How Scale-free Type-based Networks Emerge from Instance-based Dynamics

Tom Lenaerts, Hugues Bersini and Francisco C. Santos
IRIDIA, CP 194/6, Université Libre de Bruxelles, Avenue Franklin Roosevelt 50,
1050 Brussels, Belgium.
tlenaert@ulb.ac.be

Abstract

This article defines a growth model for complex networks based on the underlying dynamics reminiscent of biological systems. The model constructs a network consisting of vertices that correspond to biological types of elements and edges modelling the dynamical interactions that have occurred between instances of these types. It is shown that under endogenous birth-only dynamics a scale-free distribution can emerge with general properties similar to those in biological systems. We also show that, as in Biology, the growth model can produce networks that are organised in a hierarchical modular fashion.

Introduction

Due to recent advances in high-throughput methods that gather information of the overall structure of molecular networks, Biologists are now able to investigate the properties of these networks in terms of their structure and functionality (Alam and Arkin, 2003; Li and et al., 2003; Vidal, 2005; Uetz and Finley Jr., 2005). Yet, little is known about the mechanisms which are responsible for these properties. Consequently, models need to be designed which might make predictions about the results we can now observe using these high-throughput methods.

In statistical physics, general constructive models have been proposed to explain the origin of these biological networks of interactions (NoI) (Barabási and Albert, 1999; Dorogotsev et al., 2001; Newman, 2003; Pastor-Satorras and Vespignani, 2004; Dorogovtsev and Mendes, 2003; Solé et al., 2002; Strogatz, 2001). Yet, due to, on the one hand, the expected simplicity of the models and, on the other hand, the absence of any real-world biological semantics (Alam and Arkin, 2003), these models, like the preferential attachment model (Barabási and Albert, 1999), have suggested construction rules that are difficult to justify in a biological context. In acknowledgement of this problem, other mechanisms like the gene-duplication model (Vazquez et al., 2003; Wagner, 2003; Solé et al., 2002) were introduced to explain the structure of biological NoI like protein interaction networks.

Starting from a similar motivation as both previous models, we designed a generalised growth model for Biological NoI. The motivation for the introduction of this new model came from the shortcomings of the BA model in a biological context and the limited validity of the gene-duplication model to represent other biological NoI (Barabási and Oltvai, 2004). Moreover, since the behaviour of biological systems is defined by the interaction dynamics between the elements only a growing model that incorporates these interaction dynamics will be able to give a meaningful explanation. To achieve our goal, a growth model that uses four basic principles of biological simulations is defined. The basic principles that are examined are: (i) every node has a *different identity* based on its physical properties defining its *type* and an associated *concentration* that changes in time; (ii) since every node represents a type, biological networks are *type-based* networks instead of *instance-based* like social and technological networks; (iii) every node connects to a selected set of nodes based on mutual *attractiveness* (*affinity*) ; and (iv) which nodes will be added to the network depends on the *dynamics of the existing nodes* in the network. Especially the principles (ii), (iii) and (iv) are critical features that are completely ignored in other general growth models. When a cloning dynamics is used to introduce the new nodes into the network, the model comes close to the gene-duplication model. The difference is that attachment of the new node to the network is performed using the concentration of the biological types and an actual binding rule based on complementarity (or similarity) instead of the previous connectivity of the nodes (Vazquez et al., 2003; Wagner, 2003; Solé et al., 2002).

Using this generalised growth model we show here that the power-law or scale-free degree distribution with different γ can be obtained under certain conditions i.e. endogenous production of similar instances of the types that are present in the network and an amplification mechanism concerning the concentrations of the types. As soon as these conditions are relaxed, the network looses its scale-free properties and moves to the other extreme of the spectrum. This other extreme is obtained by an exogenous production model

that has similarities with the growth-only model defined in (Barabási et al., 1999). Furthermore, we demonstrate that this growth model produces NoI whose structure corresponds to the hierarchical and modular NoI discussed in biological literature (Newman, 2003; Ravasz et al., 2002; Rives and Galitski, 2003; Guimerà and Nunes Amaral, 2005). Such a structure may improve the overall robustness of the network is since perturbations will only have local effects.

The structure of the article is the following. First we provide an algorithmic description of the growth model and discuss, in the following section, a simplified birth dynamics that produces the nodes that might be recruited into the network. In the same section the results of this model are discussed. In a following section the structure of the NoI topology is discussed. Here we focus on the modular or hierarchical nature of the network and discuss how our results correspond to existing results in the literature. Finally we provide a conclusion.

Growing Biological Networks with Homogeneous Types

Based on the five basic ingredients that were discussed in the previous section, a growth model is defined as follows: Every type (vertex) in the network is identified by a binary string of length N so that only 2^N types are possible. Yet, since biological system consists of multiple interacting copies (instances) of the same type, whose abundance can change due to the underlying dynamics, every vertex also has a variable indicating the number of instances that are present (the concentration). The general reasoning is that for every interaction an instance has to exist. An interaction is defined by a *binding* process: two instances of particular types bind when their binary string is sufficiently complementary. Since we are using bitstrings to represent types, the hamming distance (DH) is used to calculate the difference. The resulting binding rule for an instance of type n_i to connect to another instance of type n_j is:

$$DH(n_i, n_j) > t \tag{1}$$

meaning that the Hamming distance (DH) has to be superior to a given threshold t. The value of t influences only one physical property of the NoI i.e. the maximal number of links any type in the NoI can possess. This idea of binding comes very natural in interactome modelling (Thomas et al., 2003; Maslov and Sneppen, 2002).

Because every node uses bit-strings of the same length N and the same threshold value t we refer to the types as *homogeneous*. This assumption can be relaxed since types in Biological system are not homogeneous. In this article, we discuss only this first scenario.

The general structure of the growth model with homogeneous types works as follows:

a) Initially a few types with initial concentration equal to one are recruited in the system in order to kick off the growing.

b) At each time step, do:

1. Produce a new instance x from the set of possible types T using an *instance-birth* dynamics (explained below).
2. Select a set of possible partners P of a predefined size $trials$ from the collection of types T in the current network relative to their concentrations.
3. Traverse P one by one and determine the affinity between the new instance x and the potential partner p from P. Examine the attractiveness between x and p using Equation 1.

(a) if a partner p is found then go to 5.

4. if no partner p was found go to 7.
5. If the type $t(x)$ associated with the new instance x is not yet in the network then add the type, set the concentration of $t(x)$ to 1 and connect the type $t(x)$ with the type $t(p)$ of the partner.
6. If the type $t(x)$ associated with new instance x is already in the network then increase the concentration with 1 and if the link between $t(x)$ and $t(p)$ is not yet present in the network then add it.
7. When the amount of types in the network reaches a predefined number, the simulation stops. Otherwise go back to 1.

This growth model incorporates a biological interpretation of the growth and preferential attachment rules proposed by Barabási and Albert: At each iteration new instances of a particular type are introduced to the NoI and they are added to the NoI when they can bind to other instances in the NoI. The link only appears in the network when the types of the two instances were not bound before. This part simulates the growth process of the model.

Important now is the distribution of types that are added and to which type the recruited instances are connected. In the algorithmic description, the distribution of types is determined by the *instance-birth* dynamics. This dynamics is a simplification since we only consider birth and no death of instances. A complete dynamical model would require death rate or a carrying capacity which determines the stable concentration of a particular type. This extension is left for future work. The fundamental differences with all previously suggested growth models is that preferential attachment does not depend on the current connectivity of the node in the NoI. As we will demonstrate, it is a function of the concentration of the types.

Before this growth model can be used, the value of a major parameter needs to be tuned: how many types will we select from the NoI to decide whether the new instance will be included ? Or what is the value of $trials$ (see step b2)?

Since, every instance in the models has the same probability of encountering the new instance, the partners are selected according to the concentrations of the types: a highly concentrated type has a higher probability to encounter the new instance. A small value for *trials* indicates only that there is a small chance that the node will find a partner. The bigger the value of trials the higher the possibility. As soon as the system encounters a possible partner (p) for this new instance, the type of the new instance can be added to the NoI. This assumption introduces a natural preferentiallity for types which are highly concentrated.

Further parameters are the size of the bit-string N, the amount of types that will be added to the NoI and the value of the threshold t. All experiments in this section are performed using $N = 13$, $t = 9$ and the amount of types is limited to 1000. This limit for the NoI size is selected since most of the experimental data of biological networks are of modest size in terms of number of types.

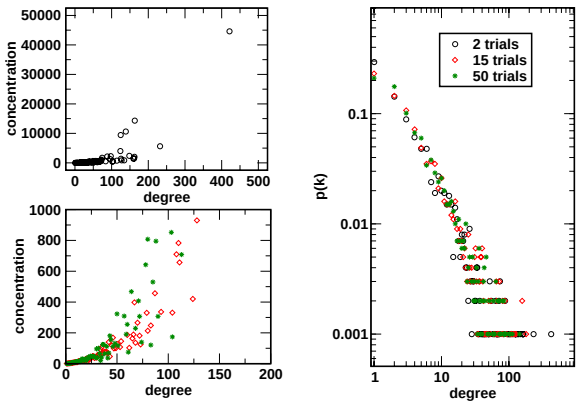

Figure 1: Results of producing a type NoI using cloning dynamics. The left plot shows the correlation between degree and type. The right plot shows the degree distribution of the NoI in terms of the number of types. $p_{mut} = 0.001$

Cloning Dynamics

Different instance dynamics which produce new instances and (possibly) types can be introduced in our growth model. Here, we discuss one production scheme: an instance of a type is cloned and its clone is possibly mutated. Mutation means that the bits of the string are switched from 0 to 1 or vice versa. Changing the bits has an effect on the attractiveness of the new instance towards other instances in the system. The probability of mutation is defined by a new parameter p_{mut}.

Every instance in the system has the same likelihood to produce a mutated clone: At every iteration of our growth model, a random instance is selected. Selecting a random instance corresponds to selecting a particular type relative to the concentrations of each type. Hence, types with high concentrations have a higher probability to produce the new clone than types with low concentrations.

The results of this node recruitment scheme are visualised in Figure 1 for different values of the parameter *trials*. As can be observed, the degree distribution becomes clearly scale-free for the case where $trials = 2$. For higher values of *trials*, the distribution shifts a bit, but remains close to the power-law distribution. These latter distributions are referred to as scale-free with exponential cut-off and are often observed in real biological data (Jeong et al., 2001). In the left-plot of Figure 1, it can also be seen that there exists a correlation between degree and concentration i.e. the higher the concentration, the higher the degree. This correlation between concentration and degree becomes very strong (polynomial) in the case where $trials = 2$. This difference can also be observed in Table 1. The high concentrations of the hubs are a consequence of the very few possibilities that cloned instances have to reconnect to the NoI. Important to note here is that the instance dynamics consists only of a birth process.

These results are a consequence of the endogenous production scheme and the amplification effect (or positive feedback) produced by the concentration dependency: All new instances are produced through cloning of an existing instance. When p_{mut} is small, this means that there is a high probability that the node which is produced, is an exact copy of the original one. In other words mutation did not change the bitstring. Since this allows a new instance to be added to the concentration of the highly concentrated node, the probability of selecting this type again for cloning increases. This explains the positive feedback mechanism: The more clones a type produces the higher the probability that it produces more clones in the future. Putting it in more popular terms, the rich get richer. This latter interpretation clearly reproduces the preferential attachment for this cloning model. Yet, and this is very important, to achieve these results the model did not make any assumptions about the degree of the type in the NoI. The time-dependent factor here is the concentration of the type which is implicitly linked to its degree.

The clonal model discussed here is of course strongly related to the gene-duplication model (Pastor-Satorras et al., 2003). In both cases it is some kind of cloning mechanism in combination with mutation which produces the new elements that are added to the NoI. The important contribution of the current model, next to providing an extension of the gene-duplication model, is that it highlights the importance of a positive feedback mechanism working on the concentration of the type to produce the power-law distribution in biological growth models. A similar amplification effect is also at work in the gene-duplication model: highly connected nodes will receive additional connections because

they were linked to nodes that are duplicated. By adding additional links the probability that they will receive even more connections is increased. The difference between the gene-duplication approach and our scheme is that we do not explicitly link this positive feedback to the degree of the node. Linking it to the concentration makes it more plausible from a biological perspective.

Apart from all that, the cloning model presented here also highlights the importance of the underlying dynamics in the formation of the NoI. Identifying this relationship with the dynamics is fundamental since it is an important source of new types that can be added to the NoI. It is therefore also an active research interest in the complex network community at the moment (Caldarelli et al., 2004).

Table 1: Physical properties of Networks produced by clonal recruitment. $< k >$ refers to the average degree, *max* refers to the maximum degree in the network, $< L >$ refers to the average path length, $< C >$ refers to the average clustering coefficient, r refers to the assortativeness of the network and γ refers the exponent of the power-law distribution that fits the data. The fifth column contains data on an experiment with a different binding rule: hamming distance smaller than a certain value t, instead of bigger than t.

	$trials = 2$	$trials = 15$	$trials = 50$	$t = 2$ and $trials = 2$
$< k >$	10.42	10.39	9.37	4.044
max	421	180	143	211
$< L >$	2.82	3.01	3.11	3.51
$< C >$	0.00765	0.00097	0.00582	0.1219
r	-0.33	-0.37	-0.36	-0.29
γ	2.7	2.4	2.6	2.3

Differences between the degree distributions shown in Figure 1 become more clear when we examine their properties listed in Table 1. On a first observation one can see that the value of *trials* has an impact on the final NoI. The average degree and max degree reduce when the number of trials increases. When comparing the scenario where $trials = 2$ with the other two, we can even conclude that this reduction is high for the value of the maximum degree. Since the degree reduces, an increase of the average path length can also be observed. In Table 1, we also added the γ values for the exponents of the power-law distribution that fits the data. It can be observed that the values are within the interval between 2 and 3, which contains the networks with the most interesting properties (Albert and Barabási, 2002; Newman, 2003).

In Figure 2, the results are shown for the second question: What happens to the NoI when the cloning process becomes more noisy? Increasing the mutation rate (value of p_{mut}) corresponds to an increase in the production of nodes that differ more from their parents, producing almost ran-

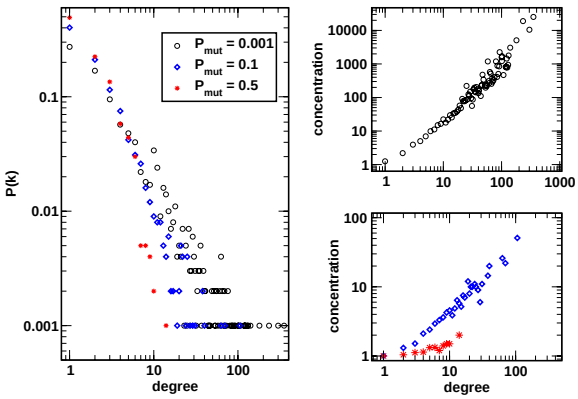

Figure 2: Results of producing a type NoI using clonal recruitment with increasing values of p_{mut}. The left plot shows the degree-distributions for three values of p_{mut}. The two right plots visualise the correlation between degree and concentration for each p_{mut} value.

dom new instances when $p_{mut} \approx 1$. The increase in mutation rate corresponds to an increase in the amount of random nodes that appear in the network. We call this source exogeneous. In the figure, one can observe that by an increasing the exogenous production, the degree distribution (left plot of Figure 2) looses its scale-free properties. This change is confirmed by the degree-concentration correlations shown in the two plots at the right in Figure 2: The top-right plot shows a polynomial relationship between degree and concentration reflecting the power-law distribution of the NoI produced for $p_{mut} = 0.01$. The bottom-right plot shows that when p_{mut} increases, the correlation becomes linear which results in a shift away from the scale-free distribution that we had before. When $p_{mut} = 0.5$, the correlation between degree an concentration becomes sub-linear indicating that we are coming close to the exponential distribution. In summary, increasing the exogenous production of nodes will remove all the hubs from the system resulting in a growth-only model.

Analysis of the Topological Organisation

One negative aspect of the first three columns in Table 1 is that the clustering coefficient is rather low. This is in contradiction with the data that have been obtained on biological networks (Newman, 2003). This problem in the current model is the consequence of the simple binding rule. Since the hamming distance has to be bigger than a particular threshold t, the mutated clone can not reconnect to the node it was produced by. This problem is not fundamental for our model and can be easily resolved by defining an inverse binding rule:

$$DH(n_i, n_j) < t \qquad (2)$$

which means that the hamming distance needs to be smaller than some value. Using this binding rule does not change anything in our discussion so far. The only difference is that now, the mutated clone can reconnect to the original one. In this way, *transitive relationships* can emerge and the clustering coefficient will increase since it counts the number of your neighbours that are also connected.

For instance, assume that in Equation 2 the value of t (threshold) is 2. This means that only nodes which differ in two 0 or 1 bit can connect with each other. When p_{mut} is very low, this will often be the case and we obtain a NoI with higher clustering coefficient. Results for an experiment with $trials = 2$ are visualized in Figure 3 and in the last column of Table 1.

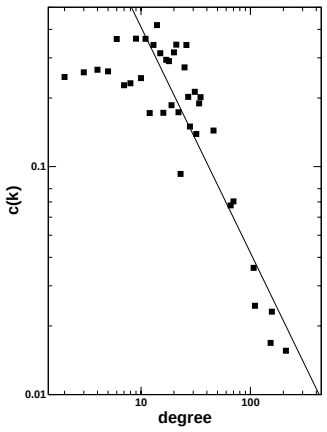

Figure 3: Results of producing a type NoI using cloning dynamics with $t = 2$. The plot shows the correlation between the clustering coefficient and the degree of the nodes of the network. The line corresponds to the scaling law $C(k) \approx k^{-1}$ which defines a law on the connectivity of the nodes in an hierarchical network. Plot shows the resulting degree distribution.

In Table 1 we can see that the average clustering coefficient has increased immensely, moving the resulting network closer realistic biological data. From Figure 3, one can derive that there exists a particular community structure in this experiment: the nodes with high degree have low clustering coefficient and the nodes with lower degree have a high clustering coefficient. This seems to indicate that the NoI has a modular structure where the hubs connect nearly decomposable sub-networks (Barabási and Oltvai, 2004). In the literature of biological networks equivalent plots have been shown for yeast protein networks (Rives and Galitski, 2003). Such a relationship between degree and the clustering coefficient provides a signature for modular structure. As already indicated in the introduction, this feature enhances the robustness of the NoI in terms of its functioning. As argued in (Maslov and Sneppen, 2002) this increase in robustness is a consequence of the lack of crosstalk between the different functioning modules in the NoI.

Furthermore, it has been suggested that biological networks are organised in a hierarchical structure, where nodes are organised in small modules which are in turn organised into larger modules (Rives and Galitski, 2003; Hallinan, 2004). In (Rives and Galitski, 2003) it is argued that this hierarchical modularity can be identified without identifying the actual modules using a scaling law for the connectivity of nodes in a hierarchical modular network:

$$C(k) \approx k^{-1} \tag{3}$$

with $C(k)$ being the cluster coefficient. As can be observed, the network produced by our model seems to come close to the suggested law, indicating that the network has a hierarchical modular structure. Note that this law does not provide any information on the form of the modularity. Hence further analysis needs to confirm the predicted outcome.

A final observation for all the previous cloning models is that the NoI are disassortative, a trait which is frequently encountered in the topological analysis of biological networks. Dissassortativeness refers to the fact that nodes with high degree are connected preferentially with nodes with low degree. The combination of modularity and the fact that the network is disassortative seems to indicate that it is not the hubs that connect the different modules in the NoI (Newman, 2003; Guimerà and Nunes Amaral, 2005). Whether this is the case also requires further analysis.

Conclusion

In this article a new growth model has been proposed which reflects naturally the kind of NoI one can find in real biological systems. The approach taken in the model is to combine general properties implicit in many biological simulations: (i) NoI consists of types which are defined by their structural properties; (ii) as a consequence biological NoI are type-based and this has some implications on the properties of the NoI; (iii) the links in the NoI are defined by complementary or matching relations and (iv) exogenous and endogenous mechanisms produce the new instances of existing or new types which are recruited into the NoI.

The experiments briefly discussed here show that a scale-free NoI is only produced in systems where new instances are recruited using an endogenous production scheme in combination with a positive feedback mechanism. When this scheme is combined with an exogenous production mechanism or the noise level in the recruitment mechanism increases, the scale-free property disappears leading, in the worst case, to an exponential distribution. Different from all previous growth models is that it is not the preference of connecting to nodes with higher degrees that leads to these

results. Type concentration determines both which types are endogenously reproduced and to whom this newly produces instance are connected. Hence, rich types are those which have a high concentrations and they will get richer when the new nodes are endogenously produced. This rich get richer phenomenon defines the positive feedback or amplification mechanism.

Finally, the growth model based on cloning dynamics (with certain binding rule) produces a NoI that has a modular and hierarchical organisation as can be observed in different biological NoI, confirming that our model produces NoI meaningful in a biological context. Yet, this feature of the model and its meaning require further investigation.

Acknowledgments

FCS acknowledges the support of *COMP*^{2}*SYS* a Marie Curie Early Stage Training Site, funded by the EC through the HRM activity.

References

Alam, E. and Arkin, A. (2003). Biological networks. *Current Opinion in Structural Biology*, 13:193–202.

Albert, R. and Barabási, L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys*, 74:47.

Barabási, L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

Barabási, L., Albert, R., and Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A*, 272:173–187.

Barabási, L. and Oltvai, Z. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5:101–113.

Caldarelli, G., Erzan, A., and Vespignani, A., editors (2004). *A virtual round table on ten leading questions for network research*, volume 38. Eur. Phys. J. B. special issue on Applications of Networks.

Dorogotsev, S. N., Mendes, J. F. F., and Samukhin, A. N. (2001). Size-dependent degree distribution of a scale-free growing network. *Phys. Rev. E*, 63:062101 1–4.

Dorogovtsev, D. and Mendes, J. (2003). *Evolution of networks: From biological nets to the Internet and WWW*. Oxford University press.

Guimerà, R. and Nunes Amaral, L. A. (2005). Functional cartography of complex metabolic networks. *Nature*, 433:895–900.

Hallinan, J. (2004). Gene duplication and hierarchicl modularity in intercellular interaction networks. *BioSystems*, 74:51–62.

Jeong, H., Mason, S., and Barabási, L. (2001). Lethality and centrality in protein networks. *Nature*, 41.

Li, S. and et al. (2003). A map of the interactome network of the meazoan c. elegans. *Science*, 303:540–543.

Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, 296:910–913.

Newman, M. (2003). The structure and function of complex networks. *SIAM*, 45:167–256.

Pastor-Satorras, R., Smith, E., and Sole, R. V. (2003). Evolving protein interaction networks through gene duplication. *Journal of Theor. Biol.*, 222:199–210.

Pastor-Satorras, R. and Vespignani, A. (2004). *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press.

Ravasz, E., Somera, A., Mongru, D., Oltvai, Z., and Barabasi, A. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555.

Rives, A. and Galitski, T. (2003). Modular organization of cellular networks. *PNAS*, 100(3):1128–1133.

Solé, R. V., Pastor-Satorras, R., Smith, E., and Kepler, T. (2002). A model of large-scale proteome evolution. *Adv. Complex Syst.*, 5:43–54.

Strogatz, S. (2001). Exploring complex networks. *Nature*, 410.

Thomas, A., Cannings, R., Monk, N. A. M., and Cannings, C. (2003). On the structure of protein interaction networks. *Biochem. Soc. Trans.*, 31:1491–1496.

Uetz, P. and Finley Jr., R. (2005). From protein networks to biological systems. *FEBS Letters*, 579:1821–1827.

Vazquez, F., Flamimi, A., Maritan, A., and Vespignani, A. (2003). Modeling of protein interaction networks. *ComplexUs*, 1:38–44.

Vidal, M. (2005). Interactome modelling. *FEBS Letters*, 579:1834–1838.

Wagner, A. (2003). How the global structure of protein interaction networks evolve. *Proc. R. Soc. Lond. B.*, 270:457–466.