Tailored Random Graph Ensembles

ES Roberts$^{1,2,3}$, A Annibale$^{1,2,3}$ and ACC Coolen$^{1,2,3,4}$

$^1$ Institute for Mathematical and Molecular Biomedicine, Hodgkin Building, King’s College London, London SE1 1UL, United Kingdom
$^2$ Department of Mathematics, The Strand, King’s College London, London WC2R 2LS, United Kingdom
$^3$ Randall Division of Cell and Molecular Biophysics, New Hunts House, King’s College London, London SE1 1UL, United Kingdom
$^4$ London Institute for Mathematical Sciences, 35a South St, Mayfair, London W1K 2XF, United Kingdom

E-mail: ekaterina.roberts@kcl.ac.uk

Abstract. Tailored graph ensembles are a developing bridge between biological networks and statistical mechanics. The aim is to use this concept to generate a suite of rigorous tools that can be used to quantify and compare the topology of cellular signalling networks, such as protein-protein interaction networks and gene regulation networks. We calculate exact and explicit formulae for the leading orders in the system size of the Shannon entropies of random graph ensembles constrained with degree distribution and degree-degree correlation. We also construct an ergodic detailed balance Markov chain with non-trivial acceptance probabilities which converges to a strictly uniform measure and is based on edge swaps that conserve all degrees. The acceptance probabilities can be generalized to define Markov chains that target any alternative desired measure on the space of directed or undirected graphs, in order to generate graphs with more sophisticated topological features.

1. Introduction

Consider looking at a network not as a set of nodes and edges, but as a collection of topological properties. Any observed set of topological features can be used to define a family of networks. Specifying a sufficient number of topological features will eventually define a network uniquely. If we meaningfully choose our topological features to relate to our understanding of the processes which formed the network, then the associated ensemble is interesting in its own right.

In this paper we will call such a family of networks a ‘Tailored Random Graph Ensemble’ (TGE). We study several aspects of TGEs. Results are derived analytically, in leading order in $N$. We look to measure the size of the ensemble via its Shannon Entropy $S$ per bond. This allows us to quantify the topological properties we have imposed from the point of view of how severely they restrict the TGE. We can also compare two TGEs via a Kullback-Leibler distance. Our approach gives a rigorous and quantitative basis to compare two networks based on their topological properties.

Finally, we consider how we might sample without bias from a TGE. To achieve this, we begin by drawing attention to (and providing a correction) to a bias in a common graph randomisation algorithm. We then show that our approach can be generalised to sample networks with desired topological properties.
2. Quantifying the topology of a network via tailored random graph ensembles

[1] and [2] derive closed form solutions for the Shannon entropy per bond

\[ S = -\frac{1}{N\langle k \rangle} \sum_c p(c) \log p(c) \]

of the maximum entropy ensemble associated with particular constraints, where \( k_i \) is the degree (number of links) of node \( i \), \( \langle k \rangle \) is the average number of links per node, \( p(c) \) is the degree distribution and \( N \) is the number of nodes. We choose to describe degree-degree correlation via the degree statistics of connected nodes

\[ W(k, k' | c) = \frac{1}{N\langle k \rangle} \sum_{ij} C_{ij} \delta_{k_i, k(c)} \delta_{k'_j, k(c)} \]

We can define the maximum-entropy ensemble of non-directed graphs constrained by a degree distribution \( p(k) \) and a degree-degree correlation \( W(k, k') \) by setting the probability of observing a network \( c \) to be

\[ p(c) = \frac{1}{Z} \prod_i [d k_i p(k_i)] \prod_{i<j} \frac{W(k_i, k_j)}{\langle k \rangle} \delta_{c_{ij}, 1} + \left( 1 - \frac{\langle k \rangle W(k_i, k_j)}{\langle k \rangle} \right) \delta_{c_{ij}, 0} \]

The partition function \( Z \) is defined in the usual way. \( \pi(k) \) is the Poissonian distribution with the same average degree, and \( W(k) \) indicates the marginal. In this case [1] find the Shannon entropy per bond of this ensemble to be

\[ S = \frac{1}{2} \left[ 1 + \log(N/\langle k \rangle) \right] - \frac{1}{\langle k \rangle} \sum_k p(k) \log \left( \frac{p(k)}{\pi(k)} \right) - \frac{1}{2} \sum_{k,k'} W(k, k') \log \left( \frac{W(k, k')}{W(k)W(k')} \right) + \epsilon_N \]

The ensemble constrained by the degree distribution only can be found by substituting \( W(k, k') = p(k)p(k') \) into equation 1. Substituting this into equation 2 can be easily demonstrated to be equivalent to removing the last term. That is, the three terms of equation 2 can be associated with the progressive tightening of the constraint on the ensemble from average degree, then to degree distribution, then additionally to degree-degree correlation.

[2] extended [1] to find a similar expression in the case of directed graphs.

3. Structural dissimilarity of graphs \( c_A \) and \( c_B \), based on information-theoretic distance between associated ensembles

[1] uses the same methods as were used to derive equation 2 to determine a Kullback-Leibler distance between ensemble \( A \) (defined through \( p_A \) and \( W_A \)) and ensemble \( B \) (defined through \( p_B \) and \( W_B \))

\[ D_{AB} = \frac{1}{2N} \sum_{c \in G} \left\{ p(c|p_A, W_A) \log \left( \frac{p(c|p_A, W_A)}{p(c|p_B, W_B)} \right) + p(c|p_B, W_B) \log \left( \frac{p(c|p_B, W_B)}{p(c|p_A, W_A)} \right) \right\} \]
This is evaluated to be

\[
D_{AB} = \frac{1}{2} \sum_k p_A(k) \log \left( \frac{p_A(k)}{p_B(k)} \right) + \frac{1}{2} \sum_k p_B(k) \log \left( \frac{p_B(k)}{p_A(k)} \right) \\
+ \frac{1}{4(k)_A} \sum_{kk'} p_A(k) p_A(k') \Pi_A(k, k') \log \left( \frac{\Pi_A(k, k')}{\Pi_B(k, k')} \right) \\
+ \frac{1}{4(k)_B} \sum_{kk'} p_B(k) p_B(k') \Pi_B(k, k') \log \left( \frac{\Pi_B(k, k')}{\Pi_A(k, k')} \right) \\
+ \frac{1}{2} \sum_k p_A(k) \log \rho_{AB}(k) + \frac{1}{2} \sum_k p_B(k) \log \rho_{BA}(k)
\]

with

\[
\Pi(k, k') = W(k, k)/W(k)W(k'), \\
\rho_{AB}(k) = \sum_{k'} \Pi_B(k, k')W_A(k')\rho_{AB}^{-1}(k')
\]

The directed case is covered in [2].

4. Constrained Markovian graph dynamics

This section will demonstrate the existence of a bias which applies to the commonly used edge-swapping algorithm for edge-preserving graph randomisation. This bias can be corrected, and a natural generalisation of the correction allows us to sample networks with more finely tailored topological properties.

The aim is to sample networks from a subset \( G[\star] \subseteq G \) without bias, where \( G \) is the space of all networks and \( G[\star] \) is the subset that satisfies some constraint \( \star \). To do this, we look to define a Markov Chain process with transition probabilities \( W(c|c') \) such that \( p_{t+1}(c) = \sum_{c' \in G[\star]} W(c|c') p_t(c') \) based on some set of auto-invertible allowed moves \( \Phi \) (excluding identity) such that \( \Phi \) is ergodic on \( G[\star] \) and the process converges to \( p_\infty(c) = Z^{-1} e^{-H(c)} \) on \( G[\star] \) for some target \( H(c) \). Setting \( H(c) = \text{const} \) would sample each member of the set with equal probability. Choosing a different form of \( H(c) \) allows us to favour certain networks.

We can express our transition probability as

\[
W(c|c') = \sum_{F \in \Phi} q(F|c') \left[ \delta_{c,Fc} A(Fc'|c') + \delta_{c,c'} [1 - A(Fc'|c')] \right]
\]

where \( A(c|c') \) is the move acceptance probability and \( q(F|c) \) is the move proposal probability. If all allowed moves \( F \) are equally probable \( q(F|c) \) can be written as \( q(F|c) = I_F(c)/n(c) \) where \( I_F(c) \) indicates if \( F \) is a valid move on \( c \), and \( n(c) = \sum_{F \in \Phi} I_F(c) \) is the number of allowed moves on \( c \).

As a sufficient condition for convergence we impose a detailed balance condition. This is equivalent to

\[
(\forall F \in \Phi)(\forall c \in G[\star]) : \quad q(F|c)A(Fc|c)e^{-H(c)} = q(F^{-1}|Fc)A(c|Fc)e^{-H(Fc)}
\]

It then follows that

\[
(\forall F \in \Phi)(\forall c \in G_F[\star]) : \quad \frac{1}{n(c)} A(Fc|c)e^{-H(c)} = \frac{1}{n(Fc)} A(c|Fc)e^{-H(Fc)}
\]

which is satisfied by

\[
A(c|c') = \frac{n(c') e^{-\frac{1}{2}(H(c) - H(c'))} + n(c) e^{\frac{1}{2}(H(c) - H(c'))}}{n(c') e^{-\frac{1}{2}(H(c) - H(c'))} + n(c) e^{\frac{1}{2}(H(c) - H(c'))}}
\]
A common example would be to sample from the set of networks with a certain degree sequence by defining a process based on degree-preserving edge swaps [3]. Most common implementations of this scheme accept all proposed moves, i.e. $A(c|c') = \text{const}$. This corresponds to $H(c) = -\log n(c)$, so would give $p_\infty(c) = \frac{n(c)}{\sum_{c' \in \mathcal{G}} n(c')}$ - a sampling bias.

To sample without bias from the ensemble of all networks which satisfy the constraint $\star$, it is necessary to set $H(c) = \text{const}$, which leads to

$$A(c|c') = \frac{n(c')}{n(c') + n(c)}$$

Suppose, however, we wanted to study an ensemble which was also tailored to certain soft constraints, such as was seen in equation 1. Using the relation $p_\infty(c) = Z^{-1}e^{-H(c)}$, we can do this by setting

$$H(c) = -\sum_{i<j} \log \left[ \frac{(k)}{N} \frac{W(k_i, k_j)}{p(k_i)p(k_j)} \delta_{c_{ij}, 1} + \left( 1 - \frac{(k)}{N} \frac{W(k_i, k_j)}{p(k_i)p(k_j)} \right) \delta_{c_{ij}, 0} \right]$$

[4] and [5] derive practical expressions for $n(c)$ for the most commonly seen degree-preserving process based on swapping pairs of edges.

5. Discussion
In this paper we have summarized some recently developed tools for studying the topology of directed and non-directed binary networks.

Firstly, section 2 provides exact closed form expressions which allow us to measure topological properties with reference to the Shannon entropy of the associated constrained maximum entropy ensembles. In a particular example, calculating a large value indicates that the constraint is not very restrictive, whereas finding a low value indicates that the constraint is strong, and networks with this collection of properties are more rare in the space of all graphs. Section 3 provides expressions that can be used to compare how similar the topology of two networks is based on the Kullback-Leibler distance between the associated ensemble. This provides a far more natural and intuitive basis for comparing networks than the more commonly seen Hamming distance.

Section 4 discusses a complementary approach to studying the topology of networks: how do we generate an example network from our ensemble in an unbiased way? This is an important consideration for numerical work, for example in comparing the frequency of motifs in a real network compared to a suitably randomised test network. We show that one of the most commonly used approaches to randomising networks is biased, and provide a simple correction to eliminate this bias.

6. Acknowledgements
Author ESR thanks BBSRC for financial support.

References
[1] A. Annibale, A. C. C. Coolen, L. P. Fernandes, F. Fraternali, and J. Kleinjung. Tailored graph ensembles as proxies or null models for real networks I: tools for quantifying structure. J. Phys. A, 42(48):485001, 2009.
[2] E S Roberts, T Schlitt, and A C C Coolen. Tailored graph ensembles as proxies or null models for real networks ii: results on directed graphs. J. Phys. A, 44(27):275002, 2011.
[3] R Taylor. Combinatorial Mathematics VIII. Springer, 1981.
[4] A. C. C. Coolen, A. De Martino, and A. Annibale. Constrained Markovian dynamics of random graphs. J. Stat. Phys., 2009.
[5] E. S. Roberts and A. C. C. Coolen. Unbiased degree-preserving randomization of directed binary networks. Physical Review E, 85:046103+, 2012.