Probabilistic generation of random networks taking into account information on motifs occurrence

Frédéric Y. Bois, Ghislaine Gayraud

Probabilistic generation of random networks taking into account information on motifs occurrence

Frédéric Y. *Bois*[a,b], Ghislaine *Gayraud*[c,d]

November 26, 2013

a. Chair of Mathematical Modeling for Systems Toxicology, Université de Technologie de Compiègne, Compiègne, France. Email: frederic.bois@utc.fr
b. INERIS, DRC/VIVA/METO, Verneuil en Halatte, France.
c. LMAC, Université de Technologie de Compiègne, Compiégne, France.
d. LS, CREST-INSEE, 3, avenue Pierre Larousse, Malakoff, France

Abstract

Because of the huge number of graphs possible even with a small number of nodes, inference on network structure is known to be a challenging problem. Generating large random directed graphs with prescribed probabilities of occurrences of some meaningful patterns (motifs) is also difficult. We show how to generate such random graphs according to a formal probabilistic representation, using fast Markov chain Monte Carlo methods to sample them. As an illustration, we generate realistic graphs with several hundred nodes mimicking a gene transcription interaction network in *Escherichia coli*.

Introduction

Graph models are essential tools for understanding and modeling complex systems of interacting variables or agents (Albert and Barabási, 2002). The global features of social, telecommunication or biological networks can only be analyzed with correspondingly large graph models. Such models consist of bonds (edges) indicating relationships between agents (nodes), with parameters quantifying the strength of the bonds. Over the years, graph models have been extensively studied both in theory (see among others Lauritzen, 1996) and in methodology (Whittaker, 1990) in different scientific areas, such as statistical physics (Gibbs, 1902), genetics (Wright, 1921), economics (Wold, 1954) or social sciences (Blalock, 1971). When there is no ambiguity about the links between nodes and their interactions strengths, complex systems are well described by deterministic networks. In the presence of ambiguity about the system, probabilistic graphs are usually considered. In that case, bonds represent stochastic links between nodes and their parameters specify conditional distributions. Probabilistic graphs offer a natural framework for statistical inference and knowledge integration. For example, in communication engineering, we may want to know what conditions the robustness of a network; in systems biology we may be interested in understanding how genes control each others.

The theory of random graphs was initiated by Erdös and Rényi in a series of papers Erdös and Rényi, 1959, 1960, 1961) in which they proposed to generate a random graph with a given number of labeled nodes by connecting every pair of nodes with probability p ($p \in (0, 1)$). The main goal of random graph theory has been to determine at which connection probability p a particular property of a graph will most likely emerge. One of the first properties studied by Erdös and Rényi is the appearance of given sub-graphs like cliques, triangles, *etc.* In some sense, the question they have addressed is relative to the structure/topology on a graph. In many settings, interest focuses first on the structure of a graph and the question arose of whether the random graphs proposed by Erdös and Rényi were able to display specific structures of real

complex networks. Over the past few years, there has been a growing interest in investigating and developing new tools and measures able to capture specific properties of real networks. Among such properties are the degree distribution, corresponding to the probability $P(k)$ that a node in the network is connected with k other nodes (Albert *et al*, 2000; Leskovec *et al*, 2010), small-world properties (Watts and Strogatz, 1998) in which most nodes are not neighbours of one another, or the node clustering coefficient (Watts and Strogatz, 1998).

Nevertheless, inference about network structures remains difficult because of the extremely large number of possible graphs, even with a modest number of nodes (Markowetz and Spang, 2007). Estimating the amount of data needed to recover the structure of a graph is also difficult, but it is clear at least in biology, that most of the current experimental designs are insufficiently powerful for that aim. In such a context, every bit of information counts. Still in biology, the relevant scientific literature indicates that all graphs are not equally plausible, some being *a priori* more likely than others. Accounting for prior knowledge is well formalized in Bayesian statistics (Robert, 2001), but the probabilistic representation of such knowledge is still a question. Mukherjee and Speed (2008) have recently proposed a set of informative priors for network structure inference. More precisely, they have considered priors able to capture information relative to existence of edges, degree distribution or sparsity structure in Bayesian networks, *i.e.*, acyclic directed graph models.

Network motifs are patterns (sub-graphs) that recur within a network much more often than expected for random graphs (Milo *et al*, 2002). It has been shown that gene transcription regulation networks, for example in the bacteria *Escherichia coli*, contain a small set of network motifs (see Alon, 2007 and references therein), suggesting that such motifs are basic building blocks of transcription networks. An important aspect of network topology inference is therefore to include the probability of occurrence of such network motifs (see Janson *et al* (2000) for an overview).

In this article, we address precisely that question. We extend the approach of Mukherjee and Speed (2008) by relaxing the acyclicity requirement which characterizes Bayesian networks and propose rigorous probabilistic representations of *a priori* information on pairwise links, degree node distribution, and network motifs. We use Markov chain Monte Carlo (MCMC) simulations to sample networks satisfying those joint distributions, for moderately large networks (several hundred to thousands of nodes). Such random networks can be used as priors for formal inference, after updating with data in a Bayesian framework. They can also be used for pure simulation purposes, *e.g.* for methods or software testing. We do not deal with data and associated likelihood (or "score") functions, but focus on probabilities. Our distributions, however, are entirely compatible with any score function and can be used for inference, in particular in a Bayesian framework. As an illustration of our results, we generate realistic graphs mimicking a gene transcription interaction network in *E. coli*. The weight of the various proposed priors is examined.

Results

Graph Models for Networks

A graph model simply consists of nodes (vertices) connected by edges (Wilson, 2012) The nodes often represent physical entities (people, genes, servers...) and the edges represent links or dependencies between them ("is a friend", "controls", "is physically connected to", ...). Nodes can be assigned attributes (*e.g.*, "on", "off") which can depend in turn on the attributes of the nodes to which they are connected. The edges may also have attributes influencing those of the nodes they connect. Edges may be "undirected" or "directed", the latter case (often noted by an arrow) introducing an asymmetry between the two nodes. For example, an arrow from node i to node j may indicate that i controls j, the reverse being not true. Directed edges can in turn be signed, indicating a positive or negative control, *etc.* Finally, graph models may have global features imposed to them. For example, we may impose no unconnected node. A commonly imposed feature is "acyclicity". In that case, the graph model cannot contain any path (succession of edges) linking any node to itself, and in particular no "auto-loop" edge from a node to itself. In the case of directed edges, paths are understood to follow the directions of the edges. A particular class of such acyclic directed graph models, Bayesian networks, has a clear probabilistic interpretation and is easily amenable to inference about network structure or parameters (Neapolitan, 2003)

Since we are interested in generating general graphs, in particular those describing genetic regulatory systems, we consider directed graph which may be cyclic (see De Jong (2002) for an overview of genetic regulatory networks modeling). More precisely, our graphs are composed of labeled nodes with directed edges. There can be two reverse edges between any two nodes and auto-loops are allowed.

Informative priors on networks

Let G be a graph described by a set $V(G) = \{v_1, \ldots, v_n\}$ of n vertices ($n \geq 2$) and a set $E(G) = \{e_{i,j} : (i,j) \in \{1,\ldots,n\} \times \{1,\ldots,n\}\}$ of directed edges and auto-loops. G may be described by its adjacency matrix A, a $(n \times n)$-matrix whose generic term is given by $a_{i,j} = \begin{cases} 1 & \text{if the edge } e_{i,j} \text{ exists} \\ 0 & \text{otherwise} \end{cases}$.

Incomplete *a priori* knowledge on such a graph can be described by a statistical distribution. Given n, the number of nodes of G, we propose to include three levels complexity in that distribution. It is only mandatory to define the first level, which does not reflect any specific structure except for the probability of presence of individual edges. We then refine it by including information on the degree distribution.The next step is to incorporate information related to the occurrences of sub-network motifs.

Priors on individual edges

As in the random graphs considered by Erdös and Rényi, prior knowledge on each individual edge can be conveniently modeled by a Bernoulli distribution, assigning probability $p_{i,j}$ to the existence of a directed edge from node i to node j, that is, $e_{i,j} \sim B(p_{i,j})$ for all $(i,j) \in \{1,\ldots,n\} \times \{1,\ldots,n\}$. In this context, the adjacency matrix A related to a graph G with n nodes, becomes $A = (e_{i,j})_{\{1 \leq i,j \leq n\}}$.

For a graph of a given size n, there are n^2 individual pairwise possible links in the cyclic case and at most $n^2 - n$ in the acyclic case (auto-loops being ruled out by definition). Specifying the complete set of edges priors requires the definition of a $n \times n$ matrix $\mathbf{P} = (p_{i,j})_{1 \leq i,j \leq n}$ (with a null diagonal in the acyclic case). If the pairwise links are supposed to be independent, the probability distribution for the entire graph G is therefore

$$P_{Bern,G} = \prod_{i,j=1}^{n} p_{i,j}^{e_{i,j}} (1 - p_{i,j})^{e_{i,j}-1}. \tag{1}$$

There are various ways to choose or elicit values for the individual prior probabilities $p_{i,j}$, which we will further discuss in our application but intuitively they are related to the weight of the prior evidence (*e.g.*, p-values (Bernard and Hartemink, 2005) we have on the existence of given edges. Other distributions could be used, such as a =multinomial if we had chosen to give a sign to the edges (to indicate positive or negative interactions, for example). But the principle would remain the same, and in the absence of precise prior information, a Bernoulli prior is probably all we can specify.

Priors on degrees' counts

The degree $deg(v)$ of a vertex v is the total number of edges to which vertex v participates. The degree distribution of a graph G is a function $P(d)$ expressed in terms of $|\{v \in V(G) : deg(v) = d\}|$, the total number of vertices having degree d. In many biological networks (Jeong *et al*, 2000), it appears that the degree distribution has a power-law tail, which means that $P(d) \propto d^{-\gamma}$, with $\gamma > 0$. Such networks are called scale-free (Barabási and Albert, 1999). Then, we define the probability distribution of a graph G as follows

$$P_{deg,G} \quad \propto \quad P_{Bern,G} \times \prod_{i=1}^{n} \left(\sum_{j \in \{1,\ldots,n\} : \sum_j e_{i,j} > 0} e_{i,j} \right)^{-\gamma}. \tag{2}$$

Since we do not impose that every node should be linked to another one, the degree distribution attributes implicitly a weight of one to any isolated node. It entails that the probability, with respect to degrees, of an empty graph (without any connection between its nodes) is one.

Here again, other distributions, even empirical (as defined by an histogram) and reflecting better the degree distribution of a given class of graphs, could be used if enough information was available.

Priors on motifs

One important local property of networks is the eventual occurrence of motifs. Motifs are defined here following Milo *et al* (2002), Alon (2007), Shen-Orr *et al* (2002), Milo *et al* (2004) and Kashtan *et al* (2004), as over-represented sub-graphs compared to what is found in an Erdös and Rényi random graph. Some motifs have a notable importance in biological networks because they can carry out specific information-processing functions, and hence may help in understanding the global behavior of such networks (Masoudi-Nejad, 2012). For example, there are thirteen possible configurations for the relationships between three nodes (see Figure 1). Among the non-degenerate configurations, only the feed-forward loop (FFL, top row, first motif on the left, Figure 1) has been found in the transcriptional regulation network of *E. coli* (Alon, 2007; Shen-Orr *et al*, 2002; Mangan *et al*, 2003). No feed-back loop (FFB, top row, second motif on the left, Figure 1) has been observed in *E. coli*. The FFL is one of the most studied network motifs in transcription interactions. It corresponds to a directed sub-graph of three nodes (genes) such that one of them is regulated by the two others, which are linked. Given that each of the regulatory interaction can either an activation or a repression, there are eight sub-types of signed FFL, two of them occurring much more frequently than the other six in transcription networks (Mangan and Alon, 2003; Mangan *et al*, 2006).

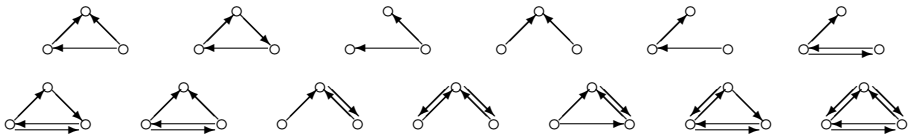

Figure 1: The thirteen possible three-node motifs.

In this paper, we consider network motifs with three nodes; auto-loops are not taken into account in such sub-graphs. We define a motif distribution based on the proportion of FBL among all three-node loops. More precisely, for a graph G let us consider N_1 the number of FBL motifs and N_2 the number of FFL motifs and note they may be expressed in terms of $(e_{i,j})_{\{1\leq i,j\leq n\}}$ as follows, $N_1 = \sum_{(i,j,k):i\neq j,k\neq j,i\neq k} e_{i,j}e_{j,k}e_{k,i}$ and $N_2 = \sum_{(i,j,k):i\neq j,k\neq j,i\neq k} e_{i,j}e_{j,k}e_{i,k}$.

For a graph G with a total number $N_1 + N_2$ of motifs of type FBL and FFL, we place a beta-binomial probability with parameters u and v on N_1, $BB(N_1|u, v, N_1 + N_2)$. The prior for graph G is then

$$P_{Motif,G} \quad \propto \quad P_{Bern,G} \times C_{N_1+N_2}^{N_1} \frac{B(N_1 + u, N_2 + v)}{B(u, v)}, \tag{3}$$

where $C_{N_1+N_2}^{N_1}$ is a Binomial coefficient and $B(\cdot, \cdot)$ denotes the Beta function. The choice of a Beta-Binomial distribution is justified by the fact that we may not know the exact proportions of FFL and FBL, but simply have observations about the numbers of such loops in some actual network we base our prior on. Obviously, as we will discuss later, other motifs could be tracked and entered in the definition of a graph probability, using a similar device.

Piecing together a global prior

In addition to the probabilities $P_{Bern,G}$, $P_{Deg,G}$ and $P_{Motif,G}$ defined by (1), (2) and (3) respectively, we consider one more graph distribution $P_{Total,G}$ which combines all informative priors independently. Therefore:

$$P_{Total,G} \propto \prod_{i,j=1}^{n} p_{i,j}^{e_{i,j}}(1-p_{i,j})^{e_{i,j}-1} \times \prod_{i=1}^{n} \left(\sum_{j\in\{1,\ldots,n\}:\sum_j e_{i,j}>0} e_{i,j} \right)^{-\gamma} \times C_{N_1+N_2}^{N_1} \frac{B(N_1+u, N_2+v)}{B(u,v)}. \quad (4)$$

Application to a Biological Network

Transcriptional regulatory networks orchestrate the gene expression of cells. In such networks, the nodes are operons (one or more genes transcribed on the same mRNA template). Edges go from operons encoding a transcription factor to operons directly regulated by that factor. Shen-Orr *et al.* (2002) developed and applied motif-detection algorithms to the transcriptional regulation network of *E. coli*. They extracted data from the RegulonDB transcriptional database (Salgado *et al*, 2013), and enhanced them with additional transcription factors and interactions described in the literature. We used here the latest version of the dataset (version 1.1, made publicly available by Dr. U.Alon).

A graph representation of the *E. coli* transcriptional regulatory network is shown on Figure 2. It contains 423 nodes, all connected, with 578 directed edges. That is actually only 0.32% of the number of possible edges, indicating that the network is sparse.

We report here the results for Alon's full size network (of 423 nodes). We investigate here which elements of our prior knowledge on *E. coli* regulatory network features are the most important to simulate realistic networks.

In a first set of simulations we assigned "vague" priors to the edge probabilities, setting them all to the same value (equal to $578/423^2$, *i.e.*, 0.0032). With that prior, all connections are equally probable, and their expected number is equal to the one observed for *E. coli*. In *E. coli* the degree distribution follows approximately a power law with an exponent of 1.7 (see Figure 6), so we set γ to that value when the degree distribution was in effect. In addition, *E. coli* regulatory network is known to contains 42 FFLs and no FBL. To allow flexibility in the prior and allow some occurence of FBL motifs, we set equation (3) parameter u to 2 and parameter v to 50 in our simulations. That implies an expected proportion of only two percents FBLs.

In the second set of simulations we use the same priors for degree distribution and motifs frequencies, but used informative priors to individual edge probabilities: The edges reported by Alon *et al.*, were assigned probability 0.95. Non-reported, therefore hypothetical, edges were assigned probability 0.00016. Together those probabilities lead again to 578 expected edges.

In all cases, three MCMC chains of 2 billions iterations were run independently. Convergence of the edge probabilities (according to Gelman and Rubin's criterion) was always attained after at most 1 billion iterations. The degree distribution (*e.g.*, Figure 6) and motif frequencies (*e.g.*, Figure 4) also converged within that time frame: Results from the three independent chains basically overlap, except in the case of rare events (with frequencies less that one in 10,000), where Monte Carlo sampling uncertainty becomes noticeable. We therefore discarded systematically the first billion simulations and base all the following results on the second billion. We optimized our MCMC sampling C code and simulations are rather fast. Running 2 billions iterations to generate a random graph with 423 nodes takes about 2 minutes on a Intel Core 2 Duo machine clocked at 2.13 GHz. It takes actually little more time to sample graphs with a thousand nodes. Overall, the time it takes to update all the elements of the adjacency matrix is approximately proportional to the number of its elements, and therefore proportional to the square of the number of nodes for the graph considered. In our implementation, memory requirements are simply proportional to the number of nodes and minimal.

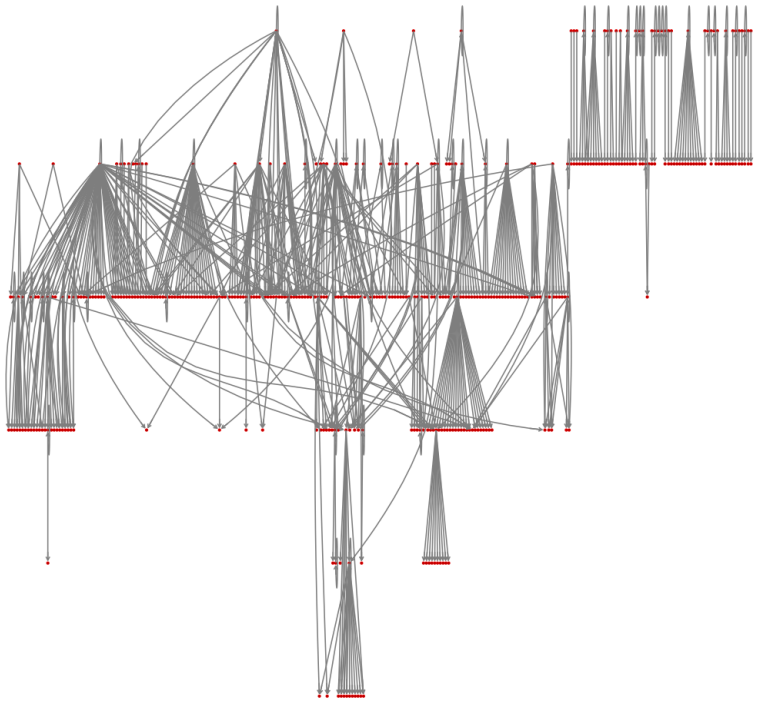

Figure 2: *E. coli* transcriptional regulatory network, as reported in Shen-Orr *et al* (2002), with minor updates (see text).

Figure 3 shows samples of networks generated using the above vague prior on individual edges. Used alone, that prior gives all edges the same probability and the resulting network has little structure, except that the expected number of edges is respected (Figure 3, panel A). The proportion of FBLs is 1/4, as expected in an unconstrained setting (there are two possible FBLs and six possible FFLs for each triplet of nodes). Adding the prior component on degree distribution imposes a major change in network shape. The number of edges is similar, but the structure becomes hierarchical (panel B). Placing a prior on the proportion of FBLs and FFLs, in addition to the prior on individual edges, has little visible impact on the structure (panel C), but the proportion of loops is now much lower and close to its expected value. Finally (panel D) putting the three priors together gives us again a hierarchical structure, but with the correct proportion of FBLs.

Figure 4 shows in more details how the number of feedback and feed-forward loops is influenced by the specification of prior knowledge in the context of a vague specification of individual edge probabilities. The hierarchical structure imposed by the degree distribution (prior "B") leads to a much lower number of loops, but without altering the ratio of FBLs to FFLs (25% in the case of prior "A" and "B"). In contrast, imposing a prior on the proportion of FBLs can later reduce both the ratio (4% in the case of prior "C" and 2% with prior "D") and the number of loops in the network.

If we now turn to networks simulated with an informative prior on individual edges (Figure 5), we see a striking difference with Figure 3. The structure of those networks, even if random, is quite close to the actual *E. coli* transcriptional regulation network (Figure 2). The difference between the networks with a prior on degrees (panels B and D) or without (panels A and C) is now more subtle. Actually the prior on individual edges is strong enough to impose a correct distribution of degrees, even if the degree distribution is not specified (see Figure 6). In that Figure, a deviation of the actual number of degrees from the power law, for high degrees, can be observed and is well simulated. A similar behavior of degree distribution can

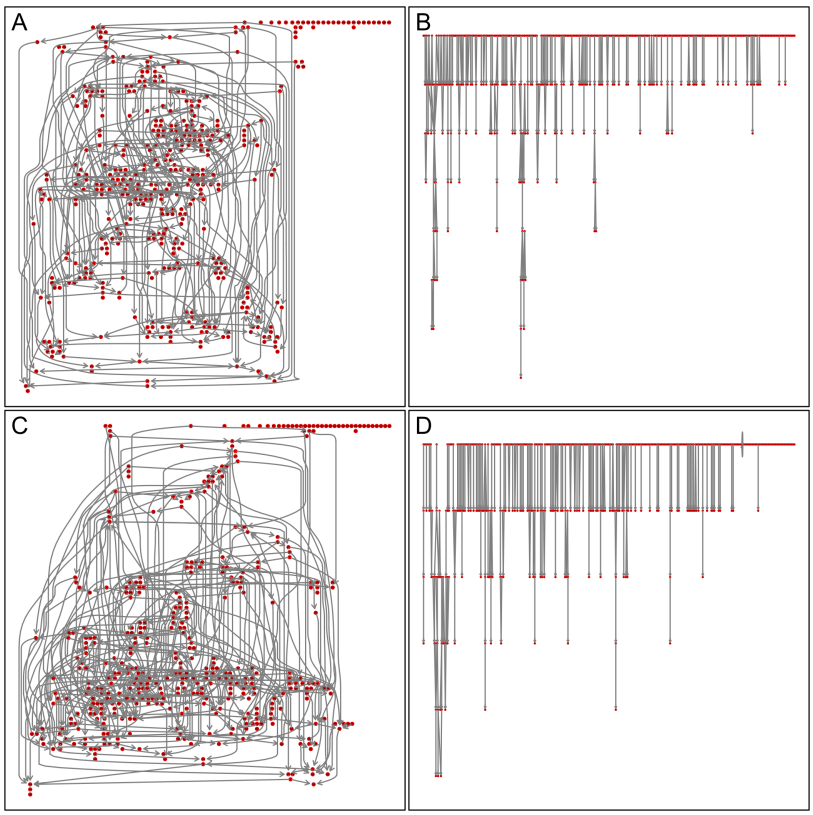

Figure 3: Transcriptional regulation networks generated using a vague prior on individual edges. Red dots: 423 genes in each network. Panel A: prior on individual edges only; B: prior on individual edges and degree distribution; C: prior on individual edges and the proportion of feed-back loops; D: all three priors together.

be found in (Dobrin *et al*, 2004). In Figure 5, the frequency of motifs is also controlled directly by the edges probabilities: The number of FBLs is about constant at 1.0 10^7, for approximately 3 10^{10} FFLs, hence a proportion of 0.05%. The differences between chains are small and all those results have a 5% CV.

Discussion

There are two important applications to the generation of semi-random graphs with known properties: *i.* Simulating actual networks for hypothesis testing, software bench-marking, statistical power calculations *etc.* (Van den Bulcke, 2006); *ii.* Assessing whether a graph is coherent with our prior knowledge in numerical data analytic methods such as Bayesian network modeling, Gaussian graphical methods *etc.* Such methods, in their naive implementation, are known to suffer badly from the curse of dimensionality. However, prior knowledge about network structure increases daily, at least for biological networks, and a proper accounting of such knowledge is our only hope to redeem the curse we face.

We have extended here the results presented in Mukherjee and Speed (2008) by including a flexible specification of edge probability, via Bernoulli priors, and by defining a prior on network motifs. In doing so we have dropped the commonly used distance penalty from a pre-specified reference network (whereby "distance" correspond to the number of differing edges between a proposed graph and the reference graph). Such a distance penalty is simple to specify and compute, but has several drawbacks: It is quite coarse and does not give a weight to the various edges, while in fact we may be more or less certain about some of them. Therefore it treats in the same way edges known to be absent or present, and edges for which the we do not know whether they are absent or present. Also, only one reference network is specified and this limits the number of questions we can ask. It is finally more an *ad hoc* penalty function than a proper distribution,

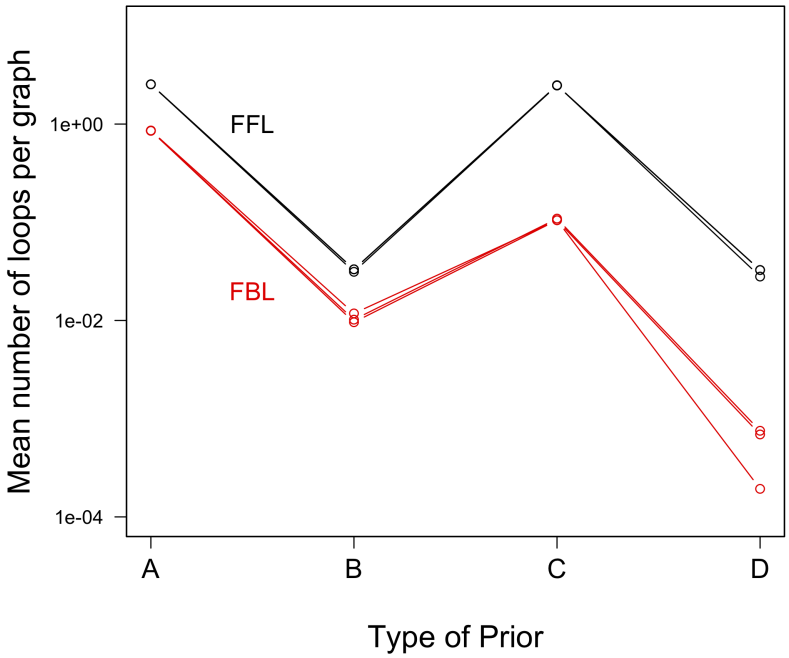

Figure 4: Motifs frequencies in transcriptional regulation networks generated using a vague prior on individual edges. A: prior on individual edges only; B: prior on individual edges and degree distribution; C: prior on individual edges and the proportion of feed-back loops; D: all three priors together.

although that may be seen as a purely technical argument. In any case, an edge by edge prior probability assignment is not much more difficult to specify, it can be simplified by giving default vague probability values if information is lacking, and can be quite powerful when information is available.

There are, however, cases where higher levels of structure are important. In fact, we can hypothesize that higher levels are more important than we usually suspect: That is the whole point of systems biology. Power law degree distribution is a well known characteristic of biological networks. The frequency of occurrence of particular network motifs is another point in case. Fascinating recent work has addressed the question of degree distribution (Leskovec *et al*, 2010), but the problem remains for motifs since no direct sampling or generative method is available in that case. We have shown here how to use MCMC sampling to obtain the desired random graphs, even for realistically large number of nodes. An important point is that a sample of random graphs is much more informative than a single approximate, or even "best", estimate graph. With a single graph, all sense of uncertainty is lost, and only over-confidence is gained. Ensemble results are much more robust and useful; but they can be cumbersome and the question is how to best handle them.

Stochastic simulations also give insights about the relative weights of the various components of our prior knowledge. A first point is that prior knowledge about the probable number of edges, or at least about network sparsity, can strongly constrain the set of admissible networks. At least, that is the case when implemented in a form of Bernoulli priors on individual edges. Actually, more flexible priors (*e.g.*, hierarchical) could be used instead of Bernoulli to allow more uncertainty about that expected number of edges. Specific knowledge about subsets of high probability, or conversely low probability, edges is also very informative. The degree distribution of a network may appear as a weak predictor of its structure since, for example, with a given number of nodes, two graphs with the same degree distribution may have completely different edge lists. However, we found that specifying a degree distribution has a visible impact on the

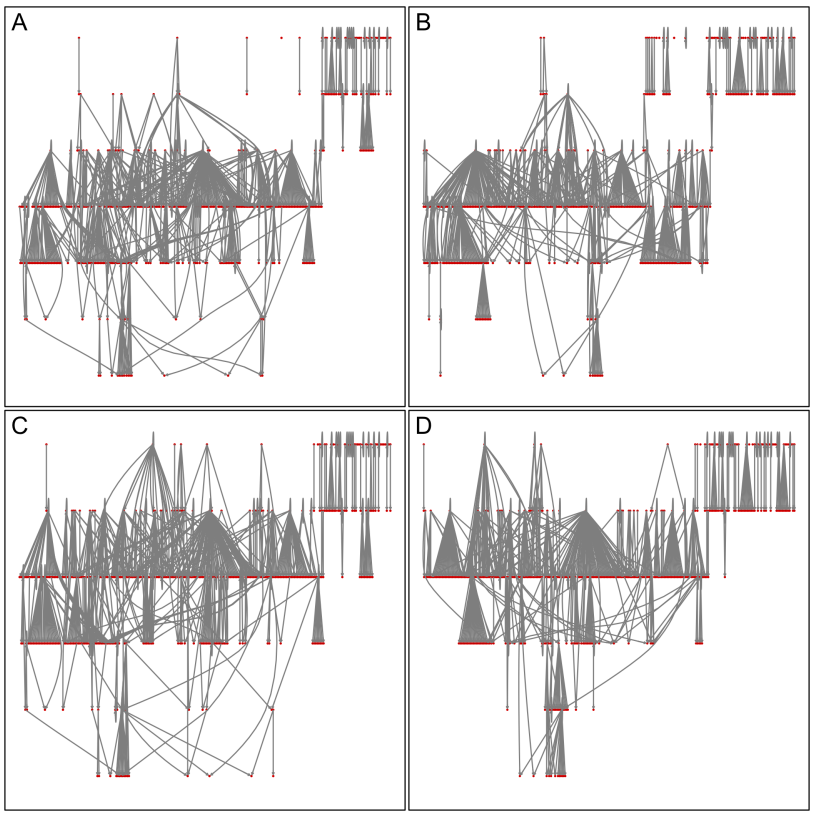

Figure 5: Transcriptional regulation networks generated using informative priors on individual edges (on the basis of *E. coli* network). Red dots: 423 genes in each network. Prior on individual edges (A), on individual edges and degree distribution (B), on individual edges and the proportion of feed-back loops; D: all three priors together.

network structure. Here again it would be easy to be more flexible about that distribution. Imposing a prior on the occurrence of specific motifs can be important in terms of functionality, but leads to more subtle modifications. Note however, that we imposed a distribution on the relative frequency of two loop motifs, rather than on there absolute number. That would be an easy extension, which could have profound consequences on the network structure. Overall, there are many variables with which the prior can play, and even potential conflicts between components of our prior knowledge, in particular if a strongly informative prior is placed on individual edges. Such conflicts can be hard to figure out without the help of simulations, because our intuition often fails in high dimension and in the case of graphs (Helbing, 2013). In that respect, the possibility to perform quickly billions of iterations for network of sizes commensurable to those of genomes is encouraging. A word of caution is in order here, however: With 423 nodes, a billion simulations correspond to $10^9 \times 423^{-2}$, *i.e.* about 6000 full updates of the network. With 10^4 nodes we might have to go to trillions of simulations to get to convergence and this would currently take us a day and half of computation, although GPU computing, for example, could again increase speed.

To be more precise about the relative weight of the different priors would have required some evaluation of the number of possible graphs, or some form of enumeration of the number of different graphs sampled during stochastic simulations. However there are, for example 2^{178929} different directed graphs with 423 nodes, and tracking the list of graph sampled would entailed considerable time and memory capacity.

Obviously, it would be interesting to extend those results to the important problem of statistical learning of the network structure from data acquired in large scale genotyping or phenotyping studies, for example. In a Bayesian context, that simply entails the computation of a data likelihood function, or its marginalisation. However, the models currently favored for that purpose: Gaussian graphical networks (Krumsiek *et al*,

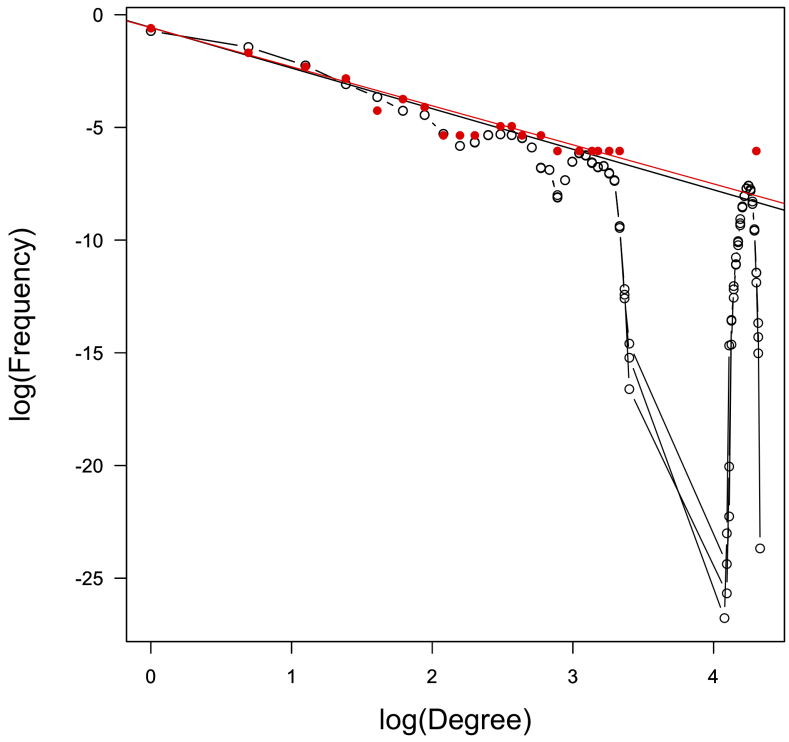

Figure 6: Degree distribution in *E. coli* actual transcriptional regulation network (in red) and in Monte-Carlo sampled networks (in black). Informative distribution on individual edges. The dip for high degrees due to a deviation of reality from the power law assumption

2011; Liu *et al*, 2012) and Bayesian networks (Mukherjee and Speed, 2008), are either undirected or acyclic, respectively. Dealing with undirected graphs would be easy, but the acyclicity of Bayesian networks is quite restrictive. Loops motifs cannot exist formally in such models, unless they are made dynamic, and checking for acyclicity at every iteration imposes significant computational burden. The possibility of using hybrid models (Silva and Ghahramani, 2009) is an interesting possibility to explore.

Materials and methods

Graph probabilities such as $P_{Total,G}$ are only defined up to a multiplicative constant. In that case, a simple way to generate sample graphs according to that probability distribution is to use the Metropolis-Hasting sampler (Casella and Robert, 2004). From a current graph G, with total probability $P_{Total,G}$ (eq. (4)), a graph $\tilde{G}$ is proposed by first selecting two nodes, say v_i and v_j of G (v_i may be equal to v_j in case of auto-loops) and then by deciding on the presence of an edge from v_i and v_j by a random Bernoulli draw with edge probability $p_{i,j}$. That amounts to sampling $\tilde{G}$ from the Bernoulli prior on edges defined in eq. (1). The total probability $P_{Total,\tilde{G}}$ of $\tilde{G}$ is computed using eq. (4) and $\tilde{G}$ is accepted with a probability equal to $\min(1, P_{Total,\tilde{G}} \,/\, P_{Total,G})$. In case of rejection, $\tilde{G}$ is discarded, and G is again the current sample (that implies that the same graph can be drawn several times in succession). The procedure is iterated as many times as it is needed to reach convergence in probability to the target distribution sought. Convergence can be checked by running several simulation chains and computing Gelman and Rubin's $\hat{\mathrm{R}}$ criterion (Gelman and Rubin, 1992) on each element of the graphs' adjacency matrix, or by monitoring the degree distributions obtained, or the motifs probabilities in those independent chains.

A C language version of the algorithm has been implemented as a module of the free *GNU MCSim* software (Bois, 2009) (http://www.gnu.org/software/mcsim). That software was used for all the simulations presented here. Graphs were produced with R, version 2.14 (R Development Core Team, 2011).

Acknowledgments

The research leading to these results has received funding from the scientific council of the Université de Technologie de Compiègne (project Prior-Motives) and the Innovative Medicines Initiative Joint Undertaking, under Grant Agreement number 115439 (StemBANCC), resources of which are composed of financial contribution from the European Union Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution. This publication reflects only the author's views and neither the IMI JU nor EFPIA nor the European Commission are liable for any use that may be made of the information contained therein.

Author Contributions

F.B. and G.G. designed and performed research, and wrote the paper.

References

Albert R, Barabási AL (2002) Statistical mechanics of complex networks. *Reviews of Modern Physics* **74:** 47–97

Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. *Nature* **406:** 378–382

Alon U (2007) Network motifs: theory and experimental approaches. *Nature Reviews Genetics* **8:** 450–461

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

Bernard A, Hartemink AJ (2005) Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pacific Symp. Biocomput.:* 459–470

Blalock HM (1971) *Causal models in the Social Sciences*, MacMillan, London

Bois FY (2009) GNU MCSim: Bayesian statistical inference for SBML-coded systems biology models. *Bioinformatics* **25:** 1453-1454

Casella G, Robert CP (2004) *Monte Carlo Statistical Methods*, Springer-Verlag, Berlin

De Jong H (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* **9:** 67–103

Dobrin R, Beg QK, Barabasi AL, Oltvai ZN (2004) Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. *BMC Bioinformatics* **5:** 10

Erdös P, Rényi A (1959) On random graphs. *I. Publ. Math. Debrecen* **6:** 290–297

Erdös P, Rényi A (1960) On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5:** 17–61

Erdös P, Rényi A (1961) On the strength of connectedness of a random graph. *Acta Math. Acad. Sci. Hung.* **12:** 261–267

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences (with discussion). *Stat. Sci.* **7:** 457–511

Gibbs JW (1902) *Elementary Principles in Statistical Mechanics*, New York: Scribner's, 1902. Reprint, Wood-bridge, Conn.: Ox Bow Press, 1981, vi–viii

Helbing D (2013) Globally networked risks and how to respond. *Nature* **497:** 51–59

Janson S, Rucinski A, Luczak T (2000) *Random graphs*, Wiley

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL (2000) The large-scale organization of metabolic networks. *Nature* **407:** 651–654

Kashtan N, Itzkovitz S, Milo R, Alon U (2004) Efficient sampling algorithm for estimating sub-graph concentrations and detecting network motifs. *Bioinformatics* **20:** 1746–1758

Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology* **5:** 21

Lauritzen SL (1996) *Graphical models*, Oxford Science Publications, vol. 17, United Kingdom: Clarendon Press

Leskovec J, Chakrabarti D, Kleinberg J, Faloutsos C, Ghahramani Z (2010) Kronecker graphs: an approach to modeling networks. *J. Mach. Learning Res.* **11:** 985–1042

Liu H, Han F, Yuan M, Lafferty J, Wasserman L (2012) High-dimensional semiparametric Gaussian copula graphical models. *Annals of Statistics* **40:** 2293–2326

Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. *Proc. Natl Acad. Sci. USA* **100:** 11980–11985

Mangan S, Zaslaver A, Alon U (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J. Mol. Biol.* **334:** 197–204

Mangan S, Zaslaver A, Alon U (2006) The incoherent feed-forward loop accelerates the response-time of the gal system of *Escherichia coli*. *J. Mol. Biol.* **356:** 1073–1081

Markowetz F, Spang R (2007) Inferring cellular networks - a review. *BMC Bioinformatics* **8 (Suppl. 6):** S5

Masoudi-Nejad A, Schreiber F, Razaghi MKZ (2012) Building Blocks of Biological Networks: A Review on Major Network Motif Discovery Algorithms. *IET Systems Biology* **6:** 164–174

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

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network Motifs: Simple Building Blocks of Complex Networks. *Science* **298:** 824–827

Mukherjee S, Speed TP (2008) Network inference using informative priors. *PNAS* **105:** 14313–14318

Neapolitan RE (2003) *Learning Bayesian Networks*, Prentice Hall

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org

Robert CP (2001) *The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation*, Springer-Verlag, New York

Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, *et al.* (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.* **41:** D203–D213

Shen-Orr S, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31:** 64–68

Silva R, Ghahramani Z (2009) The hidden life of latent variables: Bayesian learning with mixed graph models. *J. Machine Learning Res.* **10:** 1187–1238

Van den Bulcke T, Van Leemput K, Naudts B, van Remortel P, Ma H.W, Verschoren A, De Moor B, Marchal K (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics* **7:** 43

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* **393:** 409–410

Whittaker J (1990) *Graphical models in applied multivariate statistics*, Wiley series in probability and mathematical statistics: Probability and mathematical statistics, Wiley

Wilson RJ (2012) *Introduction to Graph Theory*, Fifth edition, Pearson

Wold HOA (1954) Causality and econometrics. *Econometrica* **22:** 162–177

Wright S (1921) Correlation and causation. *Journal of Agricultural Research* **20:** 557–585