arXiv:1011.2336v1 [cs.SI] 10 Nov 2010

A network model with structured nodes

Pierluigi Frisco
School of Math. and Comp. Sciences, Heriot-Watt University,
EH14 4AS Edinburgh, UK,
P.Frisco@hw.ac.uk

Abstract

We present a network model in which words over a specific alphabet, called *structures*, are associated to each node and undirected edges are added depending on some distance between different structures.

It is shown that this model can generate, without the use of preferential attachment or any other heuristic, networks with topological features similar to biological networks: power law degree distribution, clustering coefficient independent from the network size, etc.

Specific biological networks (*C. Elegans* neural network and *E. Coli* protein-protein interaction network) are replicated using this model.

1 Introduction

In the last years mathematical and computer science (CS) concepts and methodologies and have been successfully used in Biology. This fascinating and fruitful combination of these disciplines has clear advantages for both of them. When biological phenomena are regarded as information processes, then they can be studied using mathematical and CS tools and concepts. This gives to Biology new ways to approach problems, solutions to them and this deepens the understanding of biological processes. At the same time CS enriches itself with new ways to define and study information process while Mathematics enriches itself with new concepts and theories.

In the last decade several studies ([2, 13, 9, 12]) showed the importance of the topology of biological networks. These results proved that biological networks are composed of motifs, that biological networks with specific functions have an abundance of certain motifs instead of others, that the number of edges for the node in the network follows specific laws, etc.

More than studying the features of empirical networks, it is also important to have algorithms able to generate networks with the same features of empirical ones. This kind of algorithms, called *models*, are an invaluable help in the generation of artificial networks and they provide insights on how certain features of complex empirical networks arise from the construction rules present in the model.

Examples of such procedures are: the *Erdös-Rényi* model [5], the *Watts-Strogatz* model [17] and the *Barábasi-Albert* model [3] and its variants [1, 14]. The E-R model allows to generate random networks able to reproduce the small-world property (short path from any node to any other node in the network) but they fail to account for the local clustering characterising many empirical networks. Both these properties are captures by the W-S model, but unfortunately it does not capture the inhomogeneous degree distribution found in

many empirical networks. The B-A model can overcome these limitations and gives rise to the degree distribution. This degree distribution is obtained using preferential attachment: the probability for a node to receive an edge depends on the number of edges the node already has. The original B-A model does not capture the independence of the clustering coefficient from the size (number of nodes) of a network. This feature is captured by a variant [14] of this model in which heuristics (replication of networks) are used.

The present study originates from the wish to create a network model able to reproduce biological networks without the use of heuristics. Despite the very many successful applications of the B-A model, it was not clear to us how preferential attachment could have been present in the evolution of, say, gene networks. Why a gene with many interactions is more likely to get even more interactions than a gene with few interactions? How can a new added gene "know" what are the genes with more interactions? In this respect, we believe that preferential attachment capture the overall effect of something more basic present in the evolution of biological networks.

The network model introduced and studied in the present paper tries to capture some basic features present in the evolution of biological networks: network growth, node structure and distance between node structures.

The node structure represents, for instance, the DNA sequence in genes, proteins' secondary structure, the personality features in humans, etc. The distance between nodes represents, for instance, the fact that proteins will interact if their tertiary structure (which depends on their secondary structure) allows it, or that two humans will be friends if the treats of their personality are somehow close.

In the following we present the model with structured nodes (Section 2), we analyse it (Section 3) and we use it to generate specific biological networks (Section 4). The paper ends with a discussion section (Section 5). Supplementary material (further technical details, generated networks, program implementing the proposed network model, etc.) is present at [11].

2 Description of the model

The *network model with structured node* (SN model) is such that each node in the network has a *structure*: a word over a specified alphabet. Given initial nodes have different structure. Nodes are added to the network one by one. Each new node has a structure given by the modification of a randomly chosen structure already present in the network. If the structure of the new node is already present in the network, then the new node is not added (that is, in the network all nodes have different structure). If the structure of the new node is not present in the network and the new node has no edge with the existing nodes, then the new node is not added (that is, isolated nodes are not allowed). Undirected edges are added to the network depending on a given distance between node structures. This process is repeated until the network

reaches a given number of nodes. A simple example follows.

Let us assume that the *alphabet* is {`A, B, C`} and that the network contains only one *initial node* with structure `ABCABC`. Edges between nodes are added only if the *Hamming distance* [18] between the structures of the nodes is at most 1.

A node can be added to the network by *mutating* one symbol in the structure of an existing node. For instance, the node `ABBABC` can be obtained mutating the third symbol of `ABCABC`. An edge is added between the two nodes (they only differ in one symbol).

A third node can be added to the network by *adding* one symbol to the structure of a randomly selected existing node. For instance, the node `ABBABBC` can be obtained adding a `B` between `B` and `C` in node `ABBABC`. An edge is added between the new node and `ABBABC` (when computing the distance between two structures exceeding symbols in the longer structure are disregarded). No edge is added between the new node and `ABCABC` because there are 2 differences in their first 6 characters.

The structure `ABCBC` can be obtained from `ABCABC` *deleting* the second `B`. The node with this new structure does not become part of the network as no edge has been added (the distance between `ABCBC` and the other structures present in the network is bigger than 1).

The structure `ABBBBABBC` can be obtained from `ABBABBC` *duplicating* the second and third `B`. The node with this new structure does not become part of the network as no edge has been added.

Input parameters define the probabilities to mutate, add, delete and duplicate node structures and their values has to sum up to 1.

We also used a Hamming distance in which the comparison between symbols considers groups of consecutive symbols. The order of the symbols present in each such group is irrelevant to the distance. For instance, let us consider the two structures `ABBABC` and `BABCAB`. If the *unit distance* is 1 (i.e., symbols are compared one by one), then the distance between the two structures is 5 as the only matching symbol is the `B` in the third position. If the unit distance is 2 (i.e., pairs of symbols are compared), then the distance between the two structures is 2. This is because the first two pairs are considered equal (`AB` and `BA` differ only in the order of the symbols), and the other two pairs are different in the symbols they contain. If the unit distance is 3 (i.e., triplets of symbols are compared), then the distance between the two structures is 0. This is because the first triples are considered equal (`ABB` and `BAB` differ only in the order of symbols) as well as the second triple (`ABC` and `CAB` differ only in the order of symbols).

An edge between two nodes is present only if their distance is smaller/equal than the value of the input parameter *maximum distance*.

When unit distance is bigger than 1, then it is possible to have a *file matches* indicating how the different groups of symbols can be matched to eachother. In other words, a file matches behaves as the genetic code: it denotes which tuples of symbols have to be regarded as equal (in the same way different codons translate in the same amino acid). For instance, let unit distance be 2, the

alphabet be {`A, B`}, and the file matches be:

```
AB =
BA =
AA = BB
BB = AA
```

With this file matches, the strings `ABBB` and `ABAA` have distance 0. This is because the first pair (`AB`) is the same in both strings, while the second pair (`BB` and `AA`) is defined by the file matches to be equal. Without the file matches, the two string have distance 1 (due to the second pair).

We call *instance* a set of input parameters. The complete list of input parameters together with their description can be found in the user manual of the program implementing the SN model [11].

3 Analysis of the model

We assessed our network model over the following network topological features [10]. Given an undirected network G with N nodes and k edges we denote by $\langle k \rangle$ the *average degree*, by L the *average path length*, by C the *average clustering coefficient*, by $P(k)$ the *degree distribution* and by $C(k)$ the *clustering coefficient distribution*.

We also considered the:

3-node *motifs distribution*, that is the number (normalised to 1) of triples of nodes having no edge, only 1 edge, only 2 edges and 3 edges between themselves;

path length distribution, denoted by $PL(\ell)$, relating the number (normalised to one) of paths having a certain length ℓ;

heterogeneity index, denoted by $\rho(G)$ (where G is the network), a new formulation of Randić index introduced in [7, 6]. In [7, 6] it is also shown that the Barabási-Albert model is not able to generate network with a heterogeneity index as high as the one found in biological networks.

We compared the network generated by an instance our the SN model with the network generated by the Barabási-Albert model (our implementation of this model is based on the Fortran implementation present at [16]). For this purpose we run the Barabási-Albert model starting with a clique of 6 nodes and adding 6 edges for each new added node. We also run the following instance of our network model: initial node `ABCDEFGHILMN`, alphabet `A, ..., T`, probability to mutate 1 (which implies that the length of the node structures is equal to the one of the initial node), Hamming distance having unit distance 2 and maximum distance 2. We run these simulation for 3000 iterations storing the resulting intermediate networks every 500 iterations. These tests run 100 times for different random seeds.

Figure 1 shows how the average degree, average path length and average clustering coefficient change in the Barabási-Albert model and in the SN model.

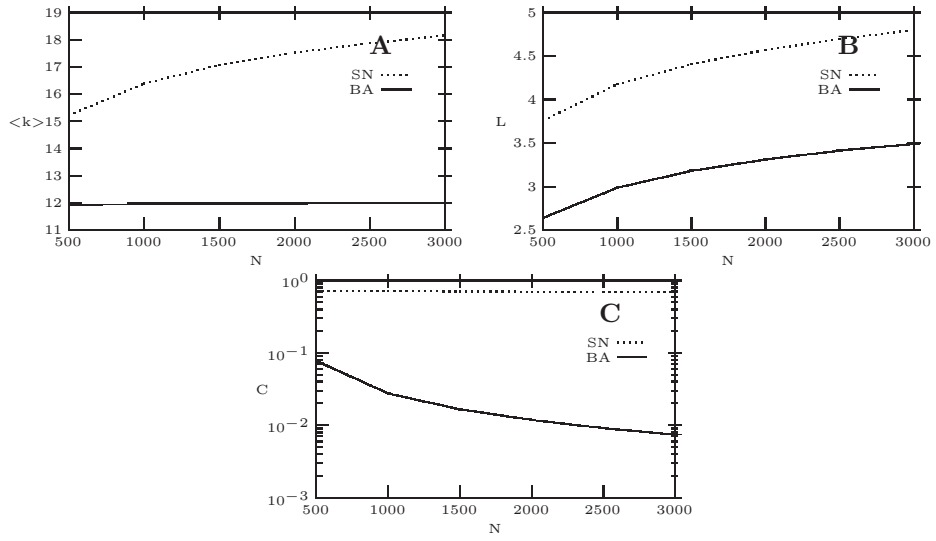

Figure 1: **(A)** average degree, **(B)** average path length and **(C)** average clustering coefficient of a growing Barabási-Albert model network (BA) and a growing SN model network.

The average path length follows the same curve in both models and the average degree slowly grows in the SN model while it remain constant in the Barabási-Albert model. The major difference is present in the clustering coefficient: in remains constant in the SN model while it decreases fast in the Barabási-Albert model. It is known that empirical networks have a clustering coefficient independent from their size and in [15] a variant of the Barabási-Albert model generating networks with a power law degree distribution and a clustering coefficient independent from the size of the network was presented. The motif distribution was similar in both model (data not shown).

It is well known that the Barabási-Albert model generates networks with a degree distribution following a power law $P(k) \sim k^{-\gamma}$. The same holds true for the considered instance of the SN model (this is not true for all instances of the SN model).

In both models the exponent of the power law does not change during growth. Anyhow, in the considered instance, the degree distribution of the networks generated by the SN model is not following a power law in its initial phases. This is shown in Fig. 2A where it can be seen that only after $k = 5$ the degree distribution follows a power law. This difference with the Barabási-Albert model is mainly due to the fact that in the Barabási-Albert model each new added node

has a fixed number of edges (6 in the case considered by us), while this request for a minimum number of edges is not present in the SN model.

We run another instance of the SN model for 55000 iterations and then we let all nodes having less than 5 edges to be removed from the generated networks together with their edges. The resulting networks, having around 3000 nodes, have a power law degree distribution Fig. 2B.

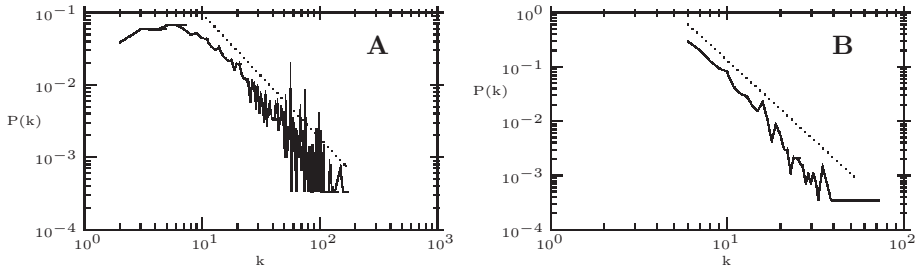

Figure 2: Degree distribution of two instances of the SN model: **(A)**: keeping all nodes (final trend having slope $\gamma = -1.72$), **(B)** removing nodes with less than 5 edges together with their edges (trend having slope $\gamma = -2.98$).

As already said, in the SN model new nodes are added to the network first selecting the structure of a node already in the network, then changing it and finally adding it to the network. In order to study if the selection criteria had an influence on the generated network we run a variant of the SN model. In this variant new nodes are added all at once and their structure is a modification of the structure of the nodes initially present. So, the difference is that only the initial structures are used as template for the new added structures.

Network generated in this way showed a very low average clustering coefficient (around 0.002) and their degree distribution did not follow a power low (see supplementary material [11]).

From this we conclude that, in the SN model, the incremental addition of new nodes to the network (as opposed to the addition of the nodes all at once) is a necessary element in order to have a power low degree distribution and a high average clustering coefficient.

4 Reproduced networks

Different instances allow the SN model to generate networks with different topological features. In this section we describe how this model can generate networks having topological features similar to the ones of some empirical networks. We followed a trial-and-error approach in order to find the input parameters to 'fit the networks': manually testing different instances until a 'sufficiently good' result was found. We are confident that a different approach, either analytical

or based on heuristics (i.e., evolutionary algorithms [4]), could lead to better results.

All the (input and output) files related to the reproduced network described in the following are available from [11]. We successfully generated networks having strong similarities with the MRSA gene network (data not shown).

C. Elegans neural network. In [17] it is reported that the neural network of *C. Elegans* has 282 nodes (neurons), an average degree of 14, an average path length of 2.65 and an average clustering coefficient of 0.28. We were not able to retrieve a description of this network, so we only tried to match the just given network topological features.

Using the described network model we run tests having: {`A, T`} as alphabet, `ATATATATATAT` as structure of the only initial node, probability to mutate equal to 1, unit distance equal to 2 and nodes have a common edge only if their Hamming distance is smaller than 1. The generated networks with 282 nodes have (average on 100 tests): 13.94 as average degree, 3.61 as average path length and 0.3 as average clustering coefficient.

The 100 generated networks have a low variance on these values: 61% have at most 10% discrepancy from the average results (see supplementary material [11]).

E. Coli **protein-protein interaction network**. In [8] the protein-protein interaction (PPI) network of *E. Coli* was published. The biggest connected component of this network consist of 230 nodes, it has an average degree of 6.04, an average path length of 3.78, an average clustering coefficient of 0.22 and a heterogeneity index of 0.24.

Using the SN model we run tests having: {`A, T, C`} as alphabet, `ATCATCTCATCACT` as structure of the only initial node, probability to mutate equal to 0.4, probability to duplicate equal to 0.6, unit distance equal to 2, using the file matches considered to reproduce the MRSA gene network and nodes have a common edge only if their Hamming distance is smaller/equal than 1.

The generated networks with 230 nodes have (average over 100 tests): 6.03 as average degree, 3.85 as average path length, 0.47 as average clustering coefficient and 0.26 as heterogeneity index. The variance of these networks is rather big: only 3% have at most 10% dicsrepancy from the average results while only 24% have at most 20% discrepancy from the average results (see supplementary material [11]). In general, we noticed that the probability to duplicate increases the heterogeneity index of a network but also increases the variance of the generated networks.

The degree distribution of the given network follows a power low with trend $\gamma = -1.06$. The degree distribution of the networks generated by us also follow a power law but (average over 100 tests) with trend $\gamma = 0.72$. It is remarkable that the SN model is able to generate networks with a small number of nodes having a power law as degree distribution.

The motif distribution of the generated networks having at most 20% discrepancy from the average results is equal to the one of the give network. The

clustering coefficient distribution of the given network has a trend $\gamma = -0.52$ while similar trend for the generated networks having at most 20% discrepancy from the average results is $\gamma = -0.47$. The path length distribution of the given network and of one of the generated networks having at most 20% discrepancy from the average results is depicted in Fig. 3.

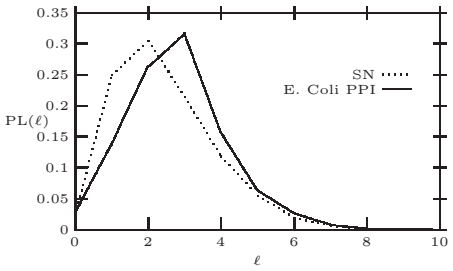

Figure 3: Path length distribution of the given *E. Coli* PPI network and a typical outcome of the SN model (SN).

5 Final remarks

In this section we give some thoughts about the SN model and we suggest possible directions for research on this model.

It seems that the SN model shows that preferential attachment is not necessary to generate networks having a power law degree distribution. Can it be the preferential attachment is somehow 'hidden' in the SN model? We think that preferential attachment is 'hidden' in the combination of structured nodes and Hamming distance. Anyhow, it is surprising that these two simple concepts (Hamming distance is rather simple when compared to the reasons behind the interactions in gene and protein networks) can behave as preferential attachment and, in some cases, (as for the average clustering coefficient being independent from the network size or the high heterogeneity index) be better in reproducing empirical networks.

The study on the SN model is still in its very early stages in order to allow us to say something new about biological networks. We do not think that the SN model can recreate all empirical networks or all features of some empirical network, anyhow, it is interesting to note that this model can recreate a broad range of topological features present in empirical networks of different nature. Of course, the big number of input parameters (and their domain) of the SN model allows to 'tune' some of the features of the generated networks more than what possible with other network models.

As we said, we used a 'trial-and-error' approach in order to find the instances of the SN model generating the networks considered by us. For some of these networks we got pretty close results, for others less good results. We did not

aim to have an exact match for the empirical networks considered by us, we aimed to show the broad range of topological features that can be matched by the SN model.

It is definitely interesting to study the classes of networks that can be generated by the SN model upon changes in its input parameters. Moreover, extensions to the model will allow it to generate directed networks, to evolve networks or to use the generated networks for other studies. For instance, one might need to have networks with a specific motif distribution in order to study their dynamics.

Acknowledgements We gratefully acknowledge Ian Overton for providing the MRSA network.

References

[1] R. Albert and A.-L. Barabàsi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[2] U. Alon. *An introduction to Systems Biology*. Chapman & Hall, 2006.

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

[4] K. Deb. *Multi-objective Optimization Using Evolutionary Algorithms*. Wiley-Blackwell, 2008.

[5] P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[6] E. Estrada. Quantifying network heterogeneity. Technical Report 25, University of Strathclyde, 2010. Available from `http://www.mathstat.strath.ac.uk/research/reports/2010`.

[7] E. Estrada. Randić index, irregularity and complex biomolecular networks. *Acta Chimica Slovenica*, 57:597–603, 2010.

[8] G. Butland et al. Interaction network containing conserved and essential protein complexes in excherichia coli. *Nature*, 433:531–537, 2005.

[9] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[10] B. H. Junker and F. Schreiber, editors. *Analysis of Biological Networks*. Wiley-Blackwell, 2008.

[11] Supplementary material. `http://www.macs.hw.ac.uk/~pier/download.html`.

[12] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303:1538–1542, 2004.

[13] 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, 2002.

[14] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002.

[15] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002.

[16] NWB Team. Network Workbench Tool. Indiana University and Northeastern University, 2006. `http://nwb.slis.indiana.edu`.

[17] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[18] Wikipedia. `http://www.wikipedia.org`.