THE EXCHANGEABLE GRAPH MODEL

Edoardo M. Airoldi, Harvard University *(airoldi@fas.harvard.edu)*

Extended Abstract. Mapping connectivity patterns in a graph onto the space of binary strings via the simplest possible model would enable comparisons between different statistical models of pairwise measurements. Such a strategy would also leads to calculations for assessing the statistical significance associated with the observed overlap between two cliques in a graph. The *exchangeable graph model* maps a graph over N nodes to a set of node-specific binary strings, to support these two analyses in practice (Airoldi, 2006).

The data generating process instantiating an exchangeable graph model for a matrix of binary observations on pairs of N nodes is specified as follows,

1. For each node in $n \in \mathcal{N}$

   1.1. Sample node-specific binary strings $\vec{b}_n \sim Uniform$ (vertex set of K-hypercube),

2. For node pair $n, m \in \mathcal{N} \times \mathcal{N}$

   2.3. Sample the binary physical binding event $x_{nm} \sim Bernoulli\ \big(\ q(\vec{b}_n, \vec{b}_m)\ \big)$,

where $\vec{b}_{1:N}$ are binary strings K-bit long, and q is function that projects binary strings into the $[0, 1]$ interval. This generating process leads to weakly dependent edges; the edges are conditionally independent given their binary string representations, technically they are *exchangeable*. In this sense, an *exchangeable graph model* provides the minimal step-up in complexity from the random graph model (Erdös and Rényi, 1959; Gilbert, 1959).

Briefly, the number of bits captures the complexity of a graph. For instance, for $K < N$ the model provides a parsimonious representation of the graph. For directed graphs the function q is asymmetric in the arguments. The sparsity of the bit strings can be controlled with a hierarchical construction based on a distribution on the unit hypercube (Airoldi, 2009). In an exchangeable graph model there are two main sources of variability: (i) the probability of an edge decreases with the number of bits K, as more complexity reduces the chances of an edge, and (ii) the probability of an edge increases with $1/\alpha$, as concentrating density in the corners of the unit K-hypercube improves the chances of an edge. While this model does not quite fit the definition of non-homogeneous models of Bollobás et al. (2007), it is tractable enough to allow the analysis of the giant component, albeit approximately, by leveraging the branching process strategy similar to the one developed by Durrett (2006). As in Durrett's analysis, the giant component emerges because a number of smaller components must intersect with high probability. In addition, the giant component has a peculiar structure in exchangeable graph models; connected components are themselves connected to form the giant component a soon as bit-strings that match on two bits appear with high probability. For an illustration see figure 1, where nodes that *bridge* two connected components are evident in the left panel. In the Figure, there are no nodes that bridge three components, as having bit-strings that match on three bits is an unlikely in this parameter setting.

In practice, given a graph we can infer the corresponding set of binary strings from data. The likelihood that correspond to an exchangeable graph model is simple to write,

$$\ell(Y|\alpha) = \int d\,\vec{b}_{1:N}\ \Big(\ \prod_{n,m} Pr\ (Y_{n,m}|\vec{b}_n, \vec{b}_m, q) \prod_{n} Pr\ (\vec{b}_n|\alpha)\ \Big),$$

and we can apply sampling or variational inference techniques (Airoldi, 2007).

The exchangeable graph model allows to assess the complexity of an observed graph leveraging notions in information theory. For instance, we can use MDL (i.e. the minimum description length principle) to decide how many bits we need to explain the observed connectivity patterns in a graph, with high probability. We can also quantify how much *information* is retained at different bit-lenghts, and plot the corresponding *information profile* for $K < N$, and an *entropy histogram* for any given value of K.

The exchangeable graph model allows comparison of any set of statistical models that are proposed to summarize an observed graph. As an illustration, consider an observed graph G and two alternative models A and B. Rather

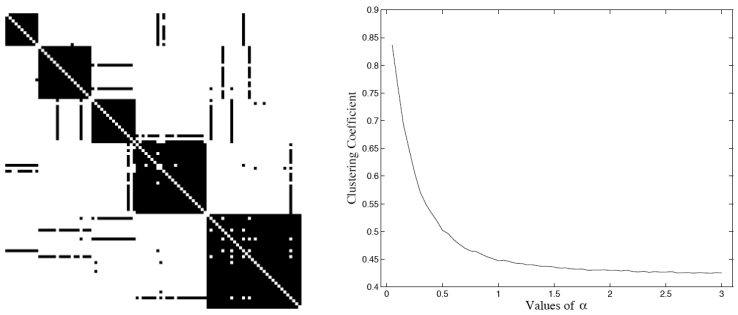

Figure 1: *Left.* An example adjacency matrix that correspond to a fully connected component among 100 nodes. *Right.* The clustering coefficient as a function of α on a sequence of graphs with 100 nodes.

than comparing how well models A and B recover the degree distribution of G, or any other set of graph statistics, and independently of whether it makes sense of not to directly compare the two likelihoods of A and B (in fact, these models need not have a likelihood), we can proceed as follows.

1. Given a graph G, fit models $A(\Theta_a)$ and $B(\Theta_b)$ to obtain an estimate of their parameters.
2. Sample M graphs at random from the support of $A(\Theta_a^{Est})$ and $B(\Theta_b^{Est})$.
3. Compute the distributions of summary statistics based on notion from information theory, such as information profile and entropy histogram, corresponding to the $2M$ graphs sampled from A and B.
4. Compare models in terms of the distribution on the statistics above, such as the complexity of the two models' supports, the similarity between the complexity of G and the models' complexity, and so on.

Last, the exchangeable graph model allows to evaluate the distribution of the number of bit-strings with I matching bits, for any integer $I < K$. From a theoretical perspective, this distribution leads to expectations on the number of nodes that bridge I communities, where the members of each community have only one out of I matching bits. In practice, we may want to specify K in advance so that each bit corresponds to a well defined property. For instance, in applications to biology nodes may correspond to proteins and the K bits encode presence/absence of specific protein domains. The distribution on the number of I matchings leads to p-values that summarize how unexpected it is to observed binding events among a set of proteins that share a certain combination of domains.

Overall, the exchangeable graph model introduces weak dependence among the edges of a random graph in a controlled fashion, which ultimately leads to a range of more structured connectivity patterns and enables model comparison strategies rooted in notions from information theory. The focus here is not on modeling per-se. In fact, the model is kept as simple as possible. Rather, the exchangeable graph model provides a bridge between graph connectivity and node attributes to support graph model comparison and significance analysis of communities overlap.

References.

E. M. Airoldi. *Bayesian mixed membership models of complex and evolving networks*. PhD thesis, School of Computer Science, Carnegie Mellon University, December 2006.

E. M. Airoldi. Getting started in probabilistic graphical models. *PLoS Computational Biology*, 3(12):e252, 2007.

E. M. Airoldi. A family of distributions on the unit hypercube. Technical report, Harvard university, Department of Statistics, March 2009.

B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

R. Durrett. *Random Graph Dynamics*. Cambridge University Press, 2006.

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

E. N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30:1141–1144, 1959.