties that exist in real networks by comparing them to synthetic networks. However, there has been less focus on comparing real networks to other real networks. How do Autonomous Systems routing networks compare to social networks? How can we tell that one biological network looks different from other biological networks? How can we quantify how similar two networks are? One way to capture the difference between two networks is to define an "appropriate" notion of distance. This project proposes using the 2-norm distance between vectors to capture network similarity. In particular, we present the results of using the 2-norm distance as a measure of similarity in two cases: a) when networks are represented as 4 dimensional vectors based on their 2x2 stochastic Kronecker initiator matrices, b) when networks are represented as 8 dimensional vectors based on the frequencies of their 3- and 4-size motifs. We show that the motif frequencies are able to capture the similarity among the following classes of networks in our data set: road networks, AS routing networks, biological networks, p2p networks, social networks, citation networks, collaboration networks and item co-purchasing networks.

General Terms

Network Analysis

Keywords

Network similarity, Kronecker graphs, Network motifs

1. INTRODUCTION

A large body of previous work focuses on understanding how real networks differ from synthetic networks. A common line of analysis in many of these studies is the following: 1) Identify a set of structural properties of real networks. 2) Understand the network evolution processes that would give rise to such properties. 3) Propose a model that will generate networks with these properties. [1] is a good example of this line of analysis. The authors observe that while many real networks have power-law or (scale-free) degree distributions, random networks do not exhibit this property. They then propose the Barabási-Albert network generation model that replicates this property. Similarly [4] observes that the communities in real networks are much tighter at small sizes than the communities in random networks. Along with this observation, they propose the forest-fire model to generate such networks. [2] is a great survey of properties observed in real networks and the corresponding generative models.

In most of these studies, no distinction is made between various types of real networks. The term "real networks" keeps appearing and reappearing in these studies to refer to the data the authors used in their studies. However, the data available to the networks community consist of hundreds of networks, some of which are much different than others. The motivating questions for this project are: Are there different classes of real networks? If so, what are the fundamental differences between these classes? In order to be able to answer these questions, we have to answer an even higher level question: How can we compare real networks to each other? One way to answer this question is to define a notion of distance between networks as a measure of similarity. Once we have a notion of distance, we can compute the distances between each pair of networks and run clustering algorithms on these distances. Once we compute some clusters, we can then try to understand the fundamental differences between them.

In this project we represent networks as vectors in two different ways and use the 2-norm distance between vectors as a measure of similarity. First, we represent each network as a 4 dimensional vector based on the 2x2 stochastic Kronecker initiator matrix of the network. Second we represent each network as an 8 dimensional vector, based on the frequencies of the 3- and 4-size motifs in the network. The main contribution of this project is to show that when networks are represented as vectors of their motif frequencies, 2-norm distance is able to capture the similarity of most classes of networks we use: road, AS routing, biological, p2p, social, citation, collaboration, and item co-purchasing networks. The Kronecker matrix representation also captures the similarity between road networks, AS routing networks, and item co-purchasing networks but the remaining classes of networks are not well distinguished from each other.

The rest of this paper is organized as follows: Section 2 describes the two ways we represent networks as vectors, the two ways we visualize similarity between vectors and the data set we use in our experiments. In Section 3 we present the results of our visualization methods. In Section

4, we discuss future work.

2. METHODOLOGY

The outline of our basic methodology is the following: We experiment with two different ways of representing networks as vectors: a) We represent each network as a 4 dimensional vector based on the 2x2 stochastic Kronecker initiator matrix of the network. b) We represent each network as an 8 dimensional vector, based on the frequencies of the 3- and 4-size motifs in the network. Once we represent each network as a vector, we take these vectors and visualize the similarity between them in two different ways: a) We reduce the dimensions of each vector by principal component analysis (PCA) to 2 and plot these points in 2D. b) We construct similarity graphs. We used the following 46 networks in our experiments:

- 6 AS routing networks
- 9 p2p file sharing networks (these are actually 9 snapshots of the Gnutella network)
- 6 biological networks
- 3 road networks of different states in the US
- 3 academic citation networks
- 4 academic collaboration networks
- 3 Amazon item co-purchasing networks
- 5 social networks
- 4 web graphs (each is inferred by a different organization)
- 3 communication networks

2.1 Representing networks as vectors

2.1.1 *Networks as 2x2 Stochastic Kronecker Initiator Matrices*

Kronecker graphs [3] are random graphs generated by Kronecker multiplication of probability matrices. In [3], the authors prove that this model is able to capture many properties that exist in real networks: small diameter, heavy-tails for the degree distributions, heavy-tails for the eigenvalues and eigenvectors, and the densification and shrinking of diameters over time. They also present an algorithm called KRONFIT. KRONFIT takes a target real network $G(V, E)$, a column size c, and a row size r as inputs, and estimates a $c \times r$ initiator matrix I, the recursive multiplication of which will generate a network that will look like the input target network. They experimentally show that KRONFIT is able to accurately mimic the properties of large target networks even by a 2x2 initiator matrix.

Our first way of representing each network as a vector is the following: For each network N, we compute the 2x2 initiator matrix I by using KRONFIT. Once we have I, we represent N as a 4 dimensional vector defined as: $v_N = (I[1,1], I[1,2], I[2,1], I[2,2])$. Each cell in I corresponds to one dimension in the vector v_N. Kronecker initiator matrices has the desirable property that every cell is a probability and has a range of [0, 1]. Thus all the values of v_N are between 0 and 1. Table 1 lists sample 4 dimensional Kronecker vectors and the networks they correspond to.

Table 1: Sample networks as 4 dimensional Kronecker vectors

	Cell(1,1)	Cell(1,2)	Cell(2,1)	Cell (2,2)
AS-1	0.94	0.65	0.49	0.05
Web-1	0.69	0.67	0.66	0.08
P2P-1	0.72	0.39	0.39	0.53
Soc-1	0.77	0.51	0.50	0.40

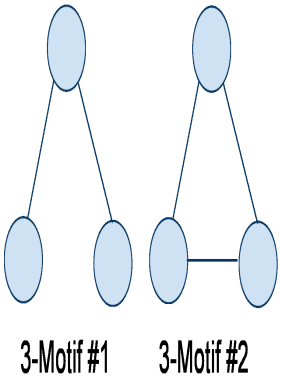

Figure 1: 2 possible 3-size motifs for undirected graphs.

2.1.2 *Networks as frequencies of their 3- and 4-size motifs*

Network motifs [5] are small, fixed-size, connected subgraphs within a graph. Figure 1 shows the two 3-size motifs for undirected graphs (there are only 2 ways 3 nodes can be connected in an undirected graph: a) the open triangle, b) the closed triangle). In [5], the authors identify some motifs that occur in real networks at a significantly higher frequency than in random networks. Again the focus of comparison in this study is between real networks and random networks.

We take this idea and use it to represent real networks as vectors in the following way: We first compute the frequency of each 3- and 4-size motif. There are 2 3-size motifs and 6 4-size motifs in undirected graphs. We take these 8 frequencies as the values of a vector in 8 dimensions. Each motif frequency corresponds to a dimension. Note that by definition all frequencies are between 0 and 1. Moreover the frequencies of the 2 3-size motifs and the 6 4-size motifs add up to 1. Table 2 lists some sample 8 dimensional motif frequency vectors.

2.2 Methods to visualize similiarity between networks

Once we have our vector sets, we use PCA and similarity graphs to visualize the distances between the vectors.

2.2.1 *PCA in 2D*

Table 2: Sample networks as 8 dimensional motif frequency vectors

	3M #1	3M #2	4M #1	4M #2	4M #3	4M #4	4M #5	4M #6
AS-1	0.98	0.02	0.75	0.18	0.04	0.01	0.005	0.0003
Web-1	0.79	0.21	0.58	0.07	0.16	0.02	0.11	0.05
P2P-1	0.99	0.01	0.41	0.57	0.01	0.005	0.0007	0.00001
Soc-1	0.98	0.02	0.48	0.45	0.05	0.0005	0.006	0.002

Given our vectors in either 4 or 8 dimensions, we run standard principal component analysis to project each vector onto a 2 dimensional space. Let's consider the 8 dimensional case as an example. We construct a 46×8 matrix M, where each row is a vector. We extract the mean matrix from M and get D. We compute the covariance matrix $C = D'D$ of D. We compute the eigenvectors of C and the corresponding eigenvalues. We take the 2 eigenvectors with the 2 highest eigenvalues and project each original vector onto this space. Once we have the projections, we plot them in 2D. This gives us a non-rigorous way to see which networks are close to each other and hence similar. Figure 2 and 3 show the 2D PCA plots for the Kronecker and motif frequency vectors respectively. We should note that we also projected onto 3D in a similar fashion but there was not a significant change in the visual plots for either case. Hence, we are only presenting the 2D results in this write-up.

2.2.2 *Similarity Graphs*

Another method to visualize the similarities between networks is to construct similarity graphs. A similarity graph $S(V, E)$ is a graph where each network is a node and there is an edge (u, v) between two networks if the 2-norm distance between their vector representations is less than a given threshold τ. We experimented with different values for τ, and 0.08 for the Kronecker and 0.05 for the motif frequency vectors gave good clusterings in the computed similarity graphs respectively. Figures 4 and 5 show the similarity graphs for these two cases. We note that to keep the similarity graphs visually cleaner, we are not plotting the isolated nodes.

3. RESULTS

The main result of this project is that motif frequencies and the 2-norm distance together capture the similarity amongst all classes of networks except the web and communication networks. The similarity graph for this method has 9 connected components corresponding to 8 classes of networks (2 of the components are p2p networks): road, AS routing, biological, p2p, social, citation, collaboration, and item copurchasing networks. We can see this on the 2D PCA graph as well. Note that each of these classes are visually very close to each other and far from other clusters of points. The only two classes of networks, which motif frequencies are unable to cluster, are the web and the communication networks. Web graphs are spread out in the middle of the PCA graph. One of the communication graphs is on the top left corner, isolated from other networks. The other is in middle of the graph, again isolated from other networks. We should note that the 2D PCA graph for the motif frequency vectors does not contain 1 web, 2 social and 1 communication networks. These were the largest networks in our data set and at the time of this write-up the jobs computing their motif counts had not completed. These clustering of networks fit our preclustering of these networks very well. We believe this is a desirable property for any distance measure, as we would intuitively expect, for example, all road, p2p, and citation networks to look similar internally and different from each other.

Kronecker vectors are also able to cluster road, AS routing, and the biological networks (some of which do blend in with other networks). However one of the connected components in the similarity graph contains a wide variety of networks: p2p, communication, biological, collaboration, citation, and the web networks. We tried different values for τ to get a more refined clustering but observed roughly one of the three cases: a) When $\tau < 0.07$, most nodes are isolated and there is only a few very small connected components. b) When $0.07 \leq \tau \leq 0.11$, we get a similarity matrix that looks like the one in Figure 4. c) When $\tau > 0.11$, the similarity graph looks like a clique, where most networks connect to most other networks. We think $\tau = 0.08$ gives a good representative best clustering (we realize these terms are subjective and we see this as a shortcoming of our study).

4. FUTURE WORK

One other method that we experimented with but have not presented in this write-up is representing networks as vectors of their network properties. In this method we compute some network properties for each network and consider each property as a dimension. In particular we computed the following 5 properties: effective diameter, average degree of the nodes, average clustering coefficient, fraction of nodes in the largest connected component, largest singular value of the adjacency matrix. We then represented each network as a 5 dimensional vector. This method did not produce good clusters with our choice of 5 properties. One problem with this attempt is that, unlike the two representations presented here, the values for each dimension can take on a different range of values. However we still think this is a promising direction. We believe whatever information is carried in motif frequencies would be reflected in some combination of network properties. Returning back to this direction and experimenting with more properties could yield good results in the future. Aside from this, we believe there are 5 immediate further steps we should take in this study:

- Increasing our data set from 46 to hundreds of real networks. We also think that adding some random networks to our data set will give us insights about the actual meaning of the 2-norm distance.
- Extending the 4-size motifs to 5- and 6-size motifs.
- Computing the average inter-cluster vs. intra-cluster distances to quantify tight and loose clusters. This

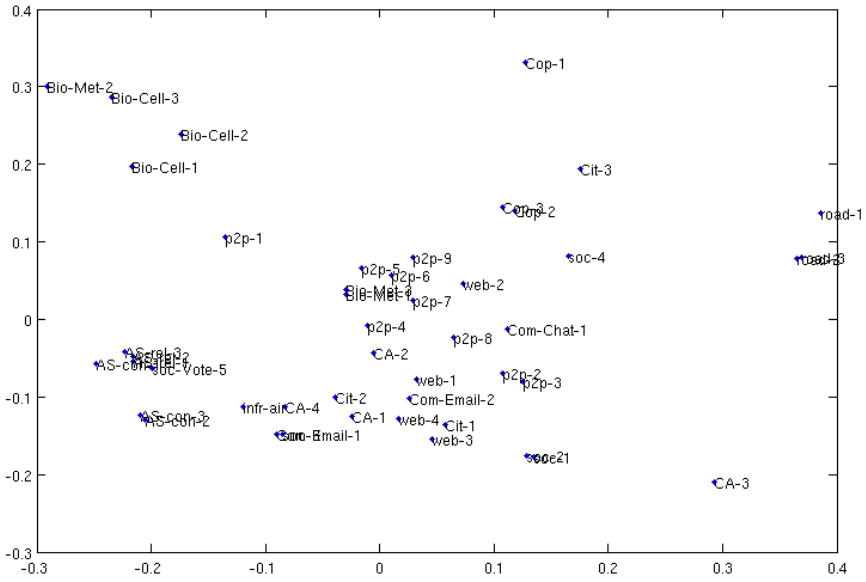

Figure 2: 2D PCA results for the Kronecker vector representation.

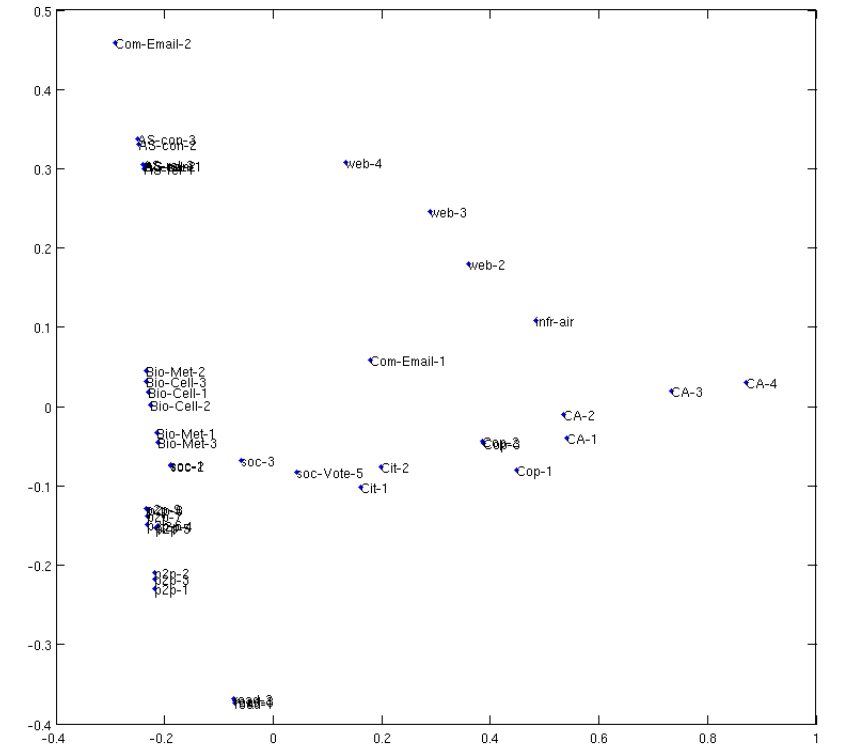

Figure 3: 2D PCA results for the motif frequency vector representation.

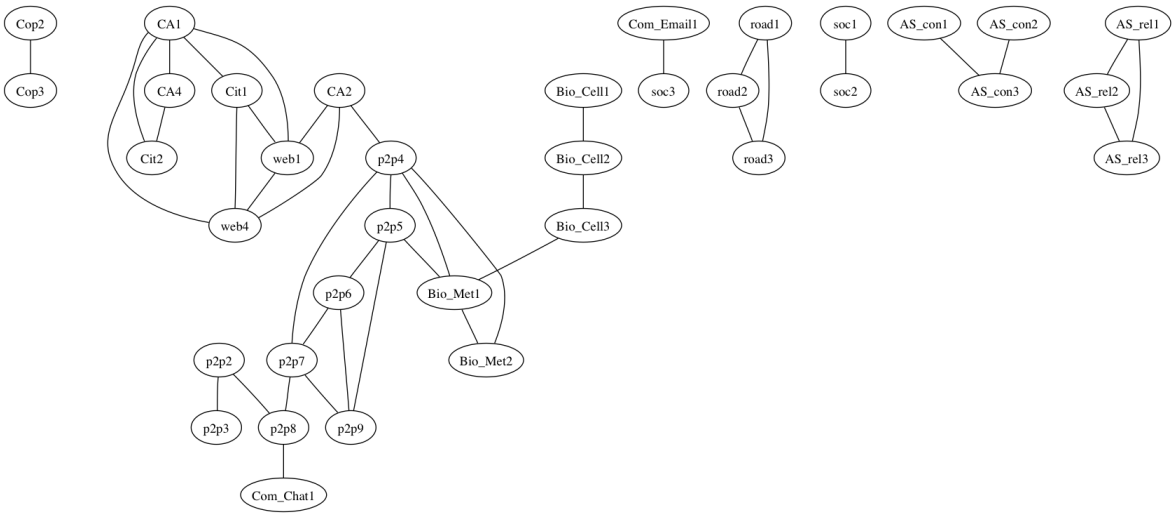

Figure 4: Similarity graph results for the Kronecker vector representation with $\tau = 0.08$.

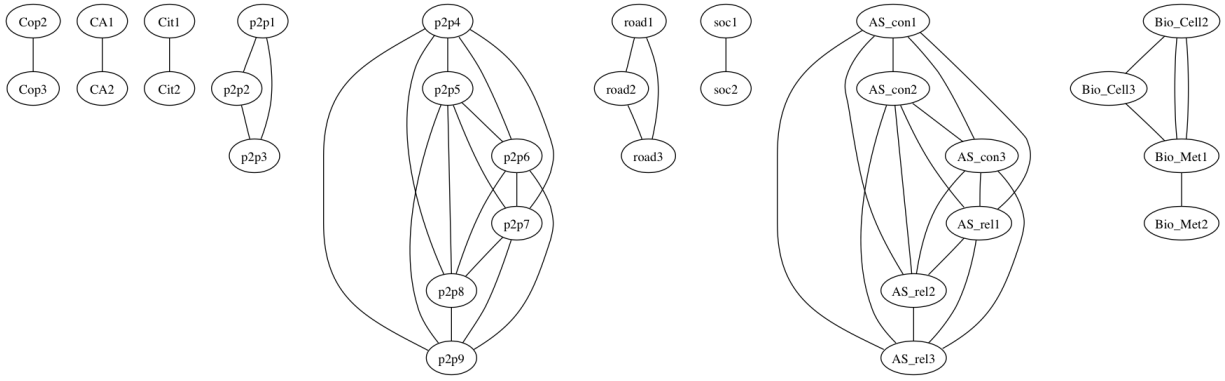

Figure 5: Similarity graph results for the motif frequency vector representation with $\tau = 0.05$.

could also give us a method to pick our τ in a more principled way.

- Constructing decision trees to compare the precision of how well the two vector representations (and possibly the third method we outline above) are able to correctly label each network.

Understanding how we can compare different real networks is the first step in discovering different classes of real networks. If we can uncover the topology of real networks, we can start asking more interesting questions: What fundamental differences are underlying the different classes of networks? Which models perform well on which classes of network? How can we model each class of networks? We believe these are fundamental questions we should be asking to understand how networks evolve in the real world.

5. ACKNOWLEDGEMENTS

I am grateful to Maks Ovsjanikov for his help in running PCAs on our data set.

6. REFERENCES

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[2] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys*, 38(1):2, March 2006.

[3] D. C. J. Leskovec, J. Kleinberg, C. Faloustos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *arXiv*, 0812.4905, August 2009.

[4] J. Leskovec, A. D. Kevin Lang, and M. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *arXiv*, 0810.1355, October 2008.

[5] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, October 2002.