A Random Dot Product Graph Model for Weighted and Directed Networks

Bernardo Marenco*‡, Paola Bermolen*‡, Marcelo Fiori*‡, Federico Larroca* and Gonzalo Mateos†
*Facultad de Ingeniería, Universidad de la República, Uruguay
‡ Centro Interdisciplinario en Ciencia de Datos y Aprendizaje Automático (CICADA), Universidad de la República, Uruguay
†Dept. of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA.
Emails: {bmarenco,paola,mfiori,flarroca}@fing.edu.uy and gmateosb@ur.rochester.edu

***Abstract*—In its most basic form, the Random Dot Product Graph (RDPG) model assigns a low-dimensional vector to each vertex, and postulates that an edge between any two nodes exists with probability given by the inner product of said vectors. Recently, this latent position model has been extended to account for weighted graphs (the so-called Weighted (W-)RDPG), now embedding each node with a sequence of vectors. For a given node pair, the inner product between the k-th elements of the respective sequences specifies the k-th moment of the edge weight's distribution. However, graphs adhering to this nonparametric W-RDPG model are constrained to be undirected and homophilic (i.e., the adjacency matrix must be positive semi-definite in expectation). In this work, we extend the model's expressivity by proposing a variant for directed graphs, which may also include heterophilic nodes. To this end, we endow each vertex with two sequences, respectively modeling the node's incoming and outgoing connectivity behavior. We propose an embedding algorithm to estimate the latent nodal sequences from an observed adjacency matrix, and also discuss graph generation when the latent positions are given. The effectiveness of the novel weighted and directed (WD-)RDPG model is illustrated via several test cases, including both synthetic and real-life networks.**

***Index Terms*—graph representation learning, node embeddings, directed graphs, weigthed graphs, graph generation.**

I. Introduction

The Random Dot Product Graph (RDPG) model has emerged as a popular latent position model for the analysis and generation of undirected and unweighted network graphs. In this model each node $i \in \{1, 2, \ldots, N\}$ is associated to a latent vector $\mathbf{x}_i \in \mathbb{R}^d$ (which is unobserved and may be interpreted as an embedding, tipically $d \ll N$), and the probability of existence of an edge (i, j) between nodes i and j is the inner product between $\mathbf{x}_i$ and $\mathbf{x}_j$. In other words, the entry A_{ij} of the symmetrix adjacency matrix $\mathbf{A} \in \{0,1\}^{N\times N}$ has a Bernoulli($\mathbf{x}_i^\top \mathbf{x}_j$) distribution. For $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N\times d}$, the outer product $\mathbf{X}\mathbf{X}^\top$ is an $N \times N$ matrix collecting all edge formation probabilities.

The widespread adoption of this model resides in its simplicity and interpretability, without compromising expressive power. For instance, Erdös-Rényi (ER) or Stochastic Block Models (SBMs) graphs, as well as other more sophisticated models can be seen as particular cases of RDPG; see e.g., [1]. With regards to interpretability, since the connection probability is given by the inner product of the embeddings, the affinity between the corresponding nodes is directly captured by their alignment. For $d \leq 3$, we may rely on visual inspection of the vector representations to identify similar nodes. In general, we may screen for community structure, or, carry out angle-based clustering of nodes in latent space [2], [3].

This "vanilla" RDPG model, however, presents certain shortcomings. Most germane to the contributions in this paper, it can only describe graphs which are undirected, unweighted, and homophilic (meaning the adjacency matrix must be positive semi-definite (PSD) in expectation, cf. $\mathbb{E}[\mathbf{A}] = \mathbf{X}\mathbf{X}^\top$). Generalizations have been proposed to tackle *some* of these limitations *individually*, for instance the Generalized (G)RDPG [4], or the Weighted (W-)RPDG [5].

Here we propose a more general model to account for weighted and directed (di)graphs, allowing also for heterophilic relations [6]. We model the first k moments of the edge weight distributions via a sequence of k latent position matrices $\mathbf{X}[k] \in \mathbb{R}^{N\times d_k}$. For a given node pair, the inner product of the respective embeddings specifies the k-th moment of the incident edge weight. Directionality and heterophily of links is achieved by encoding the mean ($k = 1$) and higher-order moments using two matrix sequences, $\mathbf{X}^l[k]$ and $\mathbf{X}^r[k]$. This way, the outer products $\mathbf{X}^l[k](\mathbf{X}^r[k])^\top$ need not be symmetric, and are devoid of spectral (sign) restrictions.

We develop a spectral embedding algorithm to estimate the latent nodal sequences, from an observed adjacency matrix adhering to the novel weighted and directed (WD-)RDPG model. Unlike [7] that only models $\mathbb{E}[\mathbf{A}]$, we show that embedding the first few ($k \geq 1$) moments of $\mathbf{A}$ offers a richer representation of weighted network structure. Conversely, if the sequences of latent positions are given (perhaps as outputs of the aforementioned embedding step), we discuss how to draw graph samples from the corresponding WD-RDPG.

The rest of the paper is organized as follows. In Section II we formally define the WD-RDPG model. We then describe how to infer the latent positions from an observed graph, and analyze embeddings of a United Nations (UN) voting dataset (Section III). In Section IV we discuss graph generation and use global migration data to assess model goodness-of-fit. Conclusions and limitations are presented in Section V.

This work was partially funded by CSIC (I+D project 22520220100076UD) and ANII (FCE-1-2023-1-176172).

II. Weighted and Directed RDPG Model

A. Weighted RDPG

The W-RDPG model was introduced for application purposes in [5], and formally defined as a generative model in [8], where we also establish the consistency and asymptotic normality of the embeddings. Specifically, we propose assigning a sequence of vectors to each node, which are related to the moment-generating function (MGF) of the weight distribution.

Specifically, each node is endowed with a sequence of latent positions $\mathbf{x}_i[k] \in \mathbb{R}^{d_k}$ that determine the k-th moment of the weighted adjacency matrix as $\mathbb{E}\left[A_{ij}^k\right] = \mathbf{x}_i^\top[k]\mathbf{x}_j[k]$, for $k \in \mathbb{N}_+$. Given the sequence $\mathbf{X} := \{\mathbf{X}[k]\}_k$, with $\mathbf{X}[k] = [\mathbf{x}_1[k], \dots, \mathbf{x}_N[k]]^\top \in \mathbb{R}^{N \times d_k}$, the W-RDPG model specifies the MGF of the symmetric adjacency matrix $\mathbf{A}$ as

$$\mathbb{E}\left[e^{tA_{ij}}|\mathbf{X}\right] = \sum_{k=0}^{\infty} \frac{t^k \mathbb{E}\left[A_{ij}^k\right]}{k!} = 1 + \sum_{k=1}^{\infty} \frac{t^k \mathbf{x}_i^\top[k]\mathbf{x}_j[k]}{k!}, \quad (1)$$

with independent edge weights A_{ij}. For $\mathbf{A}^{(k)} := \mathbf{A} \circ \mathbf{A} \circ \cdots \circ \mathbf{A}$ (k times), where $\circ$ is the entry-wise (or Hadamard) product, (1) implies that $\mathbf{M}[k] := \mathbb{E}\left[\mathbf{A}^{(k)}\right] = \mathbf{X}[k]\mathbf{X}^\top[k]$. The "vanilla" RDPG is recovered when $\mathbf{x}_i[k] = \mathbf{x}_i$, $\forall k > 0$.

There have been other RDPG generalizations to the weighted case. In [9] and [10], adjacency matrix entries are generated from a given parametric distribution $F_{\boldsymbol{\theta}}(A_{ij})$, with $\boldsymbol{\theta} \in \mathbb{R}^L$. In this framework each node now has L latent vectors $\mathbf{x}_i[l] \in \mathbb{R}^{d_l}$ ($l = 1, \dots, L$). For distribution $F_{\boldsymbol{\theta}}(A_{ij})$, where the l-th entry of $\boldsymbol{\theta}$ is $\theta_l = \mathbf{x}_i^\top[l]\mathbf{x}_j[l]$ and the A_{ij} are independent. Once more, the RDPG is recovered by letting $F_{\boldsymbol{\theta}}(A_{ij})$ be a Bernoulli(θ) distribution. However, this implies that all edges must share the same weight distribution, albeit with potentially different parameters. This limitation may be partially overcome by considering a mixture distribution. Nonetheless, a key restriction remains: $F_{\boldsymbol{\theta}}(A_{ij})$ must be chosen *a priori*, further limiting the model's flexibility and applicability.

Remark 1 (Comparison with [7])**.** A different RDPG model for weighted graphs was put forth in [7], where each node has a single associated latent position $\mathbf{z}_i \in \mathcal{Z}$, which is endowed with a probability distribution F. It is postulated that, given a family $\{H(\mathbf{z}_1, \mathbf{z}_2) : \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}\}$ of symmetric real-valued distributions, there exists a map $\phi : \mathcal{Z} \mapsto \mathbb{R}^d$ such that if $A_{ij} \sim H(\mathbf{z}_i, \mathbf{z}_j)$, then $\mathbb{E}\left[A_{ij}\right] = \phi^\top(\mathbf{z}_i)\mathbf{I}_{p,q}\phi(\mathbf{z}_j)$, where $\mathbf{I}_{p,q}$ is a diagonal matrix of p ones and q minus ones, such that $p + q = d$. This diagonal matrix facilitates modeling heterophilic (or disassortative) behavior, as in [4]. Interestingly, one can consistently recover $\mathbf{x}_i = \phi(\mathbf{z}_i)$ via the adjacency spectral embedding (ASE) of $\mathbf{A}$, and the estimated $\{\mathbf{x}_i\}$ asymptotically follow a multivariate Normal distribution. However, one can only recover $\phi(\mathbf{z}_i)$, i.e., the latent positions for the mean matrix $\mathbb{E}\left[\mathbf{A}\right]$. Accordingly, this model cannot discriminate between pairs of edges that are associated with different distributions sharing a common mean.

Unlike existing models, W-RDPG does not require a priori specification of the weight distribution, which can be either discrete or continuous. We model the mean ($k = 1$) of the weight distribution as well as its higher-order moments ($k > 1$), thus enhancing the model's discriminative power; see [5]. Up to this point, the moment matrix $\mathbf{M}[k] = \mathbf{X}[k]\mathbf{X}^\top[k]$ is by definition PSD, limiting the analysis to undirected and homophilic (or assortative) networks. Next, we lift these restrictions and extend our model to weighted digraphs. In Section III, we discuss how to estimate the latent vectors from an observed graph, with statistical guarantees as $N \to \infty$.

B. A new model for weighted digraphs

For digraphs, the idea of using a pair of nodal latent positions to model the in and out connectivity behavior of each vertex was first introduced in [11]. Combining this idea with our W-RDPG model, consider that each node has two sequences of latent positions $\mathbf{x}_i^l[k] \in \mathbb{R}^{d_k}$ and $\mathbf{x}_i^r[k] \in \mathbb{R}^{d_k}$. Together, they determine the k-th moments of the weighted adjacency matrix as $\mathbb{E}\left[A_{ij}^k\right] = (\mathbf{x}_i^l[k])^\top \mathbf{x}_j^r[k]$, for $k \in \mathbb{N}_+$.

Given the sequences $\mathbf{X}^l := \{\mathbf{X}^l[k]\}_k$, with $\mathbf{X}^l[k] = [\mathbf{x}_1^l[k], \dots, \mathbf{x}_N^l[k]]^\top \in \mathbb{R}^{N \times d_k}$ and $\mathbf{X}^r := \{\mathbf{X}[k]\}_k$, with $\mathbf{X}^r[k] = [\mathbf{x}_1^r[k], \dots, \mathbf{x}_N^r[k]]^\top \in \mathbb{R}^{N \times d_k}$ the WD-RDPG model specifies the MGF of the adjacency matrix as

$$\mathbb{E}\left[e^{tA_{ij}}|\mathbf{X}^l, \mathbf{X}^r\right] = \sum_{k=0}^{\infty} \frac{t^k \mathbb{E}\left[A_{ij}^k\right]}{k!} = 1 + \sum_{k=1}^{\infty} \frac{t^k (\mathbf{x}_i^l[k])^\top \mathbf{x}_j^r[k]}{k!},$$

where again the entries A_{ij} are independent. The "vanilla" RDPG is recovered when $\mathbf{x}_i^l[k] = \mathbf{x}_i^r[k] = \mathbf{x}_i$, $\forall\, k > 0$, and the W-RDPG is obtained by setting $\mathbf{x}_i^l[k] = \mathbf{x}_i^r[k], \forall\, k > 0$.

III. Inference Under the WD-RDPG Model

Let us now discuss the inference problem. That is, how to estimate the sequence of embeddings $\{\hat{\mathbf{X}}^l[k], \hat{\mathbf{X}}^r[k]\}_k$ given an observed adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, which may be asymmetric. Recall that in this case we have that the expected value of the entry-wise k-th power of the adjacency matrix is precisely $\mathbf{M}[k] := \mathbb{E}\left[\mathbf{A}^{(k)}\right] = \mathbf{X}^l[k](\mathbf{X}^r[k])^\top$. Consequently, we can view $\mathbf{A}^{(k)}$ as a noisy observation of $\mathbf{M}[k]$. Following the ASE approach for RDPG inference [2], for target k and prescribed d_k we form the least-squares (LS) estimator:

$$\{\hat{\mathbf{X}}^l[k], \hat{\mathbf{X}}^r[k]\} \in \underset{\mathbf{X}^l, \mathbf{X}^r \in \mathbb{R}^{N \times d_k}}{\operatorname{argmin}} \left\| \mathbf{A}^{(k)} - \mathbf{X}^l (\mathbf{X}^r)^\top \right\|_F^2. \quad (2)$$

In words, $\hat{\mathbf{M}}[k] = \hat{\mathbf{X}}^l[k](\hat{\mathbf{X}}^r[k])^\top$ is the best rank-d_k approximation to the entry-wise k-th power adjacency matrix $\mathbf{A}^{(k)}$, in the Frobenius-norm sense.

One possible solution to (2) is provided by the Singular Value Decomposition (SVD) of $\mathbf{A}^{(k)}$, which we write as $\mathbf{A}^{(k)} = \mathbf{U}[k]\mathbf{D}[k]\mathbf{V}^\top[k]$. Given a prescribed embedding dimension d_k, we define $\hat{\mathbf{D}}[k]$, $\hat{\mathbf{U}}[k]$ and $\hat{\mathbf{V}}[k]$ by keeping only the d_k most significant singular values and the corresponding columns of $\mathbf{U}[k]$ and $\mathbf{V}[k]$. We thus arrive to the following ASE solution:

$$\hat{\mathbf{X}}^l[k] = \hat{\mathbf{U}}[k]\hat{\mathbf{D}}[k]^{1/2} \quad \text{and} \quad \hat{\mathbf{X}}^r[k] = \hat{\mathbf{V}}[k]\hat{\mathbf{D}}[k]^{1/2}.$$

Note that "backwards-compatibility" is enforced by the choice of the factor $\hat{\mathbf{D}}[k]^{1/2}$. This way, if $\mathbf{A}^{(k)}$ is symmetric, we have

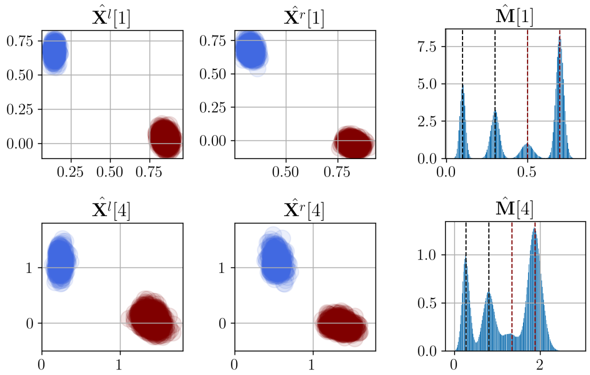

Fig. 1. Inference results for a two-class SBM with Gaussian weights and $N = 2000$ nodes. The two rightmost plots show histograms of the estimated $\hat{\mathbf{M}}[k]$ and the vertical lines indicate the true moments. Embeddings and moments are accurately estimated up to $k = 4$.

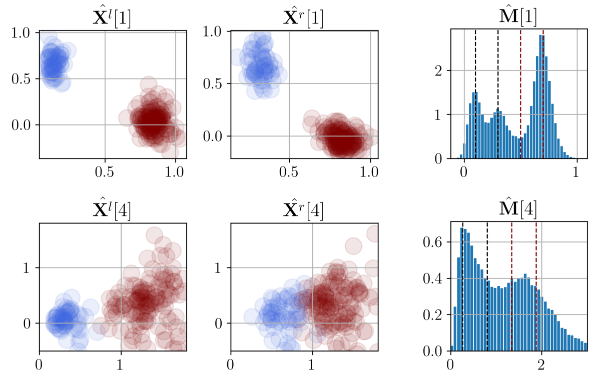

Fig. 2. Inference results for a two-class SBM with Gaussian weights and $N = 200$ nodes. The accuracy of the estimated embeddings degrades when compared to the $N = 2000$ setting. In particular, the embeddings corresponding to $k = 4$ are quite noisy.

that $\hat{\mathbf{X}}^l[k] = \hat{\mathbf{X}}^r[k]$ as in W-RDPG embeddings. Regarding the embedding dimension d_k, it can be chosen through the elbow rule on the scree plot of $\mathbf{A}^{(k)}$'s singular values, as in ASE numerical packages [12]. This will be the method of choice for the ensuing numerical experiments.

Synthetic test case. Consider a relatively simple but illustrative example. We have a large SBM graph consisting of $N = 2000$ nodes, where community c_1 has 70% of the nodes. The interconnection probability is given by the matrix $\mathbf{\Pi} = \left(\begin{smallmatrix} 0.7 & 0.3 \\ 0.1 & 0.5 \end{smallmatrix}\right)$ and all the weights are normally distributed with a mean 1 and standard deviation 0.5.

The estimated embeddings $\hat{\mathbf{X}}^l[k]$ and $\hat{\mathbf{X}}^r[k]$ corresponding to moments $k = 1, 4$ are shown in Fig. 1. The correctness of these embeddings is verified in the rightmost panels of Fig. 1, which depict histograms of $\hat{\mathbf{M}}[k]$ superimposed to the true moments. Note how the embeddings seem to follow a multivariate Normal distribution, and (as expected) the accuracy of the estimate $\hat{\mathbf{M}}[k]$ tends to decrease with k.

This last observation is further corroborated in Fig. 2, where we have repeated the previous experiment, but using only $N = 200$ nodes. We find the embeddings for $k = 1$ are still relatively well estimated. However, the limited amount of data hinders estimation of the embeddings corresponding to the fourth moment. A more thorough analysis of the relation between the estimation's accuracy and the number of nodes (i.e., the sample size) is subject of ongoing investigation. In any case, these finite sample effects will be revisited in the following numerical example.

UN voting data. Let us now discuss a real-life example which will also help us illustrate two advantages of the proposed WD-RDPG model: its interpretability and its ability to cluster nodes beyond their mean behavior; see Remark 1. In particular, we will consider UN General Assembly voting data [13]. For each roll call and member country, the dataset indicates if the country was present and if so the corresponding vote (either 'Yes', 'No', or 'Abstain') for each proposal. We represent an year's worth of votes as a bipartite digraph, where nodes are both member countries and roll calls, and an edge from a country to a roll call may have a weight of either 1, -1 or 0 (corresponding to a 'Yes', 'No', or, either 'Abstain' or absent vote, respectively).

The estimated embeddings $\hat{\mathbf{X}}^l[1]$ and $\hat{\mathbf{X}}^r[1]$ (i.e., the first moment) for $d = 2$ are shown in the top of Fig. 3. Note that $\mathbf{x}_i^l[k] = \mathbf{0}$ for nodes i corresponding to roll calls (roll calls do not vote) and likewise $\mathbf{x}_j^r[k] = \mathbf{0}$ for countries j (countries are not voted), so we are omitting those components in Fig. 3. Furthermore, countries are depicted with circles and roll calls with diamonds. The colors are chosen according to a Gaussian Mixture Model clustering, to ease visualization.

Apparently, countries can be broadly categorized into three groups. The first group, marked by a red ellipse in Fig. 3 (top) and exemplified by the USA, stands in stark opposition to the majority of roll calls. This is evidenced by the predominantly negative inner product between the nodal embeddings of these countries and most roll calls, suggesting they have opposed the majority of proposals. The second group, highlighted by a green ellipse and including Uruguay, aligns with most roll calls, indicating consistent affirmative voting patterns.The third group, represented by a yellow ellipse and typified by France, occupies an intermediate position between the other two.

Regarding roll calls, they can be arguably clustered into three groups. We will focus on the rightmost group in Fig. 3 (top), enclosed by a dashed ellipse. Note that this group of roll calls is orthogonal to the green group of countries, meaning that their *expected* weight is approximately zero. This could mean that these countries mostly abstained to these roll calls (i.e., a large probability for a weight equal to zero), or that they actually had a more balanced distribution among the three possibilities (even including not abstaining at all, but being equally likely to oppose or support the roll calls).

In order to elucidate between these two possibilities, we have to resort to the embeddings corresponding to $k = 2$; see Fig. 3 (bottom). This second moment evidences that this group is actually conformed by two subgroups of roll calls. A first

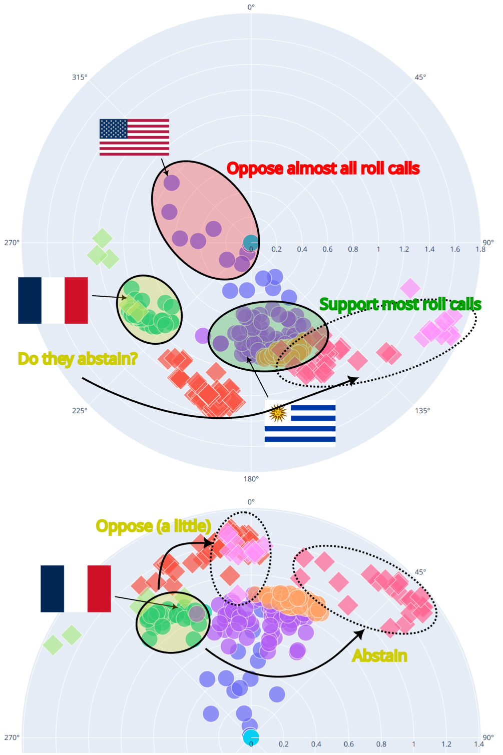

Fig. 3. $\hat{\mathbf{X}}^l[k]$ and $\hat{\mathbf{X}}^r[k]$ for $k = 1$ (top) and $k = 2$ (bottom) corresponding to the UN dataset for the year 2003. Countries (●) and roll calls (◆) are colored using a Gaussian Mixture Model clustering (for $k = 1$). Selected countries and the inferred behavior of the corresponding clusters are indicated.

one which is still orthogonal to the vector representation of the green group of countries, indicating that these countries did abstain in these roll calls. The relative alignment to the rest of the roll calls, together with our previous observations for $k = 1$, suggest that this group of countries actually tend to oppose this second subgroup of roll calls.

IV. Graph Generation Under the WD-RDPG Model

In the "vanilla" RDPG model, if we have access to the latent positions for every node, then it is straightforward to sample graphs adhering to the model: we draw $\frac{N(N-1)}{2}$ independent Bernoulli random variables, each with parameter given by the inner product between the latent positions of nodes incident to each edge. Here we outline the necessary steps to generate graph samples from an WD-RDPG model.

We will assume that we have access to a finite subset of the latent positions sequences (or estimates for that matter), i.e., we know $\mathbf{X}^l[k]$ and $\mathbf{X}^r[k]$, $k = 0, 1, \dots, K$ for some finite integer K. The problem is then to generate an adjacency matrix $\mathbf{A}$ such that for all $1 \leq i, j \leq N$, it holds that $\mathbb{E}\left[A_{ij}^k\right] = (\mathbf{x}_i^l[k])^\top \mathbf{x}_j^r[k]$ for all $k = 0, 1, \dots, K$.

If we make the extra assumption that A_{ij} (the weight of edge (i, j)) follows a discrete distribution that takes distinct values $v_0, v_1, \dots, v_Q$ with probabilities $p_0, p_1, \dots p_Q$[1], then we are after a solution to the following system of equations:

$$\begin{cases} \quad p_0 + \quad p_1 + \cdots + \quad p_Q &= M_{ij}[0] \\ v_0 p_0 + v_1 p_1 + \cdots + v_Q p_Q &= M_{ij}[1] \\ \vdots & \vdots \\ v_0^K p_0 + v_1^K p_1 + \cdots + v_Q^K p_Q &= M_{ij}[K] \end{cases} \tag{3}$$

where $M_{ij}[k] = (\mathbf{x}_i^l[k])^\top \mathbf{x}_j^r[k]$ as per the WD-RDPG model. Furthermore, if the values $v_0, v_1, \dots, v_R$ are known, then (3) is a linear system of equations $\mathbf{V}\mathbf{p} = \mathbf{m}$, where $\mathbf{p} = [p_0, p_1, \dots, p_Q]^\top$ is the probability mass function, $\mathbf{m} = [M_{ij}[0], M_{ij}[1], \dots, M_{ij}[K]]^\top$ collects the given (or estimated) moments, and $\mathbf{V}$ is the Vandermonde matrix

$$\mathbf{V} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ v_0 & v_1 & \dots & v_Q \\ \vdots & \vdots & \ddots & \vdots \\ v_0^K & v_1^K & \dots & v_Q^K \end{pmatrix}.$$

Thus, if $K = Q$ (i.e., we are given the same amount of moments than symbols), the system has a unique solution $\mathbf{p} = \mathbf{V}^{-1}\mathbf{m}$. Note that every Vandermonde matrix is invertible if the v_i's are distinct, which ensures us existence of a unique solution. Also, the inverse of such a Vandermonde matrix has a well known closed form expression, which allows us to avoid the cost of naively inverting $\mathbf{V}$. Given the estimated discrete edge-weight distribution $\mathbf{p}$, one can sample edge weights A_{ij} to generate graphs.

Global migration data. We now illustrate this method using a global migration dataset from 1990 [14]. This dataset provides estimates of the number of people who migrated between each of the 232 countries and regions, capturing flows in both directions.

Our digraph will thus have the countries (and regions) as nodes, but instead of using the raw numbers for edge weights, we will consider $(Q{+}1)$-quantiles. That is, $A_{ij} = 0$ will indicate relatively low number of migrants from country i to country j, whereas the opposite is true if $A_{ij} = Q$. This way we have discretized the weight's distribution, and by changing Q we can test the effect of having different number of symbols for a given graph size. In particular, we tested $Q = 2, 3, 5$.

For each value of Q we estimate the sequence $\{\hat{\mathbf{X}}^l[k], \hat{\mathbf{X}}^r[k]\}$ for $k = 1, \dots, Q$ so that we have as many moments as symbols. We then estimate the edge weights' distribution following the aforementioned method (separately for each edge) and generated 100 graphs by sampling $N \times (N-1)$ random variable per graph. The results are compared to the original migration graph in terms of the nodal in-degree and out-degree sequences, as well as the betweeness centrality of the vertices, in what could be interpreted as a goodness-of-fit test of the generated graphs. The resulting histograms for each value of Q and each network statistic are depicted in Fig. 4.

Note how for both $Q = 2, 3$ (top and middle panels), the generated digraphs closely follow the original network's structure (at least in terms of the three network statistics considered). However, as we move to $Q = 5$, which requires

[1] Both the v_i's and the p_i's might depend on i and j. We are not making this explicit in our notation for ease of exposition.

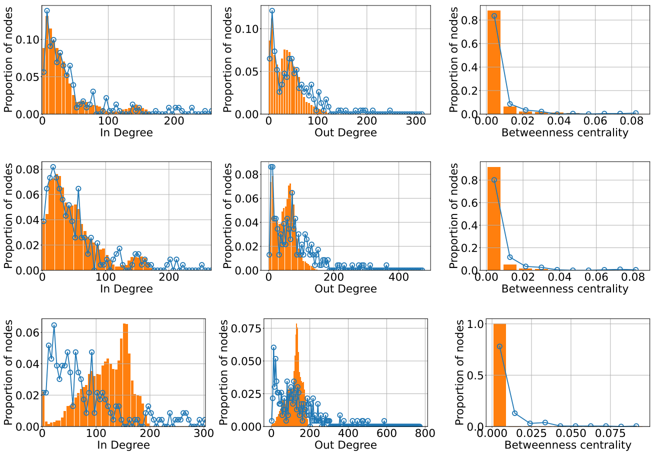

Fig. 4. Goodness-of-fit test for the generative method on the UN migration data example using $Q + 1$ symbols, with Q equal to 2 (top), 3 (middle) and 5 (bottom). As the number of symbols increases, and so does the required number of moments to be estimated, the generated digraphs deviate from the original network's structure.

good estimates of up to the fifth moment of the weights' distribution, the generated graphs fail to reproduce the structural properties in the UN migration graph. The reason behind this poor performance was discussed in the synthetic example of Section III. Given a graph of certain size N, the embedding method can produce accurate estimates of $\hat{\mathbf{M}}[k]$ up to some order k. Here, we find this value is around $k = 4$ and the moments are overly noisy for the system (3) with $Q = 5$.

V. Conclusions, Limitations, and Future Work

In this paper we have presented a variant of the popular RDPG model to represent weighted digraphs. Edge weights can have any distribution (as long as all their moments are well-defined) and we do not require any *a priori* specification of this probabilistic law. Furthermore, the directed model can seamlessly accommodate heterophilic behaviour in the nodes.

We also introduced an inference method based on the SVD decomposition of the entry-wise powers of the adjacency matrix. Additionally, we highlighted how our moment-based approach preserves the interpretability of the "vanilla" RDPG, while enabling the clustering of nodes beyond their mean behavior. Studying the theoretical properties of our estimator, such as characterizing its consistency and distribution, represent a natural next step of our research. Further extensions could explore how to incorporate missing (i.e., unobserved) edges in the adjacency matrix, building on existing work in the context of the (directed) RDPG [15], [16].

We have also discussed an associated graph generation procedure, to sample WD-RDPG graphs when edge weights are discrete random variables. The proposed method assumes that we have estimated as many moments as possible symbols, and when feasible, the generated graphs faithfully replicate structural properties of the training graph. However, we have empirically shown that, as expected, given a certain graph size we can only accurately estimate moments up to a certain order. It would thus be interesting and useful to design an algorithm that can generate samples with only a few moments of the underlying edge weight distribution.

The other aspect that deserves attention in our generative model is its scalability. The proposed approach estimates the distribution for each edge and subsequently samples the corresponding weight. This means we need to estimate $N \times (N - 1)$ distributions, and for each new graph, we then generate $N \times (N - 1)$ samples. However, if the graph follows a structure such as an SBM, where groups of nodes exhibit similar behavior, we could cluster these nodes and estimate the weight distributions collectively for each group. This approach has the potential to both enhance estimation accuracy and significantly reduce computational costs. Exploring this idea through empirical and theoretical studies represents an intriguing direction for future research.

References

[1] A. Athreya, D. E. Fishkind, M. Tang, C. E. Priebe, Y. Park, J. T. Vogelstein, K. Levin, V. Lyzinski, and Y. Qin, "Statistical inference on random dot product graphs: A survey," *J. Mach. Learn. Res.*, vol. 18, no. 1, p. 8393–8484, Jan. 2017.

[2] E. Scheinerman and K. Tucker, "Modeling graphs using dot product representations," *Comput. Stat*, vol. 25, pp. 1–16, 2010.

[3] V. Lyzinski, M. Tang, A. Athreya, Y. Park, and C. E. Priebe, "Community detection and classification in hierarchical stochastic blockmodels," *IEEE Trans. Netw. Sci. Eng.*, vol. 4, no. 1, pp. 13–26, 2017.

[4] P. Rubin-Delanchy, J. Cape, M. Tang, and C. E. Priebe, "A statistical interpretation of spectral embedding: The generalised random dot product graph," *J. R. Stat. Soc., B: Stat. Methodol.*, vol. 84, no. 4, pp. 1446–1473, 2022.

[5] B. Marenco, P. Bermolen, M. Fiori, F. Larroca, and G. Mateos, "Online change point detection for weighted and directed random dot product graphs," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 8, pp. 144–159, 2022.

[6] S. Luan *et al.*, "The heterophilic graph learning handbook: Benchmarks, models, theoretical analysis, applications and challenges," *arXiv:2407.09618 [cs.LG*, 2024.

[7] I. Gallagher, A. Jones, A. Bertiger, C. E. Priebe, and P. Rubin-Delanchy, "Spectral embedding of weighted graphs," *J. Am. Stat. Assoc.*, vol. 119, no. 547, p. 1923–1932, 2023.

[8] B. Marenco, P. Bermolen, M. Fiori, F. Larroca, and G. Mateos, "Weighted random dot product graphs," *Electron. J. Stat.*, 2024 (journal paper in preparation).

[9] D. R. DeFord and D. N. Rockmore, "A random dot product model for weighted networks," *arXiv:1611.02530 [stat.AP]*, 2016.

[10] R. Tang, M. Tang, J. T. Vogelstein, and C. E. Priebe, "Robust estimation from multiple graphs under gross error contamination," *arXiv:1707.03487 [stat.ME]*, 2017.

[11] C. E. Priebe, Y. Park, M. Tang, A. Athreya, V. Lyzinski, J. T. Vogelstein, Y. Qin, B. Cocanougher, K. Eichler, M. Zlatic, and A. Cardona, "Semiparametric spectral modeling of the drosophila connectome," *arXiv:1705.03297 [stat.ML]*, 2017.

[12] J. Chung, B. D. Pedigo, E. W. Bridgeford, B. K. Varjavand, H. S. Helm, and J. T. Vogelstein, "Graspy: Graph statistics in Python." *J. Mach. Learn. Res.*, vol. 20, no. 158, pp. 1–7, 2019.

[13] E. Voeten, A. Strezhnev, and M. Bailey, "United Nations General Assembly Voting Data," 2009. [Online]. Available: https://doi.org/10.7910/DVN/LEJUQZ

[14] United Nations, Department of Economic and Social Affairs, Population Division, "Trends in International Migrant Stock: The 2015 Revision," 2015, pOP/DB/MIG/Stock/Rev.2015. [Online]. Available: https://networks.skewed.de/net/un_migrations

[15] M. Fiori, B. Marenco, F. Larroca, P. Bermolen, and G. Mateos, "Algorithmic advances for the adjacency spectral embedding," in *Proc. of European Signal Process. Conf.*, 2022, pp. 672–676.

[16] ——, "Gradient-based spectral embeddings of random dot product graphs," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 10, pp. 1–16, 2024.