Relational models for generating labeled real-world graphs

Christoph A. Lippert, Nino Shervashidze, Oliver Stegle, Karsten M. Borgwardt
Max Planck Institute for Developmental Biology
Max Planck Institute for Biological Cybernetics
Spemannstraße 41, 72076 Tübingen, Germany
[firstname].[lastname]@tuebingen.mpg.de

Abstract

Analyzing and understanding the structure of social networks and other real-world graphs has become a major area of research in the field of data mining. An important problem setting is the creation of realistic synthetic graphs that resemble real-world social networks. While a range of efficient algorithms for this task have been proposed, current methods solely take the network topology into account ignoring any node labels. By applying concepts from relational learning we propose a probabilistic approach to synthetic graph generation *with node labels*.

1 Generation of real-world graphs

There are various motivations for the generation of synthetic graphs. Here, we are interested in two types of graph generation: First, given a graph G, we would like to be able to perform graph anonymization, that is to generate a graph G_1 of the same size (number of nodes) as G. The objective in graph anonymization is that G_1 share topological properties and exhibits similar node labels as the original graph G. Second, given a graph G, we would like to synthesize a larger graph G_2 of size $>> G$ that is similar to G. Again this synthesized graph shall be similar in terms of topology and labels and additionally exhibit typical properties of large real-world networks, such as a small diameter.

Traditional models for generating graphs are rather simple. For example the Erdös-Renyi model [1] only has the edge probability p_e as a single parameter. Another prominent graph generation model is preferential attachment [2], in which new nodes prefer to attach to existing nodes with high degree. A more recent approach, KronFit [3], is based on fitting an $N_1 \times N_1$ probabilistic initiator matrix Θ to the original graph and approximating it by iteratively computing the Kronecker product of Θ with itself. However, as in real-world graphs node labels usually correlate with topology, it would be desirable to have models that consider the generation of *labeled* graphs.

Our approach towards labeled graph generation builds upon concepts from relational learning. A starting point for our method development is the Infinite Relational Model (IRM) [4, 5].

2 Infinite Relational Model

The underlying principle of this family of models is to infer a block stochastic model of graph structure. The goal is to partition relations in an observed network by assigning nodes to clusters. Nodes that share a similar connectivity structure and similar labels are grouped together in the same clusters which leads to an informative representation of the underlying network structure. The IRM allows for an arbitrary number of clusters by deploying a Dirichlet process on the cluster variable z. The probability of a relation $R_{i,j}$ between two nodes i and j is entirely determined by their cluster membership

$$p(R_{i,j}|z_i, z_j) = \text{Bernoulli}(R_{i,j}|\eta(z_i, z_j)), \tag{1}$$

where $R_{i,j}$ is the relation status between node i and j, either exhibiting a link (true) or not (false). In an IRM, the prior probability of $\eta(z_i, z_j)$ only depends on a shared Beta distributed prior, which is identical for

Categories: graphical models, data mining
Presented by: Christoph A. Lippert (preference: oral)

all pairs of clusters

$$\eta(a,b) \sim \text{Beta}(\beta_1, \beta_2), \tag{2}$$

where a and b represent two clusters and β_1, β_2 are hyperparameters of the Beta distribution (Beta), hence influencing how likely relations between clusters are.

We represent node labels as N_f independent binary features, attached to every node i, $\mathbf{F}_i = \{F_{i,1}, \ldots, F_{i,N_f}\}$. The features of node i are Bernoulli distributed

$$P(\mathbf{F}_i | z_i) = \prod_{f=1}^{N_f} \text{Bernoulli}\left(F_{i,f} | \theta(z_i)_f\right), \tag{3}$$

where similar to the relation probabilities, the feature probabilities $\theta(z_i)_f$ depends on the cluster assignment. Again a beta prior is put on the feature probabilities

$$\theta(z_i)_f \sim \text{Beta}(\Theta_1^f, \Theta_2^f), \tag{4}$$

which is chosen to reflect the data statistics, i.e. how many nodes overall have a specific feature set on or off.

Once the cluster distribution fits the original graph, a random graph can be generated by straightforwardly sampling labeled nodes and relations from the resulting model. The block structure of a network drawn from the IRM nicely resembles community structures present in the original graph. However, due to the fact that the between cluster edge probabilities $\eta(a,b)$ are sampled mutually independent for each pair of clusters a and b, artificial graphs generated by the IRM do not capture other statistical patterns of connectivity present in real-world graphs. For example, to realistically model the degree distribution it may be desirable to have clusters that account for the existence of a small number of nodes with high degrees (hubs) and a larger number of nodes with low degrees.

3 Infinite Network Model

In order to better model global connectivity patterns of real-world graphs, we propose the Infinite Network Model (INM) which generalizes the IRM such that every cluster carries an individual "connectivity prior" in form of a Beta distribution. We assume that the probability of a relation between any two clusters a and b calculates as the product of the two respective Beta distributions.

$$\eta(a,b) \sim \text{Beta}(\beta_1^a, \beta_2^a)\text{Beta}(\beta_1^b, \beta_2^b). \tag{5}$$

To complete this extra level of hierarchy, we put Gamma priors on the beta parameters of every cluster a.

$$\beta_1^a \sim \Gamma(k_1, s_1), \qquad \beta_2^a \sim \Gamma(k_2, s_2), \quad \forall \,\text{clusters } a. \tag{6}$$

As a result of this connectivity prior, the model not only describes the interaction probability between two clusters, but also whether members of a cluster are more or less likely to form links to any other cluster in the network. That means that the INM has to ability to describe clusters of low or high connectivity, and hence low or high degrees.

In extensive experimental evaluation using real world graphs of different sizes we demonstrate the ability and performance of the INM model. Graphs generated by the INM show realistic topological properties, such as degree distribution and graph spectrum, as well as node labels capturing the original graph. The additional flexibility introduced by the INM yields a significant improvement that is most pronounced when degrees and labels in the original graph exhibit correlation. In comparison to existing approaches our model outperforms state-of-art methods for both, graphs with and without node labels.

References

[1] P. Erdös and A. Renyi. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Science*, 5:17–67, 1960.

[2] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[3] Jure Leskovec and Christos Faloutsos. Scalable modeling of real graphs using kronecker multiplication. In *ICML*, pages 497–504, 2007.

[4] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, 2006.

[5] Zhao Xu, Volker Tresp, Kai Yu, and Hans-Peter Kriegel. Infinite hidden relational models. In *UAI*, 2006.