Graph representation using mutual information for graph model discrimination

Francisco Hawas
Applied Mathematics and Statistics
Stony Brook University
Stony Brook, USA
francisco.hawas@stonybrook.edu

Petar M. Djurić
Electrical and Computer Engineering
Stony Brook University
Stony Brook, USA
petar.djuric@stonybrook.edu

***Abstract*—We present a novel approach of graph representation based on mutual information of a random walk in a graph. This representation, as any global metric of a graph, can be used to identify the model generator of the observed network. In this study, we use our graph representation combined with Random Forest (RF) to discriminate between Erdös-Renyi (ER), Stochastic Block Model (SBM) and Planted Clique (PC) models. We also combine our graph representation with a Squared Mahalanobis Distance (SMD)–based test to reject a model given an observed network. We test the proposed method with computer simulations.**

***Index Terms*—Network Topology, Graph Theory, Complex Networks, Mutual Information.**

I. Introduction

Graph structures allow us to explore how different entities interact with each other; they give us another layer of information to the usual vector of individual characteristics of independent entities in the system. This new layer helps us understand how connections emerge and disappear and/or how different individual actors interact between each other given their distinct attributes.

In this study we introduce a graph model representation tool based on mutual information of a random walk on a graph. This allows us to describe each graph structure with a two dimensional vector that, as we show, can be used to discriminate between models or reject a model. As observed by [9], the decay in mutual information of a random walk in a graph carries a signal regarding the structure of the network. Our approach tries to exploit this finding in order to obtain meaningful information regarding the structure of the observed graph.

There are various approaches that use graph representations to classify graphs. In [3], the authors propose a two-step procedure and use global characteristics. In their first step, they compute known graph features such as degree centrality, betweeness centrality, closeness centrality and others, and in the second step, they classify the graphs using Random Forests (RF). As in [3], the method from [6] is also a two-step procedure. The authors introduce Walk2Vec and Walk2Vec-SC, which exploit random walks, sparse coding and pooling to generate graph features, which are then classified using RF. In [10], the spectrum of a graph using different graph representations is analyzed. The representations include Laplacian, normalized Laplacian, signless Laplacian, adjacency matrix and heat kernel. The graph spectra are classified by neural networks. In [8], the authors combine the spectral density of a graph and the Kullback-Leibler divergence in order to obtain the best model within a certain family of network models. They also use the Jensen-Shannon divergence to measure distance between graphs structures and to test whether or not two models have zero distance between them. Unlike [10], in [8], the graph representation is the adjacency matrix only and one computes the spectrum density from it. In [7], the authors expose the structure of the graph by the use of a combination of operations on the adjacency matrix, which is transformed to attributes that are then classified using Support Vector Machines (SVM). In [5], as in [3], the classification is based on computed features of a graph. Specifically, graphlet counts are exploited to characterize different structures. The tests in the paper are mostly for sparse graphs generated from a preferential attachment variant. In [1], the authors include 47 measures that combine local and global graph characteristics. Again, these measures are applied to characterize different graph structures.

Our contribution is to develop a new global graph representation measure which carries enough information to discriminate between models. Thus, it can be used as a classification method. Further, our graph representation can be used to reject a model when implemented with standard outlier detection techniques.

The paper is organized as follows. In the next section we provide the problem statement. In Section III, we describe the proposed method, and in Section IV, we describe results of the method. We conclude the paper with Section V.

II. Problem Statement

In many cases, in network science, available data are static, that is, we do not know how an observed graph was formed. In other words, we do not know the order of arrival of edges nor nodes. Thus, we only have a graph at hand and would like to know if a particular mechanism (model) has generated the graph. So basically, here we deal with significance testing.

The authors thank the support of the National Science Foundation under Award CCF-1618999.

ISBN 978-90-827970-1-5 © EURASIP 2018

At the end, we either reject the model or do not reject it. Further, we may have several mechanisms as candidates for generative models. In other words, we deal with a model selection, where the task is to select the best model from the considered set. If we identify the correct model, we will be able to obtain samples of graphs that share the same structure, i.e., we would be able to generate networks of the same size and same characteristics as the one that was observed.

III. Proposed Method

Our method starts with sampling a graph with the tested model. Once we sample from each model, we map each graph, as described by its adjacency matrix, to a vector in $\mathbb{R}^2$. This vector contains information regarding the structure of the graph. Finally, depending on the task, this 2-dimensional representation is used in a classification procedure to determine the best available model for the observed data, or it is used to reject a specific model believed to be the generator of the observed network. We show two applications that use our graph representation, one for model selection and the other for significance testing. Although our ideas could be extended to directed and weighted networks, we are not going to explore these cases here.

Given a graph adjacency matrix A of a graph of size $N \times N$, undirected and unweighted, we compute the transition matrix $T = D^{-1}A$ where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{N} A_{ij}$. We use this transition matrix to define a random walk on the nodes of the graph. The initial state of the random walk is denoted $x(0) \in \{1, ..., N\}$. In our case, the random walk starts at node $i \in \{1, ..., N\}$ with probability $\frac{1}{N}$. After t steps, the random walk is in node $x(t) \in \{1, ..., N\}$.

We compute the mutual information between $x(0)$, the initial state, and $x(t)$ as in [9], i.e.,

$$I(t) = I(x(0), x(t)) = \sum_{i=1}^{N} p_i \sum_{j=1}^{N} P_{ij}^{t} \log\left(\frac{P_{ij}^{t}}{p_j^{t}}\right), \quad (1)$$

where $p_i = \frac{1}{N}$ where $i = 1, 2, \ldots, N$ corresponds to the distribution of the starting node of the random walk $x(0)$, $p_j^t = \sum_{i=1}^{N} p_i P_{ij}^t$ is the unconditional probability of the random walk to be at node j after t steps, and $P_{ij}^t = (T^t)_{ij}$ is the conditional probability that given the random walk starts in node $x(0) = i$, the process is in node $x(t) = j$ after t steps.

In [9], the authors use the curve $\{I_t(t)\}_{t=1,...,t^*}$, in particular, they rely on the derivative as a way to guide their clustering algorithm because it contains information regarding the structure of the graph. Our hypothesis is that this curve could contain information about the graph model used to generate the observed network and we want to exploit it by fitting a parametric function with a small number of parameters to describe it. In the end, these parameters end up being our graph representation.

In [9], the authors stated that for a convergent Markov process $\lim_{t\to\infty} I_t = 0$ and that as $t \to \infty$, I_t decays monotonically. Based on these results, we propose the following simple parametric function for $I(t)$:

$$f(t) = ae^{-bt} \quad (2)$$

For a specific graph, the parameters (a, b), we hypothesize, carry information on the model generator of a graph. We would like to be able to derive a theoretical relationship between a specific random graph model and the corresponding distribution of (a, b). Since we do not have this relationship, we sample the parameters in question by drawing for each graph model a sample of size S and compute $I(t)$ for $t = 1, ..., t^*$. Using the sample distribution of (a, b), we perform significance testing or model selection.

A. Simple Example

To clarify the main idea, we generate a small Erdős-Rényi (ER) random graph with parameter $p = 0.4$. The generated graph is represented by the following adjacency matrix:

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}. \quad (3)$$

From this matrix, we compute the transition matrix given by

$$T = \begin{pmatrix} 0 & .25 & .25 & .25 & .25 \\ .25 & 0 & .25 & .25 & .25 \\ .5 & .5 & 0 & 0 & 0 \\ .5 & .5 & 0 & 0 & 0 \\ .5 & .5 & 0 & 0 & 0 \end{pmatrix}. \quad (4)$$

Once T is obtained, we can get I_t with $t = 1, ..., t^*$, which for this example is presented in Fig. 1. The figure suggests that the function from (2) is adequate for modeling the decay of mutual information with time. Next, we fit the function in Fig.1 with the function from (2), i.e., we estimate the parameters (a, b). The estimated parameters represent the structure of this specific graph. These parameters would be one sample of the distribution of (a, b) for the model ER with $p = 0.4$ and $N = 5$.

IV. Results

A. Comparing models to Erdős-Rényi

In this section we use three models: the ER, the Stochastic Block model (SBM) and the Planted Clique (PC). The definition of these models is as follows:

- ER: $P(A_{ij} = 1) = p$ for all $i \in \{1, ..., N\}$ and $i < j$, where this last inequality comes from the fact that we are considering just undirected graphs.
- SBM: in our case we use the two-community structure. Define $C_1, C_2 \subset \{1, ..., N\}$ as the nodes in the first and second community. Let $P(A_{ij} = 1) = p_{in}$ if $i, j \in C_1$ or $i, j \in C_2$ and $P(A_{ij} = 1) = p_{out}$ if i and j belong to different communities. Again, the probabilities describe the case for $i < j$, so the graph is undirected.

ISBN 978-90-827970-1-5 © EURASIP 2018

- PC: as described in [6], we first generate an ER graph, then randomly select k nodes in the graph and connect all those k nodes together.

In the experiment, for each model we generated $S = 1,000$ graphs of $N = 1,000$ nodes with the following specific parameters:

- ER: $p = 0.05$,
- SBM: we define $\delta = p_{in} - p_{out}$, $p = \frac{p_{in}+p_{out}}{2} = 0.05$, $\delta \in \{0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09\}$, and
- PC: $p = 0.05$ and $k \in \{36, 42, 53, 58, 64\}$.

An important point is that the parameters were chosen so that the densities of the graphs were similar to make the detection of the structure of the graph more challenging.

For each model defined above we sampled $S = 1,000$ graphs, that is, we generated $1,000$ adjacency matrices. We then proceeded with computing the transition matrices based on the adjacency matrices. Next, we computed the mutual information between the initial state and the state at time $t = 1, ..., t^*$ of a random walk in the graph using (1) with $t^* = 30$. With the curve obtained by computing $I(t)$ with $t = 1, .., t^*$ for each sample of each model, we estimated the parameters (a, b) by fitting the function shown in (2). Thus, we had a sample of $14,000$ ($1,000$ for each of the 14 considered models) (a, b) parameters obtained from the simulated graphs that are shown in Fig. 2 for graphs with $N = 1,000$ nodes. In these figures, we observe that as δ decreases, it is harder to distinguish between the SBM and the ER models; also, it is harder to detect differences between the PC and the ER models than between the ER and the SBM models.

Next, we used the sample of parameters (a, b), shown in Fig. 2, obtained from the considered models (14 of them) with $N = 1,000$ nodes.

Now that we mapped each graph to a point in $\mathbb{R}^2$, we divided the sample of $14,000$ parameters (a, b) in two sets, a training set and a test set. We picked 500 samples at random from each kind of graph model, and so we ended up with $7,000$ samples in the training set. The samples that were not selected in the training set were assigned to the test set.

In our experiment, we wanted to test if our graph representation tool wass meaningful to distinguish between the ER model and another model, the SBM or PC, and we do the comparison one at a time. First we compared the ER model with the SBM model with $\delta = 0.02$. In order to do this, we took the 500 samples in the training set that came from the ER set and the 500 samples that came from the SBM set with $\delta = 0.02$ and calibrated the RF model with 100 decision trees. Then we classified the 500 samples from the ER model and the 500 samples from the SBM model with $\delta = 0.02$ coming from the test set and we used as a performance measure the Area Under the Curve (AUC) of the Receiving Operating Characteristics (ROC) curve. We compared the proposed method (PM), against those presented in [6], which we call Benchmark method (BM).

Table I shows the comparison between the ER model and each of the SBM models, using as measure the AUC. Table II shows the results for the comparison between the ER model and each of the PC models. The comparison with the BM are favorable in both cases.

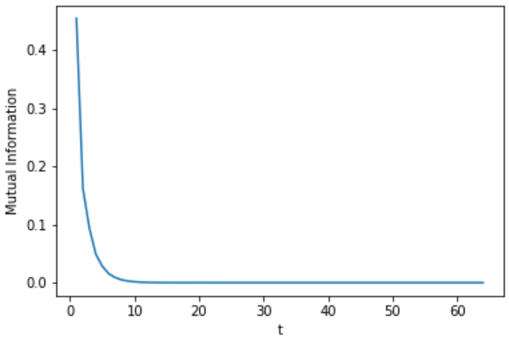

Fig. 1. Mutual information curve I_t.

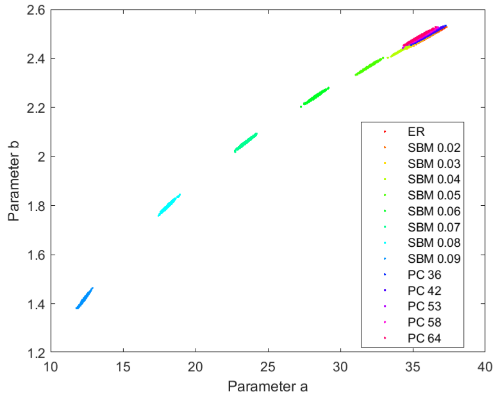

Fig. 2. Fitted parameter to I_t for $N = 1000$.

TABLE I
ER AND SBM MODEL

	Model AUC	
δ	***PM***	***BM***
0.02	72.35%	95%
0.03	99.33%	100%
0.04	100%	100%
0.05	100%	100%
0.06	100%	100%
0.07	100%	100%
0.08	100%	100%
0.09	100%	100%

[a]PM: Proposed method.[b]BM: Benchmark method.

B. Significance testing

In addition to the models used in the last section, we use the Barabási-Albert model (BA) described in [2]. This model is characterized by the number of links m added in each step of the algorithm. Every new link is connected to the existing nodes with probability proportional to the degree of the node. We take $m \in \{2, 4, 20, 50, 200\}$ in the simulations.

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE II
ER AND PC

k	Model AUC	
	PM	***BM***
36	99.99%	84%
42	99.98%	97%
53	100%	100%
58	99.99%	100%
64	100%	100%

[a]PM: Proposed method. [b]BM: Benchmark method.

In this section we assume that we have two inputs. The first input is a graph model that we call the test model and the second is an observed network. The idea now is to test if we should reject the model of the observed network. We simulated $S = 1,000$ samples from the model we wanted to test. From the graph samples, we obtained the corresponding parameters $(a_i^{gm}, b_i^{gm})_{i=1,\ldots,S}$. From the observed network, we computed $\left(a^{obs}, b^{obs}\right)$. The next step was to perform a test that assessed if we should reject the hypothesis that the observed network belongs to the proposed model. To that end, we used the squared Mahalanobis distance (SMD) which has been use as an outlier detection tool as in [4]. This distance measure is defined as:

$$SMD = (x - \mu)' \, \hat{\Sigma}^{-1} (x - \mu) \,. \tag{5}$$

In our case, $\mu \in \mathbb{R}^2$ and $\hat{\Sigma} \in \mathbb{R}^{2\times 2}$ are the mean and sample covariance matrix of the sample $(a_i^{gm}, b_i^{gm})_{i=1,\ldots,S}$ and x corresponds to $\left(a^{obs}, b^{obs}\right)$. We assume that the sample of parameters from the model are distributed according to a normal distribution. Given this assumption, the SMD is distributed according to $\tilde{\chi}_2^2$.

As before we set $t^* = 30$ and we kept the parameters of the models presented in the above section. To the previous samples of graphs, we added the BA model parameters. For each available model, we took the 1000 samples of the parameters (a, b) and computed μ and $\hat{\Sigma}$. After computing these parameters, we selected another model and for each available observation $x = (a, b)$, we evaluated SMD obtaining SMD_x according to (5). Then we checked if

$$P\left(SMD > SMD_x\right)) < \gamma, \tag{6}$$

where we set $\gamma = 0.01$. If the statement was correct, we rejected that the observation belonged to the test model, the one we used to compute μ and $\hat{\Sigma}$. We show the results in Table III, where in the first column we present the model that was used to estimate μ and $\hat{\Sigma}$ in (5), in the second column we show the model for which we take observations to use as x in (5). Since we have $S = 1,000$ observations for the model of the observed network in column 3 of Table III, we present the percentage of observations that were rejected as belonging to the test model. Because of space limitations, we only show the comparison between different models where the $\%$ of correctly rejected is less than 100%. The different SBM models are described as SBM δ, that is, SMB 0.02 is the SBM model with parameter $\delta = 0.02$. In a similar way, PC 36 corresponds to the PC model with $k = 36$. The BA models are not shown in Table III because they can be completely separated from the rest of the models and between them. The only pair of models that it is hard to reject is the ER-SBM 0.02. When the test model was the ER and the observed networks came from the SBM 0.02, we rejected it in only 8.70% of the cases. Likewise, when the test model was the SBM 0.02 and the observed network came from the ER model, we rejected it 5.90% of the time.

TABLE III
SIGNIFICANCE TESTING:$\gamma = 0.01$

Test model	Observed network	% correctly rejected
ER	SBM 0.02	8.70%
ER	SBM 0.03	99.99%
SBM 0.02	ER	5.90%
SBM 0.02	SBM 0.03	99.10%
SBM 0.03	SBM 0.02	99.10%
SBM 0.03	PC 36	99.50%
SBM 0.04	PC 42	99.00%
PC 36	SBM 0.03	99.90%
PC 36	PC 42	99.60%
PC 42	SBM 0.04	98.90%
PC 42	PC 36	84.20%
PC 53	PC 58	65.70%
PC 53	PC 64	99.80%
PC 58	PC 53	55.00%
PC 58	PC 64	81.00%
PC 64	PC 58	69.20%

In Table IV, in the second column we show the percentage of observations incorrectly rejected from the indicated model. For example, 1.50% of the observations were rejected as coming from ER while in fact that was the generating model. Since we were using $\gamma = 0.01$, the values presented in the table should be close to 1%.

TABLE IV
SIGNIFICANCE TESTING:$\gamma = 0.01$

Model	% rejected
ER	1.50%
SBM 0.02	1.00%
SBM 0.03	0.90%
SBM 0.04	0.90%
SBM 0.05	1.00%
SBM 0.06	0.80%
SBM 0.07	0.60%
SBM 0.08	1.20%
SBM 0.09	0.80%
PC 36	1.10%
PC 42	1.10%
PC 53	1.30%
PC 58	1.30%
PC 64	1.10%
BA 2	1.00%
BA 4	1.10%
BA 20	1.10%
BA 50	1.50%
BA 200	1.00%

V. Conclusions

The Markovian relaxation concept used in [9], to analyze graph structure in a clustering task, motivated us to try the graph representation shown in this paper. In summary, our graph representation corresponds to "mapping" the adjacency matrix of a graph to a vector $(a, b) \in \mathbb{R}^2$. We conjectured that the samples of (a, b) can be used for model selection and rejection because there is a distribution of values in $\mathbb{R}^2$ that is useful in identifying graph models. Extensive experiments have demonstrated that, indeed, the parameters (a, b) can be used for model selection and significance testing. At this point, even though the empirical evidence is compelling, we lack a formal theoretical proof to justify our results.

Acknowledgment

Francisco Hawas thanks Arturo Cifuentes for helpful comments regarding this manuscript and Javier Zapata who provided valuable insights in the initial steps of this research.

References

[1] E. M. Airoldi, X. Bai and K. M. Carley, "Network sampling and classification: an investigation of network model representations," Decis. Support Syst., vol. 51(3), pp. 506–518, June 2011.

[2] A. L. Barabási and R. Albert,"Emergence of scaling in random networks," Science, vol. 286, pp. 509–512, October 1999.

[3] R. S. Caceres, L. Weiner, M. C. Schmidt, B. A. Miller and W. M. Campbell, "A model selection framework for graph-based data," arXiv:1609.04859v1 [cs.SI], Sep 2016.

[4] C. Leys, O. Klein, Y. Dominicy and C. Ley,"Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance," Journal of Experimental Social Psychology, vol. 74, pp. 150–156, January 2018.

[5] J. Janssen, M. Hurshman and N. Kalyaniwalla,"Model selection for social networks using graphlets," Internet Math., vol. 8, pp. 338–363, Dec 2012.

[6] L. Li, W. M. Campbell and R. S. Caceres, "Graph model selection via random walk," arXiv:1704.05516v2 [cs.SI], May 2018.

[7] M. Middendorf, E. Ziv, C. Adams, J. Hom, R. Koytcheff, C. Levovitz, G. Woods, L. Chen and C. Wiggins, "Discriminative topological features reveal biological network mechanisms," BMC Bioinformatics, vol. 5, pp. 1471–2105, November 2004.

[8] D. Y. Takahashi, J. R. Sato, C. E. Ferreira and A. Fujita, "Discriminating different classes of biological networks by analyzing he graphs spectra distribution," PLoS ONE, vol. 7(12), e49949, December 2012.

[9] N. Tishby and N. Slonim, "Data clustering by Markovian relaxation and the information bottleneck method," in Proc. Int. Conf. Adv. Neural Inform. Process. Syst., 2000, pp. 640–646.

[10] R. C. Wilson and P. Zhu, "A study of graph spectra for comparing graphs and trees," Pattern Recognition, vol. 41(9), pp. 2833–2841, September 2008.