Relational models for generating labeled real-world graphs

Christoph Lippert CHRISTOPH.LIPPERT@TUEBINGEN.MPG.DE
Nino Shervashidze NINO.SHERVASHIDZE@TUEBINGEN.MPG.DE
Oliver Stegle OLIVER.STEGLE@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics, Tübingen, Germany
Max Planck Institute for Developmental Biology, Tübingen, Germany

Keywords: synthetic graph generation, statistical relational learning, infinite relational models

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. We propose a probabilistic approach to synthetic graph generation *with node labels*, building on concepts from relational learning.

1. Generation of real-world graphs

Given a graph G, we would like to be able to perform graph anonymization, that is to generate a graph G' of the same size (number of nodes) as G. The objective in graph anonymization is that G' shares topological properties and exhibits similar node labels as the original graph G.

Traditional models for generating graphs are based on simple graph statistics. For example the Erdös-Renyi model (Erdös & Renyi, 1960) only has the edge probability p_e as a single parameter, that is learned from data. Another prominent graph generation model is preferential attachment (Barabasi & Albert, 1999), a model that focuses on the degree of nodes and new nodes prefer to attach to existing nodes with a high degree. ForestFire (Leskovec et al., 2007) is a more recent, stepwise graph generation procedure that follows three main steps every time a new node v is added to the graph: 1) v forms a link to an existing node w, 2) v randomly selects a subset $N_{sub}(w)$ of the neighbours of w, and 3) then repeats step 2) for the nodes in $N_{sub}(w)$. Another recent approach, KronFit (Leskovec & Faloutsos, 2007), 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 approach is the Infinite Relational Model (IRM) (Kemp et al., 2006; Xu et al., 2006).

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. Employing a Dirichlet process prior on the cluster assignments, the IRM allows for an unbounded number of clusters. The probability of a relation $R_{i,j}$ between two nodes i and j is entirely determined by their cluster membership z_i and z_j.

$$P(R_{i,j} \mid z_i, z_j) = \text{Bernoulli}(R_{i,j} \mid \eta(z_i, z_j)), \quad (1)$$

where $R_{i,j}$ is the relation status between node i and j, either exhibiting a link (true) or not (false). In a traditional IRM, the prior probability of $\eta(z_i, z_j)$ solely depends on a global Beta prior that is shared among all clusters

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

Presented at *ILP-MLG-SRL*, Leuven, Belgium, 2009.

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

We consider 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

$$P(\mathbf{F}_i \mid z_i) = \prod_{f=1}^{N_f} \text{Bernoulli}\left(F_{i,f} \mid \theta_f(z_i)\right), \qquad (3)$$

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

$$\theta_f(z_i) \sim \text{Beta}(\Theta_1^f, \Theta_2^f), \qquad (4)$$

which is chosen to reflect the data statistics, i.e. how many nodes in total have a specific feature set or not.

Cluster distributions and hyperparameters are learned on a training graph. Subsequently random graphs with labels can be generated from the trained model by forward sampling from the network 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 is desirable to allow for clusters that account for 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) that 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 is given by a Beta prior taking pseudo counts from both respective clusters into account

$$P(\eta_{a,b}) \sim \text{Beta}(\eta_{a,b} \mid \beta_1^a + \beta_1^b, \beta_2^a + \beta_2^b). \qquad (5)$$

To complete the definition of this extra level of hierarchy, we put Gamma priors on the beta parameters

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

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. Hence the INM can describe clusters of low or high connectivity corresponding to low or high degrees.

4. Experiments

In our experiments, we generate synthetic graphs from three real world graphs with binary labels. **Countries** (Wasserman & Faust, 1994) is a network in which the 24 nodes correspond to countries and edges represent economic or diplomatic relationships between countries. The node labels indicate whether a country belongs to the richer half of countries (1, high GNP) or to the poorer half of countries (0, lower GNP). **Enron** (Klimt & Yang, 2004) is an email traffic network in which the 184 nodes are Enron employees and links connect employees that had email correspondence. Node labels indicate whether an employee belongs to the management level of the company (1) or not (0). In **Blogosphere** (Adamic & Glance, 2005) the 1490 nodes are political blogs and edges represent hyperlinks between them. Node value attributes indicate political leaning: left/liberal (0) or right/conservative (1).

We trained the INM on a given graph G to learn its parameter settings and then generate a new synthetic graph G' of size n. As a comparison we also performed graph generation using the IRM, KronFit and Forest-Fire. In order to simulate the generation of labeled graphs we performed random shuffling of the original node labels and randomly redistributed the labels on the synthetic graphs. We measured the ability of the methods to generate graphs that approximate the original graph in terms of the following graph properties:

In-(out-)degree distribution: the distribution of the number of incoming (outgoing) edges for all nodes in the graph.
Hop plot: the distribution of shortest distances (number of steps) between any two nodes in the graph.
Scree plot: the distribution of singular values of the adjacency matrix of the graph.
Diameter of the graph and the **effective diameter** of the graph: the maximum shortest distance between any two nodes in the graph and the 90% quantile of all shortest path distances, respectively.

Beside these purely topological criteria, we also computed a property of the synthetic graph which reflects its similarity to the original graph both in terms of topology and node labels. This was achieved by using the *graphlet kernel* (Shervashidze et al., 2009) that counts matching connected, induced subgraphs of 4

nodes in two graphs. This graphlet kernel k_g can be turned into a distance via

$$d_g(G, G') = \sqrt{k_g(G, G) + k_g(G', G') - 2k_g(G, G')},$$

where G is the original graph and G' is the synthetic graph in our case.

Results We report results for all methods in Table 1. Results are better, the smaller the entries in a table, except for the diameter. The INM outperformed KronFit in all cases except for in-degree distribution on Countries and its out-degree distribution on Blogs. The INM achieved better results than Forest Fire, except for the effective diameter, the right singular value, and the scree plot on Blogs. Also the IRM showed better results than KronFit and ForestFire in most of the tasks. Most interesting among our evaluation measures is the graphlet kernel, as it not only assesses the topological similarity to the original graph, but also the node label similarity. Here INM and IRM were also better than KronFit and ForestFire, both with random shuffling of original node labels, on all datasets. The INM was closer to the original graph than the IRM on Countries and ENRON, but worse on the Blogs dataset. This is in fact consistent with the results for the other topological features, where the INM is most of the time best on Countries and ENRON, but often worse than the IRM on Blogs. Looking at the Pearson's correlation coefficient between node labels and degrees on the three datasets (see Table 1) one can see that unlike in the other datasets there is almost no correlation in Blogs. As the INM is based on the assumption that node labels, cluster membership and degree of a node are all correlated it shows better results on Countries and ENRON where the assumption holds, while it performs slightly worse on Blogosphere.

5. Discussion

We present a relational learning approach to synthetic graph generation, given a reference graph. Unlike the numerous other approaches to graph synthesis, our method is designed to produce graphs with node labels. The additional flexibility introduced by the INM yields a consistent improvement over the IRM when degrees and labels in the original graph exhibit correlation. Synthetic graphs generated by the INM approximate both the topology and the labels of the reference graph well, outperforming the state-of-the-art.

References

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 us election. *WWW Workshop on the Weblogging Ecosystem.*

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.

Erdös, P., & Renyi, A. (1960). On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Science*, *5*, 17–67.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. *AAAI.*

Klimt, B., & Yang, Y. (2004). The enron corpus: A new dataset for email classification research. *ECML.*

Leskovec, J., & Faloutsos, C. (2007). Scalable modeling of real graphs using Kronecker multiplication. *ICML.*

Leskovec, J., Kleinberg, J. M., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *TKDD*, *1.*

Shervashidze, N., Vishwanathan, S., Petri, T., Mehlhorn, K., & Borgwardt, K. (2009). Effiient graphlet kernels for large graph comparison. *AISTATS.*

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.*

Xu, Z., Tresp, V., Yu, K., & Kriegel, H.-P. (2006). Infinite hidden relational models. *UAI.*

in-degree	Countries	ENRON	Blogs
INM	0.300	0.085	0.073
IRM	0.337	0.105	0.037
KronFit	0.292	0.175	0.333
ForestFire	0.750	0.331	0.176
out-degree	Countries	ENRON	Blogs
INM	0.187	0.090	0.082
IRM	0.246	0.097	0.061
KronFit	0.375	0.134	0.042
ForestFire	0.583	0.428	0.382
hop plot	Countries	ENRON	Blogs
INM	0.029	0.055	0.039
IRM	0.072	0.065	0.020
KronFit	0.054	0.242	0.235
ForestFire	0.453	0.059	0.658
scree plot	Countries	ENRON	Blogs
INM	0.034	0.040	0.092
IRM	0.031	0.044	0.071
KronFit	0.088	0.071	0.099
ForestFire	0.086	0.107	0.053
diameter	Countries	ENRON	Blogs
Original graph	2	4	8
INM	2	4	6
IRM	2	4	7
KronFit	2	3	6
ForestFire	4	4	12
$d_{graphlet}$	Countries	ENRON	Blogs
INM	0.114	0.059	0.327
IRM	0.141	0.084	0.086
KronFit+RS	0.240	0.091	0.434
ForestFire+RS	0.239	0.181	0.432
Correlation	Countries	ENRON	Blogs
(degree, labels)	0.46	0.43	0.01

Table 1. Distance to the original graph in terms of Kolmogorov-Smirnoff statistic for in-(out-)degree, hop plot, scree plot; Diameter (effective diameter) of the original and the synthetic networks; distance induced by the graphlet kernel (RS refers to random shuffling); correlation between degree and node labels in the original network.