IRIDIA

Université Libre de Bruxelles

Institut de Recherches Interdisciplinaires
et de Développements en Intelligence Artificielle

Growing Biological Networks: Beyond the Gene-Duplication Model

Hugues Bersini, Tom Lenaerts and Francisco C. Santos.

IRIDIA – Technical Report Series

Technical Report No.
TR/IRIDIA/2005-035

December 2005

IRIDIA – Technical Report Series
ISSN 1781-3794

Published by:

IRIDIA, *Institut de Recherches Interdisciplinaires
et de Développements en Intelligence Artificielle*
Université Libre de Bruxelles
Av F. D. Roosevelt 50, CP 194/6
1050 Bruxelles, Belgium

Technical report number TR/IRIDIA/2005-035

The information provided is the sole responsibility of the authors and does not necessarily reflect the opinion of the members of IRIDIA. The authors take full responsability for any copyright breaches that may result from publication of this paper in the IRIDIA – Technical Report Series. IRIDIA is not responsible for any use that might be made of data appearing in this publication.

Growing Biological Networks: Beyond the Gene-Duplication Model.

Hugues Bersini^{+1}, Tom Lenaerts^{+} and Francisco C. Santos^{+}.

$^{+}$IRIDIA, CP 194/6, Université Libre de Bruxelles, Avenue Franklin Roosevelt 50,1050 Brussels, Belgium.

Abstract

In this paper we propose a generalized growth model for biological interaction networks, including a set of biological features which have been inspired by a long tradition of simulations of immune system and chemical reaction networks. In our models we include characteristics such as the heterogeneity of biological nodes, the existence of natural hubs, the nodes binding by mutual affinity and the significance of type-based networks as compared with instance-based networks. Under these assumptions, we analyze the importance of the nodes concentration with respect to the selection of incoming nodes. We show that networks with fat-tailed degree distribution and highly-clustered structure naturally emerge in systems possessing certain properties: new instances need to be produced through an endogenous source and this source needs to provide a positive feedback favouring nodes with high concentration to receive new connections. Furthermore, we show that understanding the concentration dynamics of each node and the consequent correlation between connectivity and concentration is a more adequate way to capture the global properties of type-based biological networks.

Keywords: biological networks, type-based versus instance based networks, endogenous production, positive feedback, natural attractiveness.

[1] Author to whom correspondence should be addressed. email: bersini@ulb.ac.be

1 Introduction

It has been observed in Biology that in order to understand the functional properties of living systems, one needs to get an understanding of the topology of the underlying networks of interactions (NoI), the dynamics that take place on these NoI and the topological dynamics that have constructed them (Ito et al., 2000; Uetz, 2000; Ito et al., 2001; Li et al., 2004; Uetz and Finley Jr., 2005; Vidal, 2005; Berg et al., 2004). Questions concerning the global properties that organise these NoI are now asked at the level of the complete cell i.e. questions concerning the interactome. From a statistical physics perspective, it has been argued that these may be answered by studying the global physical properties of these NoI. One critical achievement is the observation that real-world biological NoI are not structured randomly. The data reveals a topology where few elements can interact with many other elements (a.k.a. hubs) whereas many others can only interact with very few (Uetz, 2000; Alam and Arkin, 2003; Barabási and Oltvai, 2004; Vazquez et al., 2003; Wagner and Fell, 2001; Wagner, 2003; Jeong et al., 2000). Frequently these NoI are referred to as scale-free networks (with or without exponential cut-off) where the degree distribution is a power-law. Moreover, general constructive models have been designed to explain the origin of these NoI (Barabási and Albert, 1999; Dorogovtsev et al., 2001; Newman, 2003; Pastor-Satorras and Vespignani, 2004; Dorogovtsev and Mendes, 2003; Solé et al., 2002; Strogatz, 2001).

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 have suggested construction rules that are difficult to justify in a biological context: The Barabási and Albert (BA) model[2] suggests a combination of growth and degree-preferential attachment to explain the origin of these scale-free NoI. Although these rules make sense in technological and social NoI (like the Internet or co-authorship networks), the preferential attachment rule poses serious difficulties in a biological

[2] The BA model for building scale-free graphs is made of two main steps: (i) at each time step a new node with m links is added to the network (*growth*) ; (ii) the probability p_i that a new vertex will be connected to a node i is $p_i = k_i / \sum ki$, k_i being the degree of node i (*preferential attachment*). The more partners a node has the more likely this same node will be the partner of a new node that enters the network. The application of this BA law during the growing of the network, instead of a pure random attachment law (where a new node would randomly connect with existing nodes), will give more chance to some nodes to acquire a larger connectivity, driving the distribution to a power-law decay for the number of nodes as a function of their number of partners (with an exponent -3) rather than an exponential decay produced by a random growing (Barabási et al., 1999).

context: How could a new biological node, discovering and observing *potential partners*, preferentially decide to connect with one of these on the basis of its connectivity? Although a human being can perform such a conscious choice while involved in the construction of any technological network or entering a sexual network, namely to express a certain preference for some nodes to attach to, this same preferential choice appears quite unlikely in natural systems growing with no human intervention since it requires high-level cognitive capacities. Moreover, most biological networks, even having a fat-tailed degree distribution, are not truly scale-free (Stumpf and Ingram, 2005; Tanaka et al., 2005). Furthermore, experimental data on biological networks should be treated carefully due to sampling effects and experimental constraints (Stumpf et al., 2005; Han et al., 2005). Having in mind these issues, the importance of the models discussed here is not only whether the resulting network is strictly scale-free or not. It should also provide all of other relevant properties .

In acknowledgement of the first problem, the biological realism of the preferential attachment, the gene-duplication model was introduced to explain the NoI of gene-regulation networks (Vazquez et al., 2003; Wagner, 2003; Solé et al., 2002; Chung et al., 2003). The basic assumption here is that the preferential attachment is a consequence of similarity between genes that produce the proteins and the initial topology of the NoI from which it is started: Proteins that have a high degree of interactions will have a higher probability to gain new incoming links since they are more likely to be connected to the protein whose gene is duplicated. The mechanism will amplify the degree differences in the initial NoI in a similar way as the Dorogovtsev-Mendes-Samukhin Model (DMS) proposed earlier in (Dorogovtsev et al., 2001) . Although the gene-duplication model has been shown to explain the scale-free structure for one particular biological area, there is no direct proof, as is argued by Barabási and Oltvai (2004), that this mechanism is the only one, or the one that explains the power laws in cellular networks. In (Berg et al., 2004) it was shown that link-dynamics occur at a much higher rate than gene-duplication, making it a more important force in the shaping of the NoI observed in the data. Furthermore, as shown in (Hallinan, 2004; Bhan et al., 2002) the clustering coefficient of the NoI produced by the gene-duplication model depends strongly on the initial seed network on which the duplication is performed.

Here another generalised growth model for Biological NoI is presented. The motivation for the introduction of this new model comes 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. The goal here is to move beyond the gene-duplication example by defining a model that encompasses a larger set of dynamic biological networks: Instead of focussing on one particular biological context, we incorporate a set of basic biological principles which lie at the foundation of an entire spectrum of dynamic models that have been constructed to gain understanding of different biological and chemical phenomena (for instance see Varela and Coutinho (1991)). The basic principles we examine 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*** network instead of *instance-based* like social and technological networks; (iii) every node connects to a selected set of nodes based on mutual *attractiveness* (*affinity*) ; (iv) certain nodes have intrinsically more ways to connect than others i.e. they are *natural hubs*; and (v) which nodes will be added to the network depends on the *dynamics of the existing nodes* in the network.

Starting from these principles, we design a generalised growth model where we show that the power-law 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). Consequently, the proposed model allows one to examine an entire spectrum of degree-distributions incorporating well-known principles from his or her favourite biological area. We believe that this model provides a richer way to discuss and evaluate biological networks than those have been proposed so far. Furthermore, the growth model discussed in this article is one of the first to rely on the underlying type dynamics to explain the topological evolution. This means that in terms of protein-protein interaction networks, we assume here that protein interactions do not occur at constant concentrations as for instance in (Berg et al., 2004). This choice was motivated by the fact that the NoI as a whole have a functional role which is influenced by the concentrations of the proteins it consists of. This will also make it harder to validate this work using biological data since high-throughput techniques like yeast two-hybrid which are used to determine protein interaction networks does not provide information on the in vivo behaviour of the proteins. For now, we only discuss a simple birth-only dynamics to create the NoI. This particular difference with existing growth models is important since it means that we do not put the emphasis on the static structure of the network; we are also interested in the underlying dynamical behaviour since we consider it to be imperative in the shaping of the NoI (Monk, 2003).

The structure of the article is the following. In the next section, the basic principles (i)-(iii) and (v) are discussed. The goal of this section is to provide a motivation for their importance and how they fit into our biological growth model. In a following section, the growth model is described and different node production schemes are discussed. Both the physical properties and the degree distributions are provided in this section. Afterward, we discuss in a separate section the importance of the concentration dynamics in the growth model. In the next section, we introduce principle (iv) and show the effect of this assumption on the properties of the network topology. Finally the article is ended with a discussion of the models.

2 Basic Ingredients of the Biological Growth Model

In this section, four of the five basic principles listed in the introduction are discussed in depth. In each section we explain why the particular principle is important for a biologically-oriented growth model. This explanation is followed by a set of growth models where we analyse the physical properties of the resulting networks. The fifth property of natural hubness is discussed in Section 5 in combination with the results produced by the associated growth model. The principles are mostly inspired by both protein-protein interaction networks and chemical reaction networks. Hence, in the following sections the different assumptions of this model will be illustrated with examples from these areas.

2.1 Defining Nodes: Types and Concentrations

In Biology, mean-field simulations are performed to attain an understanding of certain dynamical phenomena. This mean-field approach assumes that interactions between the instances in the model are equally likely. Since these instances can be clustered into different *types*, the likelihood of interaction between the types is determined by their *concentration* or frequency in the system. Although there might not be a direct causal relation, NoI of biological phenomena make a similar assumption : the nodes are types with particular physical properties and associated instance concentrations. This is different from technological and some social networks that have been analysed in the past. In the latter cases, every node corresponds to one particular computer or one particular person, whereas in for instance signal transduction networks, each node of the network is a particular protein type (see, among others, (Ito et al., 2000; Li et al., 2004; Uetz and Finley Jr., 2005)). We come back to this difference in Section 2.2, where we also briefly examine the implications of moving between these two levels of description.

Two further motivations for defining nodes as types are important. First, most experimental data on biological NoI produced by high-throughput techniques is type-based since the technique assumes that proteins appear at constant concentrations (Berg et al., 2004). Since no distinction is made between the proteins in terms of their concentrations nothing can be said about the individual interactions of the proteins. Given this fact, networks of individual protein interactions can not be derived and it is hence natural to study them at the type level. Second, many interactions between biochemical instances result in the disappearance of the original instances since they produce another instance (or instances) of another type (or types). Typical examples are chemical reactions between molecular species and complex formation in protein-protein interactions as in for instance signal transduction networks (Cho and Wolkenhauer, 2003). When an instance of protein A interacts with an instance of protein B through binding the instance will be part of a complex. As a concequence this same instance of protein A will not be able to interact with another instance of protein B or C. Hence, in such a case, all interactions are pair-wise and no structural analysis can be performed at that level. Yet the collection of all these interactions at the type-level may define a completely connected NoI whose physical properties can be studied. In that case, a type-based approach is more relevant than a instance-based approach. Both observations should also be reflected in the models that try to explain the different biological NoI.

Thus to create a growth model for a biological NoI, the resulting NoI has to be defined at the type-level instead of the instance-level. The interaction dynamics between, for instance the proteins, determines which new links appear in the functional NoI. Note that, since it is instances that appear in the system, concentration information about the types needs to be maintained to ensure the proper workings of the underlying dynamics.

The distinction between type-based and instance-based NoI has recently been touched in (Alam and Arkin, 2003). They argue that, as opposed to social and technological NoI, nodes, as in for instance metabolic networks, represent distinct chemical species. By making this observation they explicitly define the network at the level of the type and not the instance. They further state that highly linked nodes in biochemical networks like water and ADP/ATP may be understood in terms of the large number of hydrolysis and energy-utilisation reactions and the convenience of having the same component for particular functional tasks. Thus, the type will have a high connectivity as a result of its internal properties that makes it a good partner to react with. The nodes like water and ADP/ATP will be highly linked because they are needed frequently in different reactions and in order to perform these reactions both elements have to be present (or produced) in high concentration.

When one examines different biological disciplines where interactions play a role, these two observations make sense. Yet, how this distinction between instances and types may reflect on the properties of the NoI is not clear: Do any of the previous conclusions change when shifting between these two levels of description? A simple answer can be obtained if we take an instance-based NoI and map it onto a type-based NoI.

2.2 Instance Networks versus Type Networks

Suppose that a connected NoI exists at the instance level and that it has a certain scale-free degree distribution, what will happen to this distribution if it is mapped to the type-level? In other words, do switches in the level of description of the NoI of biological systems change something to the properties of the NoI? If this is the case, then a profound investigation is required into the growth models that have been proposed to explain certain degree distributions. Moreover, it requires then to explicitly define the growth models at the type-level. The motivation to go from instance to type as opposed to from type to instance is that the first is simpler. If one wants to do the same simulation from type to instance, the transformation can produces different results for the same type-network.

To get an understanding of mapping between instance and type-level, we performed a very simple experiment to investigate what happens when a scale-free instance-based NoI is transformed into a type-based NoI. Important in this simulation is that the collection of types that will be assigned to the instances can be sampled from different distributions. Here, the effects of using a uniform random and a Gaussian distribution of types are investigated. Yet, a bigger variety can be imagined. As discussed earlier, we assume that there is a connected NoI at the instance level. This assumption is purely made to simplify the discussion. As argued in the previous section, in biological systems, instances will be often consumed by the interaction.

The simple test performed here is related to other investigations on the effect of sampling in biological systems (Stumpf et al., 2005; Han et al., 2005). The importance of those analyses is that even when the real biological distribution is a scale-free, sampling this distribution and investigating the sampled distribution may produce different results. The difference with our simple experiment is that, in those sampling experiments, the instance-type distinction and consequently the distribution of type concentrations is not considered. Nevertheless, it is sufficient to say that it follows from those tests and the one performed here, that one has to be careful when discussing the networks of biological systems.

The results of the experiment are shown in Figure 1. The simulation starts from an instance-based NoI which has a scale-free degree distribution. It was produced using the well-known BA model. The scale-free distribution is visualised in the leftmost plot in Figure 1. In the middle plot of the same figure the resulting degree distributions are shown for both kinds of type-distributions. In both cases, the nodes of the network are traversed randomly and a type selected from the particular distribution is assigned to each node. The major result from this exercise is that when moving from instance to type the scale-free distribution disappears. Moreover, the resulting distribution depends clearly on the type-distribution that was used to assign types to the original nodes. This shows that one has to be careful when applying known properties from instance-based NoI to type-based NoI.

An interesting consequence of this experiment is shown in the rightmost plot of Figure 1. In both type NoI, there exists a correlation between the degree of the node and its concentration. This phenomenon is less clear when using uniform random assignment of types to the instance-based NoI than when using the Gaussian assignment. Nevertheless, nodes of higher degree tend to have a higher instance concentration. From a biological perspective, this is also what was expected: although at the type-level the nodes with high degree are rare, they have many instances because they interacted with this amount of instances of the other types. Consequently, this correlation has to be there.

The simple exercise discussed in this section indicates that a more thorough investigation of biological NoI as type-based NoI is required. Furthermore, the correlation between degree and concentration will be an important aspect in deciding the validity of the proposed growth models. Since, the type-distribution seems to play a significant role, we need to take a closer look at the mechanism that serves as the source of the instances in the growth models.

2.3 Defining Links: The Attractiveness of Types

Interactions in biological systems are restricted by the physical properties of the types in question. For instance, in protein-protein interactions, the domains on the proteins define the set of possible interactions i.e. domains bind to other domains to form complexes (Jones and Thornton, 1996). The structural properties of each domain on a protein determines which other proteins can bind to it: the domains of interacting proteins need to align in the correct way and some form of matching is required to bind them. This structural property of binding should form an intrinsic part of the types in the NoI. We will refer to the potential of types to connect to other types as their *attractiveness*.

Since, structural differences are traditionally modelled using strings of elements from a particular alphabet, the same approach will be taken here. Each type has an associated bit-string that is used to decide whether the instances of some other type can connect to it. If the difference between these structural properties of the types is above some threshold, then they are sufficiently complementary and can connect. This interpretation immediately maps on what is known about protein-protein interactions. As discussed by Jones and Thornton (1996), there is a certain fundamental characteristics to protein-protein interfaces. Two interfaces may bind but one has to know the binding strength in order to say something about the relation. In (Jones and Thornton, 1996), the change in the solvent accessible surface area (ΔASA) is used. Here we define a threshold value that will provide information on the binding strength. The complementarity of the protein-protein interface is another important issue addressed in (Jones and Thornton, 1996). There, the complementarity is determined using a gap index which is the ratio between the volume of the gap that is enclosed by the two molecules and the accessible surface area. Since bit-strings are used here to represent the interface structure, the hamming-distance function will be used to calculate this property. A similar key-lock issue can be observed in for instance immunological models (De Boer and Perelson, 1994; Detours et al., 1994; Varela and Coutinho, 1991). Although not all biological systems require this kind of binding, it is always the structural issues that decide whether the interaction occurs or not. Therefore we consider this approach general enough to model an entire spectrum of biological NoI.

By defining links as an attractiveness relationship, we introduce two assumptions. First, the binding occurs between instances of the types. This means that in the growth model, new links in the NoI will only appear when (i) the newly introduced instance represents a new type and can connect to an instance of an existing type or (ii) when the new instance belongs to a type that is already in the network but it connects to an instance of another type which was not yet bound to the first one. Second, as in protein networks, one instance can bind to multiple other instances. Which elements can bind is determined by the available domains i.e. the bit-string in our case. So we do not consider here the case where there is a one-to-one mapping between instances. The major motivation for this latter assumption is simplicity: If only pair-wise connections can be made at the instance-level, the network at the type-level can be disconnected, which complicates things a bit. Thus, to keep things simple we assume that instances can bind to different other instances simultaneously.

2.4 Node Dynamics: Source of Instances

In a dynamical system, one assumes a collection of elements and the rules that transform them. In such a system, all elements are produced from what is known about the state of the system. No elements appear that can not be explained by doing an analysis of the internal dynamics. We refer to such a system as an *endogeneous* system. At the other extreme, the internal dynamics produce nothing and all new elements are inserted from the outside. This is called an *exogenous* system.

This difference between endogenous and exogenous production is important since the types created by the internal system depend strongly on the constructive interaction rules of the biological system that is investigated. On the one hand, in for instance chemical or protein interaction systems, a reactive interaction between two instances may produce one or more new instances which can be similar or completely different from the original ones. On the other hand, in for instance immune models, randomly constructed antibodies are continuously added to the the immune system from the outside. What is the effect of these different sources on the topology of the NoI?

Additionally, endogenous approaches make assumptions about the probability of encounters between the instances in the model: The likelihood of interaction will determine whether two instances meet and one or more new instances are produced. Does this mean that an underlying concentration-based dynamics is required to produce realistic biological topologies? It was already observed in the previous section that a correlation exists between concentration and degree: the higher the degree, the higher the concentration. Can this be extended to scale-free networks? Do the hubs correspond to the nodes with the highest concentration? This feature was left unstudied in all previous growth models like the BA model, DMS model or the gene-duplication model. Yet, in our opinion, it is an essential component that brings these growth models closer to biological reality.

In light of the current questions, we will discuss three approaches to the node dynamics. The first approach assumes a uniform random introduction of nodes as in immune networks: It is an outside source that introduces new instances. This scheme will be called the *random model*. The second approach assumes that instances are introduced as mutated clones of the existing nodes in the network. This model links our work to the gene-duplication model but extends it with the basic principles we consider necessary for a biological growth model. We refer to this scheme as the *cloning model*. The final approach assumes that two instances collide producing one new instance. This final model draws its inspiration from bimolecular reactions in chemistry, genetic crossover or the binding of different proteins in for instance signalling networks. This approach is the *collision model*. Other approaches or combinations of approaches can be possible. For now we limit ourselves to these three.

These schemes were selected because they represent orthogonal approaches for producing new instances: exogenous versus endogenous node production, concentration dependent versus concentration independent production and similar versus dissimilar type production. The random model belongs to the class of schemes that produce nodes exogenously without taking into account the current concentrations of the types in the network. This means that the current state of the system does not determine what the next instance will be that is added to the system. In the cloning and collision schemes the opposite is true: the current distribution of instances determines what will be produced next. The difference between the cloning and the collision model is that the first will produce types that are very similar to the original one (if p_{mut} is low). The second scheme makes a combination of the two types and produces one new type which is related to the original ones but can be completely dissimilar. This is done by splitting the two bit-strings of the original types and combining two parts from them to create a new bit-string and hence a new type.

3 Growing Biological Networks with Homogeneous Types

Given the four basic ingredients that were discussed in the previous sections, a set of growth models using these properties is evaluated here. Every type in the network is identified by a binary string of length N so that only 2^N types are possible. As already indicated in Section 2.3, instances of a type n_i will connect with other instances of other types on the basis of their complementarity. Since we are using bit-strings, 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. We will demonstrate this in one of the following examples.

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 will be relaxed in Section 5 since types in Biological system are not homogeneous. For now, we show what can be observed in 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 one of the three birth-dynamics discussed in Section 2.4.
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 their type is added when it can connect to another type in the NoI. This part simulates the growth process of the model. Important now is the distribution of types that are added (see the discussion in Section 2.2) and to which type the recruited instances are connected. The three schemes we mentioned in Section 2.4 will each provide a different mechanism that produces the distribution of types. The fundamental difference with all previously suggested growth models is that node 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.

Since the growth model introduces new instances of a type at each iteration, the number of instances that are actually presented and added can be much higher. In Figure 2 one can observe three plots showing the speed with which new types appear in the NoI until the limit is reached for all three production schemes.

From this figure we can derive that in case of the random introduction of new nodes, the amount of types almost grows linearly with the amount of instances introduced. The collision model requires a bit more time to produce 1000 different types. The delay is a consequence of the concentration dependency in the production rule of the collision model: two instances are selected and produce an new instance that is a random combination of them. Since the new type is a random combination of the old types, new types appear still at a more or less regular basis but not as quickly as in the random model. In the cloning model, the delay is much larger. The motivation for this is similar to the collision model. Yet, the delay is much larger because newly produced instances are still very similar (or even equal) to the original type that produced them. Consequently, much more time is required to introduce 1000 different types. Of course in this last case, speed can be increased by increasing p_{mut}. Yet, this extra amount of noise will have its consequences on the degree distribution (see Section 4). This difference in convergence speed will also have its consequences on the concentrations of the types in the different production schemes. It is expected that for low mutation rates the cloning model will produce higher concentrations than the random or collision model.

3.1 The Random Model

In the Random model, new instances are introduced from the outside in a uniform random fashion. We referred to this as the exogenous production of nodes without taking into account the concentrations of the types already present in the NoI. In Figure 3 both the resulting degree distribution and the correlation between degree and concentration are visualised. As can be seen in the right plot, their are no hubs and the degree distribution follows an exponential decay. Three distributions are shown for different values of the parameter *trials*. The exact value of this parameter does not seem to have an important effect in the current model. The resulting degree distribution is not surprising since the growth model with uniform random recruitment corresponds to the BA model without preferential attachment (Barabási et al., 1999). In fact, no node is favoured during the attachment of any new node since the concentration of all nodes in the NoI approximately remains the same (left plot Figure 3). In Table 1 the physical properties of these generated networks are listed

Table 1: Physical properties of Networks with random 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 and r refers to the assortativeness of the network.

	$trials = 2$	$trials = 15$	$trials = 50$
$< k >$	2.132	2.144	2.14
max	10	12	10
$< L >$	9.17	9.24	9.91
$< C >$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
r	-0.04484	-0.01173	-0.08078

The similarity between the networks for different values of *trials* can again be observed in Table 1. They all have similar average and maximum degrees, similar average path length and average clustering coefficient and finally they have similar assortativeness. The important thing learned from this model is that connecting with higher probability to types with higher concentration does not produce long-tailed degree distributions and, consequently, no hub pops up.

3.2 The Cloning Model

Although the random recruitment of new instances is plausible in for instance the area of immune networks (Varela and Coutinho, 1991; De Boer and Perelson, 1994; Detours et al., 1994), different mechanisms are active in other biological models. As argued in Section 2.4, the dynamics that exist between the instances in the biological model can also be an important source of new instances of new and existing types. In this section, we examine a production scheme where an instance of a type is cloned and its clone is possibly mutated. Here 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, an instance is randomly 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 4 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 4, 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 2. The high concentrations of the hubs are a consequence of the very few possibilities that cloned instances have to reconnect to the NoI.

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 bit-string. 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 (in concentration) get richer. 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. As a matter of fact, this aspect relates with the notion of ***natural hubness*** that will be discussed in Section 5 since it is the concentration that produces the high degree instead of the degree itself. A highly concentrated node turns out to be a natural hub.

The clonal model discussed here is of course strongly related to the gene-duplication model (Bhan et al., 2002; Pastor-Satorras et al., 2003; Chung 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 society at the moment (Caldarelli et al., 2004).

Table 2: 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 4 become more clear when we examine their properties listed in Table 2. Different from the table in Section 3.1 is that now 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. The motivation for the reduction in the maximum degree is that the more trials you introduce to find a partner, the less the concentration plays a role in determining the hubness of the nodes. In Table 2, 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 is claimed to contain the networks with the most interesting properties (Albert and Barabási, 2002; Newman, 2003).

One negative aspect of the table 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 is was produced by. This problem is not fundamental and can be easily resolved by defining an alternative binding rule:

$$DH(n_i, n_j) < t \tag{2}$$

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 will be that now, the mutated clone can reconnect to the original one. In this way, transitive relationships can emerge and the clustering coefficient will increase. 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 5 and in the fifth column of Table 2.

From the right plot of Figure 5, one can also derive that there exists a particular community structure in this experiment. This indicates that the NoI has a modular structure (Barabási and Oltvai, 2004). 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. The combination of this 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; Guimera and Amaral, 2005). We will come back to this in the discussion section.

3.3 The Collision Model

As argued in Section 2.4, the cloning model presents only one possible mechanism for endogenous instance production based on the concentrations of the types in the network. The production mechanism does not necessarily have to produce types similar to the original one. The new types can easily be dissimilar as in for instance protein-protein interactions. The Collision Model represents this alternative idea: two instances of (possibly) different types are randomly selected from the NoI relative to their concentrations. From each instance a part is randomly selected from the bit-string and two parts are combined into a new instance with bit-string of the same length (N). The growth model is then used as before to add this newly created instance to the network (or not).

The results of this experiment for different values of $trials$ are visualised in Figure 6. A first observation is that the degree distribution is again (close to) a power-law distribution. A second observation is that again the increase in $trials$ results in a decrease in the degree of the hubs. Thus, as in the cloning model, the endogenous source for instances in combination with the preferential selection of highly concentrated nodes produces the distribution. But when comparing the actual values of the maximum degree with those in table 2 we can see that for the collision model the values are much lower. Even worse, in case of $trials = 50$, the maximum degree is only 44 which indicates that the hubs gradually disappear as we try harder to find a partner for a newly recruited instance (as before in the cloning model). Another effect of increasing the value of $trials$ is that the γ increases. For the current values, the γ is still within the interesting interval.

An additional difference from the cloning model is that now for $trials = 2$ the concentration of the hubs (number of instances) is not so excessive: When examining Figure 6 and comparing it to the degree-concentration correlation of Figure 4, one can conclude that there is a difference in final concentration values between both models for similar values of the parameter $trials$. The motivation for this difference can be explained using Figure 2. In the collision model, the required 1000 types are recruited faster than in the cloning model. As a consequence, the types in the collision model have less time to increase their concentrations before the experiment is finished. Yet the correlation between degree and concentration remains. Even-more, as in the cloning model, to obtain the power-law distribution, this correlation needs to be some kind of polynomial distribution as before (top left plot Figure 6). If it becomes linear and sub-linear (see bottom left plot Figure 6), the scale of the hubs decreases and the distributions becomes scale-free with an exponential cut-off.

In general, the current production scheme shows that creating new types which are dissimilar from the original one does not necessarily produce a strict linear power-law distribution in the log-log plot. Only when a positive feedback is present which can produce a polynomial correlation between degree and concentration, will we observe this strict linear relationship. The same observation was made for the cloning model.

Table 3: Physical properties of Networks produced by collisions. $< k >$ 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 data.

	$trials = 2$	$trials = 15$	$trials = 50$
$< k >$	7.406	4.966	3.792
max	112	71	44
$< L >$	3.6	4.14	4.74
$< C >$	$< 10^{-5}$	0.00019	0.00117
r	-0.16054	-0.05080	-0.04649
γ	2.4	2.8	2.9

4 The Influence of the Source of Instances

In the previous section, the growth model and its variations were explained and their results were discussed. In extension to this previous discussion, the importance of concentration-dependence and the mechanism of type production are examined more closely. First, one can wonder whether concentration-dependence is that important in the cloning model since the gene-duplication model easily produces a similar power-law distribution without taking concentration into account: Do we need to select a type relative to its concentration or is random selection of types enough to produce the same results? If this would be the case, the polynomial degree-concentration correlation looses its relevance as an explanation for the power-law distributions that can be observed in the cloning (and collision) model. Second, in biological systems there is no strict separation between exogenous and endogenous node production. What will happen if the cloning model is combined with the random model? Moreover, what ratio of endogenous versus exogenous production will remove the scale-freeness of the resulting degree-distribution? These two problems are discussed here.

In Figure 7, the results of an experiment useful to solve the first problem are shown. If the type is not selected relative to its concentration, then the cloning model will not produce the power-law distribution. The left plot in that figure shows the results when the type is selected randomly. This result clearly shows that without concentration-dependent selection and the positive feedback mechanisms, cloning does not produce the same results as the gene-duplication model. This implies that both the type-concentration and the degree-concentration correlations play a crucial role in explaining the observed degree distributions.

In Figure 8, the results are shown for the second problem: What happens when a mixture of endogenous and exogenous production is used to recruit new types? Increasing the mutation rate (value of p_{mut}) corresponds to an increase in the exogenous production of nodes since higher mutation rates alter the structure of cloned bit-string faster, producing almost random new instances. In the figure, it is observed that by an increase in exogenous production, the degree distribution (left plot of Figure 8) looses its scale-free properties. This change is confirmed by the degree-concentration correlations shown in the two plots at the right in Figure 8: 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 that we observed in Figure 3. In summary, increasing the exogenous production of nodes will remove all the hubs from the system resulting in a growth-only model. The same observation can be made for the collision model.

5 Growing Biological Networks with Heterogeneous Types

So far, the assumption was made that every type has similar properties: a bit-string of length N and a common threshold t. This homogenous-type scenario is not realistic from a biological perspective (Caldarelli et al., 2002). As argued in the introduction, some types are born to be hub because of their intrinsic structural properties. In this section, this notion of natural hubness and its effect on the growth models is discussed. Before performing the actual experiments, lets explore this issue a bit.

5.1 Natural Hubs: Heterogeneity in Type Properties

As argued in Section 2.3, every type has its particular physical properties that decides with whom they can interact. So far it was assumed that this is the same for every type i.e. every node uses the same threshold. From a biological perspective, this does not make much sense: For instance, molecules differ in their *reactiveness* when colliding with other molecules. Proteins have differences in the number and kind of domains which are used to bind other proteins. Hence, some types, due to their intrinsic properties, are more attractive than others. Due to these differences in attractiveness, some types are naturally inclined to become hubs of the network. Hence the hubness is not only a result of the growing history of the network (the degree) but is a consequence of nodes intrinsic properties. We refer to these types as *natural hubs*.

This idea of natural hubness has been discussed before. Alam and Arkin (2003) argue that highly linked nodes in biochemical networks like water and ADP/ATP may be understood in terms of the large number of hydrolysis and energy-utilisation reactions and the convenience of having the same component for particular functional tasks. Thus, the type will have a high connectivity as a result of its internal properties that makes it a good partner to react with.

To introduce this heterogeneity in types an additional attribute called *attractiveness threshold* (AT_i) that takes a value in the interval [1,N-1] is added to the type. This means that each type is now defined by the bit-string of size N and the corresponding attractiveness threshold. Thus instances with the same bit-string but different values for AT_i are considered to belong to different types.

This new extension to the previous model requires a new binding rule that takes into account the attractiveness threshold. Two nodes can bind if and only if:

$$DH(n_i, n_j) > min(AT_i, AT_j) \tag{3}$$

where n_i and n_j refer to the bit-strings that represent the structural properties of the types and the attractiveness thresholds (AT_i and AT_j) define the natural hubness of each type: A node with a low attractiveness threshold has much more possibilities to connect than a node characterised by the same binary string but with a higher threshold. We take the minimum of the two attractiveness values since a natural hub should have a higher probability of connecting to the presented partners. So when an instance of a type with $AT = 10$ encounters another instance with the same attractiveness threshold, the two instances have to be different in 10 bits. Yet if a natural hub instance with $AT = 1$ is recruited and its partner has an $AT = 10$, then the instance only has to differ in one bit to connect to this element. It is clear that in this case, all nodes with low attractiveness threshold will have a higher probability to connect to elements in the network.

Note that if the binding rule expressed by Equation 3 is replaced by

$$DH(n_i, n_j) < min(AT_i, AT_j) \tag{4}$$

one again obtains a system where the newly produced instances can connect to their origin as discussed in Section 3.2.

Given this new binding rule, how does this change the previous experiments and the corresponding results? Will hubs appear in the random model? Will the natural hub also become the highly concentrated node of the NoI? These are only a few question that can be asked and which we try to answer using the experiments discussed in the following section.

5.2 Growing Networks of Heterogeneous Types

As before, three modes of instance production are considered in this section: the uniform random production, production by cloning and production by collisions. The difference with previous models is that now, for every instance, we also need to determine a value for AT_i. This means that in the random model, next to the random bit-string, we also randomly select a value from the interval [1,N-1]; in the cloning model, the bit-string is cloned as before and the AT_i value of the original instance is modified to another value in the interval using the same probability p_{mut}; finally in the collision model, the bit-strings of each instance or recombined as before and the value of AT_i is determined by taking a value in between the two original ones.

Given these setups, the resulting degree distributions and degree-concentration correlations are shown in Figure 9 and Figure 10. All plots are shown here for $trials = 2$, $N = 13$ and for a network of 1000 types.

As can be seen in those figures, the results of these experiments are equivalent to those produced in the previous section. A slight difference can be observed for the collision model where the distribution is no longer scale-free. The reason for this is that in the new definition of the collision model, the natural hubs might not always be present because they were not created by the endogenous production. Moreover, since in this model a type is defined by a bit-string and a threshold value, the probability of generating instances of a type already present in the network is smaller. This feature is responsible for the increasing of the randomness of the resultant distribution of connectivities.

More importantly for the discussion on natural hubs is that we expect the natural hubs to be the nodes with the highest concentration and highest degree. Since these nodes were born to play this role in the NoI, the topological dynamics should also assign this role to them. In order to verify this hypothesis, we examined the correlation between the values of AT for each degree. Note that since multiple values exists for every degree, the averages are plotted. The results are shown in Figure 11.

In the center plot of Figure 11, the results for a randomly created NoI are shown. As can be seen there, there is a nice correlation between degree and the attractiveness threshold: the higher the degree, the lower the value of the attractiveness threshold. This indicates that it is the natural hubs that have the highest degree in the produced NoI. Note though, that these types are not really hubs in this random model since they have a relative low degree.

When examining the two other plots for the cloning and collision model we see that things are not as clear as in the random model. The noise on these plots is a consequence of the concentration dynamics that plays a role in both models. In this case, the effects of concentration and natural hubness compete between each other. Although, it is clear from figure 11 that the nodes with highest connectivity, beside having highest concentration, are nodes with low threshold. Moreover, in the cloning model, one can observe that for low degrees the values of the threshold are centered around the average. This is logical since most types will have a low degree (see degree distribution in Figure 9) and these types can have all possible attractiveness values. As we move to higher degree nodes in the cloning model, the relation with the low attractiveness values becomes clearer: there are no attractiveness values bigger than 6 for nodes with degree higher than 300. For the collision model similar things can be observed: there is a tendency toward lower attractiveness values when the types have a higher degree. Yet some outliers can be observed.

Given these results, it becomes clear that when a type is constructed with the intrinsic properties to play the role of hub in the NoI, the dynamics will also assign this role to it.

6 Discussion

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 here 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; (iv) certain types are born to be hubs and (v) exogenous and endogenous mechanism produce the new instances of existing or new types which are recruited into the NoI. This approach was introduced since the BA model defines rules which are biologically not completely realistic and the gene-duplication model focuses only on one particular biological area. Furthermore, it is not clear whether the latter model is actually responsible for the observed effect (Barabási and Oltvai, 2004).

To validate the model a set of experiments were performed using different node production schemes: a random model, a cloning model and a collision model. The first model assumes that new instances of the possible types are introduced from an outside source in a uniform random manner. The second model produces new instances from highly concentrated types using cloning and mutation. Finally the third model produces new instance by combining the structural information of two highly concentrated types. All three models provide different, sometimes orthogonal, approaches to recruit nodes into the NoI. As was discussed, the cloning model uses a similar production system as the gene-duplication model. Nevertheless, the current model goes beyond the domain for which the gene-duplication model was defined. One of the differences being that here it is the underlying dynamics that guide the NoI's growth process.

The experiments 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.

The model discussed fits into the collection of other models that have been constructed to investigate biological NoI. The most well-known is the gene-duplication model. The work by (Rzhetsky and Gomez, 2001) comes even closer to the work described here. In their work, a model of metabolic networks is proposed based on bindings between domains. The important contribution of that work is that it provides the prediction that the frequencies of distinct DNA and protein domains is also a fat-tailed distribution.

A further result is that the correlation between degree and concentration may provide a signature to distinguish between scale-free and other kinds of NoI. In case of a polynomial correlation the resulting degree distribution is certainly scale-free. An almost linear correlation indicates either that their is an exponential cut-off on the scale-free NoI or that the we are in between a scale-free and an exponential NoI. A smaller than linear or sub-linear relationship is a signature for the gradual disappearance of any kind of hubs. An important question in relation to biological NoI is whether this relation between degree and concentration can also be found. Here we want to stress that this will probably not be the case since in realistic contexts limits exist on the amount of instances that will be produced. For instance, the proteins in the proteome regulate their production by reducing the production when concentration becomes to high. In this way different switching functionalities as in the p53 system are produced Vogelstein et al. (2000). Currently we are extending the current models to a dynamics where death of instances compensates the birth-dynamics and which come closer to realistic examples.

In relation to biological NoI, obtaining a real scale-free distribution is not the main feature since their are mostly only close to scale-free and have an exponential cut-off due to finite size effects. Other features like clustering coefficient, modularity and hierarchy play also an important role. The presented model produces the expected fat-tail distributions which can have high-clustering coefficients.

The latter depends on the binding rule, i.e. whether transitive relations are positive or not and not on the initial seed of the model as is for instance the case in the gene-duplication model (Hallinan, 2004; Bhan et al., 2002). In Table 2, we saw that the average clustering coefficient increased immensely, when the binding rule was changed, moving the resulting network closer to the presence of modularity in realistic biological data. Furthermore, from Figure 5, we 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. 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.

Also, 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:

$$C(k) \approx k^{-1} \tag{5}$$

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. Further analysis needs to confirm the predicted outcome.

Finally, we showed that, natural hubs, when recruited into the NoI, will assume their role as hubs in the final topology. This means that types which possess the structural properties to turn into a hub, will become the actual hubs of the network. This feature is important since it means that in existing networks, a naturally born hub will be able to take up its position in the distribution of the network.

Although the model uses a very simple birth-only dynamics, we feel that this approach of incorporating instance dynamics with network formation will play a crucial role in finding a useful explanation for the origin, dynamics and stability of the observed NoI in biological systems.

7 Acknowledgments

We want to thank the anonymous reviewers for their patience and excellent comments on how to improve the quality of the article. FCS acknowledges the support of $COMP^2SYS$ 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.

Berg, J., Lässig, M., and Wagner, A. (2004). Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evolutionary Biology*, 4:51.

Bhan, A., Galas, D., and Dewey, T. (2002). A duplication growth model of gene expression networks. *Bioinformatics*, 18(11):1486–1493.

Caldarelli, G., Capocci, A., De Los Rios, P., and Munoz, M. A. (2002). Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.*, 89:258702.

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.

Cho, K. and Wolkenhauer, O. (2003). Analysis and modelling of signal transduction pathways in systems biology. *Biochem. Soc. Trans.*, 31(6):1503–1509.

Chung, F., Lu, L., Dewey, T., and Galas, D. (2003). Duplication models for biological networks. *Journal of Computational Biology*, 10(5):677–688.

De Boer, R. and Perelson, A. (1994). Size and connectivity as emergent properties of a developing immune network. *Journal of Theor. Biol.*, 149:381–424.

Detours, V., Bersini, H., and Stewart, J. (1994). Development of an idiotypic network in shape space. *Journal of Theor. Biol.*, 170:401–404.

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

Dorogovtsev, 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.

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

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

Han, J.-D., Dupuy, D., Bertin, N., Cusick, M., and Vidal, M. (2005). Effects of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, 23(7):839–844.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *PNAS*, 98(8):4569–4574.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. (2000). Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci.*, 97(3):1143–1147.

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

Jeong, H., Tombor, B., Albert, R., and Barabási, L. (2000). The large-scale organisation of metabolic networks. *Nature*, 407:651–654.

Jones, S. and Thornton, J. (1996). Principles of protein-protein interactions. *Proc. Natl. Acad. Sci.*, 93:13–20.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J.-F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., van den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. (2004). A Map of the Interactome Network of the Metazoan C. elegans. *Science*, 303:540–544.

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

Monk, N. (2003). Unravelling nature's networks. *Biochem. Soc. Trans.*, 31(6):1457–1461.

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.

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

Rzhetsky, A. and Gomez, S. (2001). Birth of scale-free molecular networks and the number of distinct dna and protein domains per genome. *Bioinformatics*, 17(10):988–996.

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.

Stumpf, M. and Ingram, P. (2005). Probability models for degree disributions of protein interaction networks. *Europhysics Letters*, 71(1):152–158.

Stumpf, M., Wiuf, C., and May, R. (2005). Subnets of scale-free networks are not scale-free: Sampling properties of networks. *PNAS*, 102(12):4221–4224.

Tanaka, R., Yi, T.-M., and Doyle, J. (2005). Some protein interaction data do not exhibit power law statistics. *FEBS Letters*, 579:5140–5144.

Uetz, P. (2000). A comprehensive analysis of protein-protein interactions in saccharomes cerevisiaie. *Nature*, 403:159–166.

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

Varela, F. and Coutinho, A. (1991). Second generation of immune network. *Imunology today*, 12(5):159–166.

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.

Vogelstein, B., Lane, D., and Levine:, A. J. (2000). Surfing the p53 network. *Nature*, 408:307–310.

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

Wagner, A. and Fell, D. A. (2001). The small world inside large metabolic networks. *Proc. R. Soc. Lond. B.*, 268:1803–1810.

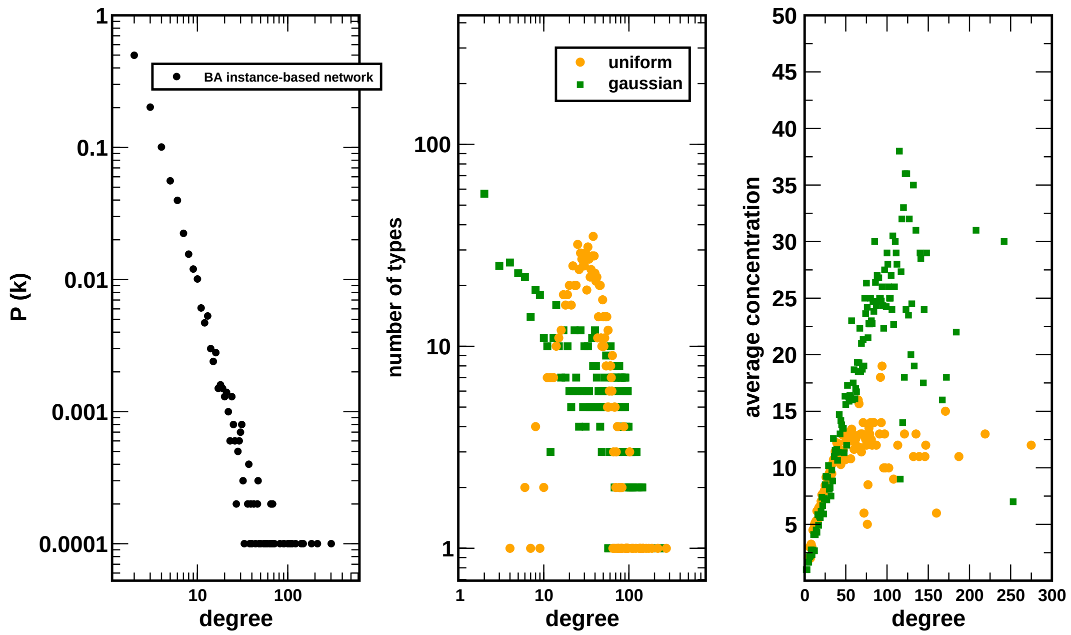

Figure 1: Results of producing a type NoI from a scale-free instance NoI. In the leftmost plot, the original scale-free instance NoI (produced using the BA model) is shown (log-log plot). The middle plot shows the degree-distribution of the type NoI when types are, on the one hand, assigned in a uniform random manner and, on the other hand, assigned using a Gaussian distribution. The rightmost plot shows for each type-network the correlation between degree and concentration.

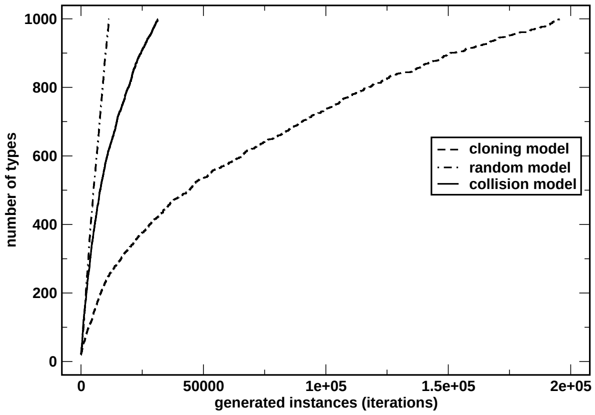

Figure 2: Visualization of the rate of appearance of types over time. Time corresponds only to the number of instances presented to the NoI.

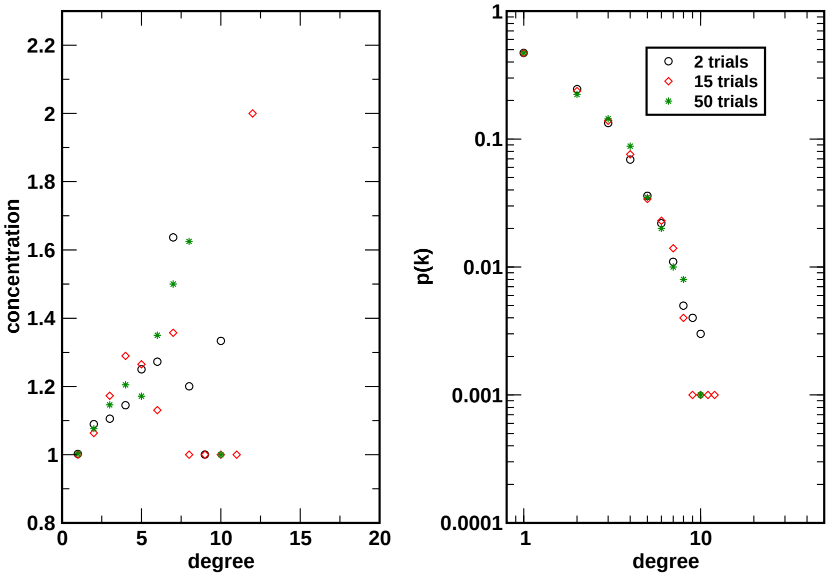

Figure 3: Results of producing a type NoI using random recruitment. The left plot shows the correlation between degree and type. The right plot shows the degree distribution of the NoI in terms of the number of types.

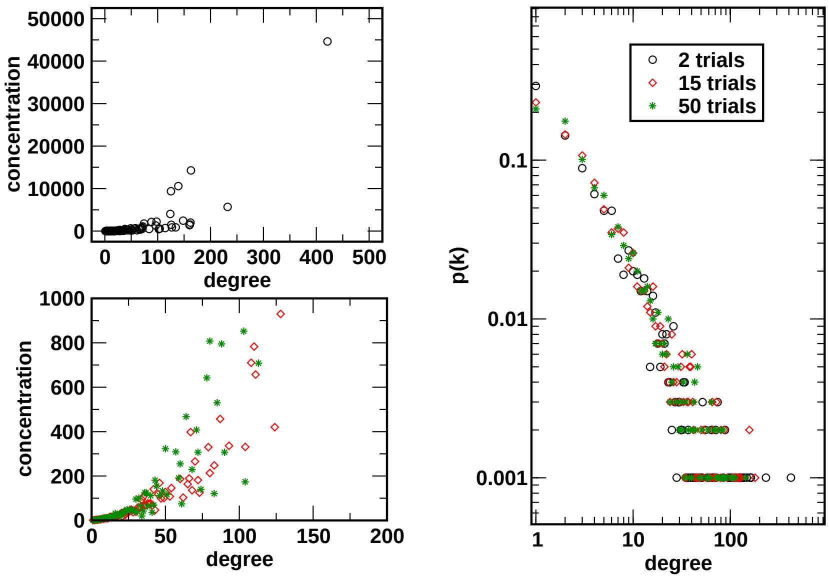

Figure 4: Results of producing a type NoI using cloning and mutation. The left plot shows the correlation between degree and type. The right plot shows the degree distribution of the NoI in terms of the number of types. $p_{mut} = 0.001$

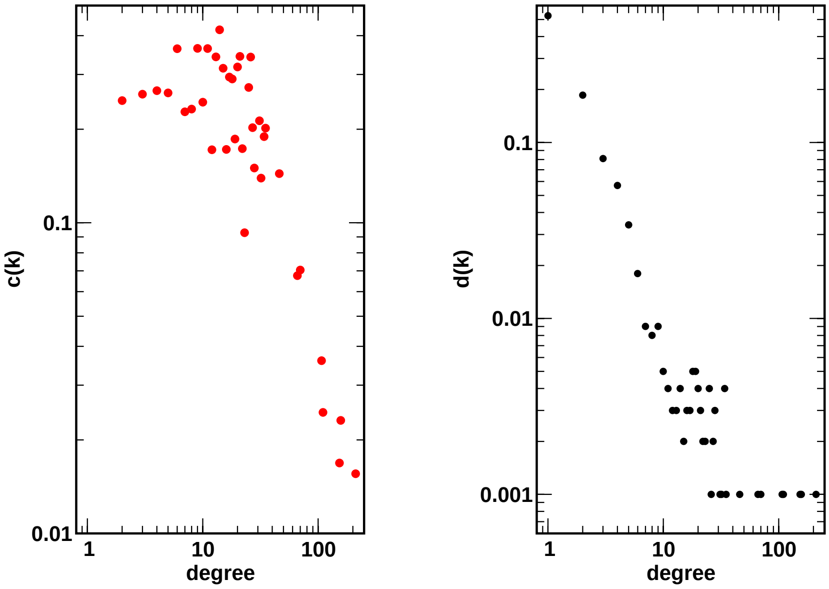

Figure 5: Results of producing a type NoI using clonal recruitment with $t = 2$. The left plot shows the correlation between the clustering coefficient and the degree of the nodes of the network. The right plot shows the resulting degree distribution.

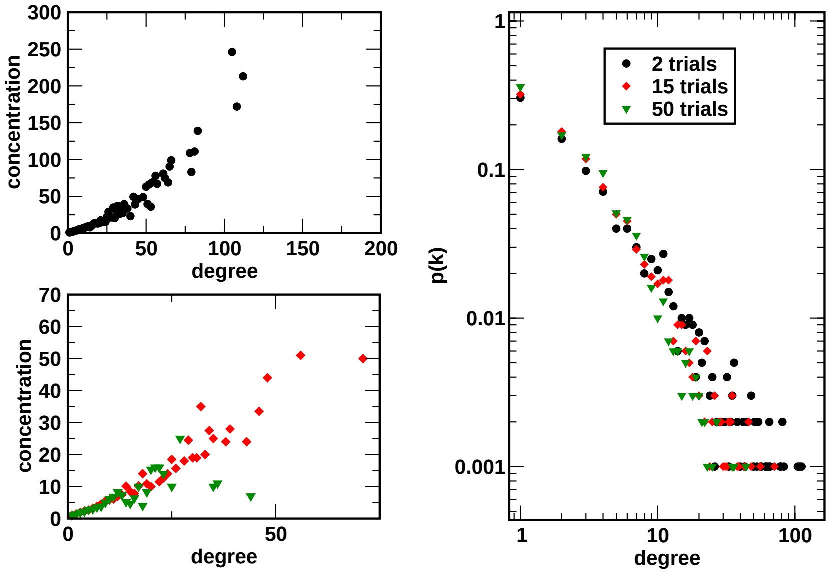

Figure 6: Results of producing a type NoI based on the collision model. The left plot shows the correlation between degree and type. The right plot shows the degree distribution of the NoI in terms of the number of types.

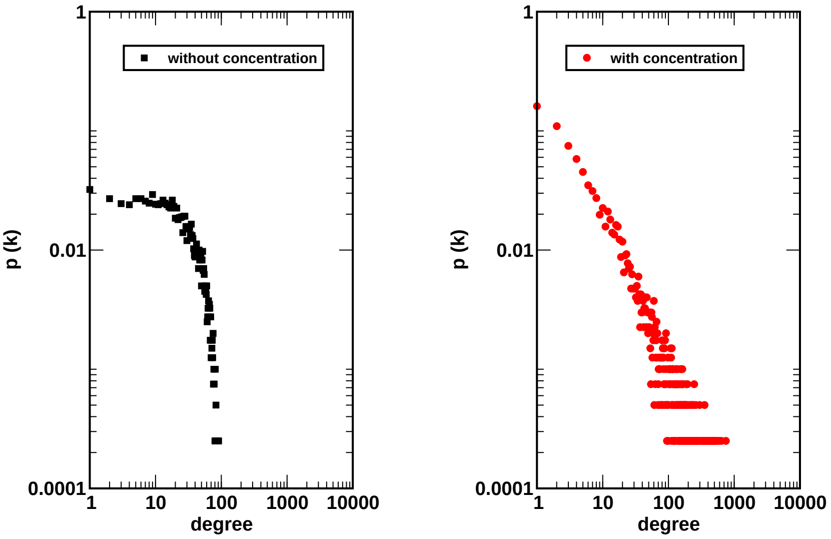

Figure 7: Results of producing a type NoI using clonal recruitment. The left plot shows the degree distribution of the NoI for the clonal model where concentration does not play a role in deciding which type will produce a clone. The right plot shows the degree distribution of the NoI when concentration is used to determine which type will produce a new instance.

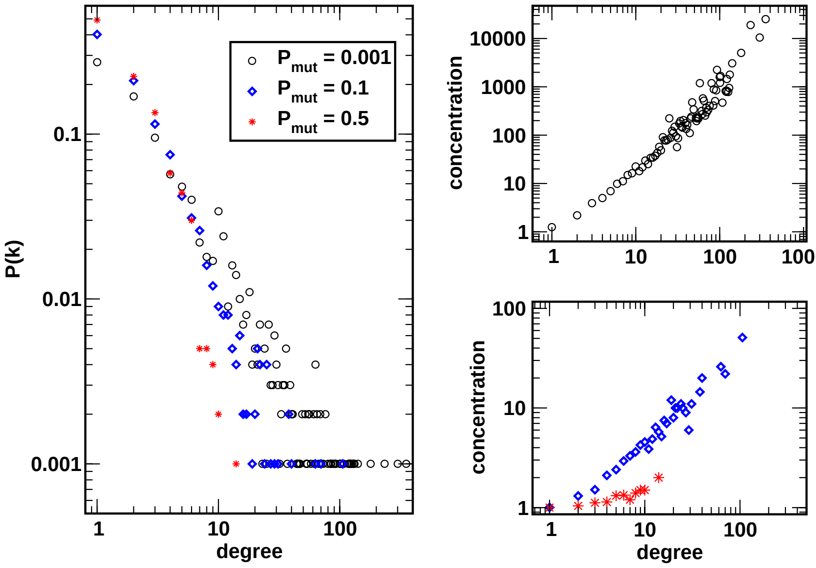

Figure 8: Results of producing a type NoI using clonal recruitment with increasing values of p_{mut}. The left plot shows the degree-distributions for three values of p_{mut}. The two right plots visualise the correlation between degree and concentration for each p_{mut} value.

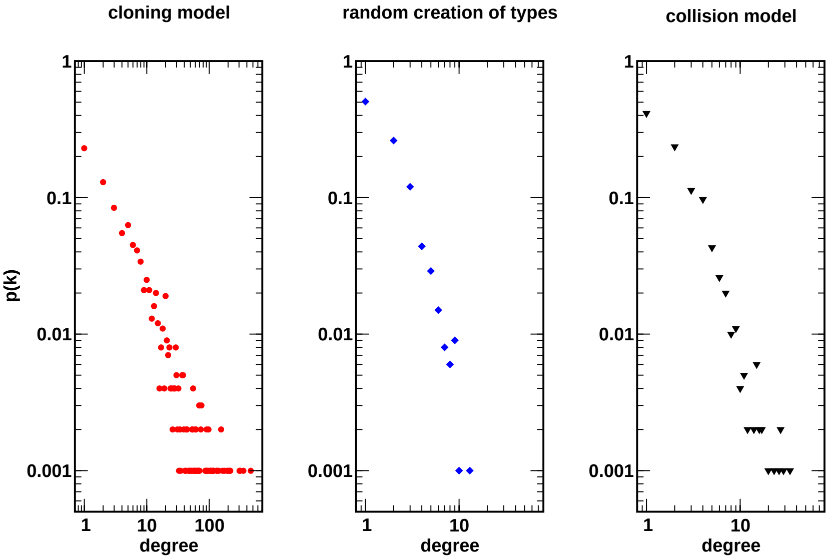

Figure 9: Degree distributions of NoI with natural hubs for all three production schemes.

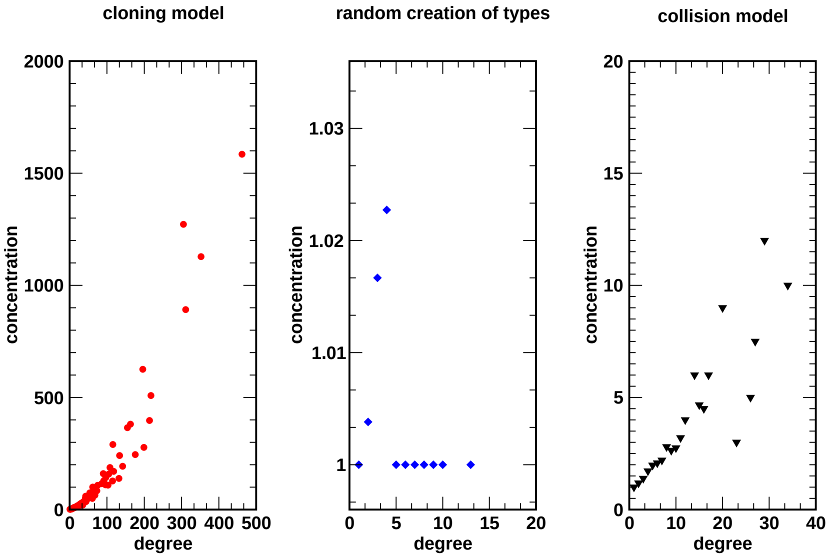

Figure 10: Correlations between degree and concentrationsof NoI with natural hubs for all three production schemes.

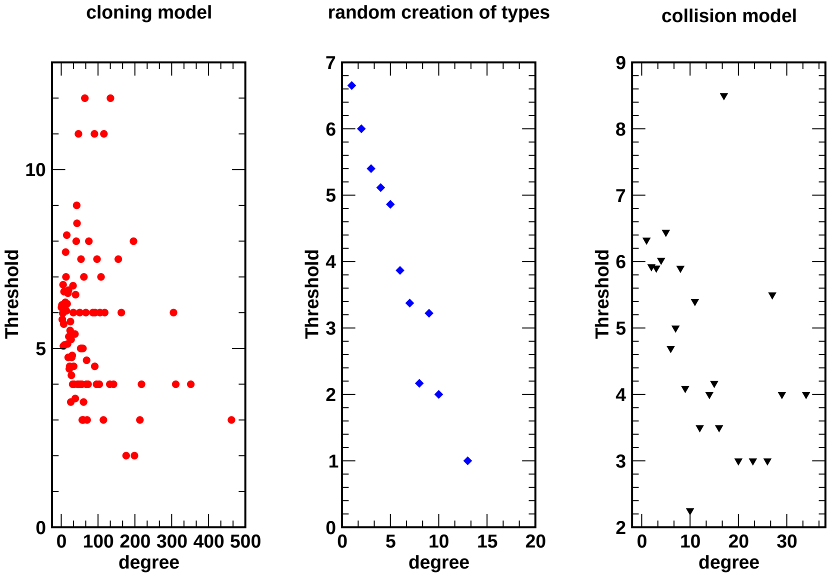

Figure 11: Correlations between AT_i and degree of NoI with natural hubs for all three production schemes.