Generalized Random Dot Product Models for Multigraphs

Daryl DeFord
Advisor: Dan Rockmore
Dartmouth College Department of Mathematics

Abstract

This work presents a generalization of the random dot product model for networks whose edge weights are drawn from a parametrized probability distribution. We focus on the case of integer weight edges and show that many of the results for traditional dot product networks can be extended to this setting, particularly with respect to small world metrics. We show that our model outperforms the binary version of the dot product model for community detection problems on weighted networks and exhibit a stress function for dimension selection.

Motivation

Complex networks are used to model many types of physical and social systems. However, the process of extracting a useful network model from a noisy data set requires making a series of decisions that determine the properties of the eventual network. Some of these choices are suggested by the data, such as the categorization of nodes and edges, while others may be determined by the mathematical tools available, such as the choice between digraphs and undirected models or between simple networks and weighted networks. Finally, some choices, such as the selection of thresholding parameters, are influenced by many factors and can change the resulting network in subtle and complicated ways [9].

Our contribution is a generative model for weighted networks that incorporates the weights directly in order to both construct more accurate null models and provide a geometric framework for studying properties of individual networks. The dot product formulation gives natural interpretations of the angle and magnitude of the vectors in the latent embedding in terms of similarity and centrality respectively. This method is also valuable for modeling networks derived from time series data, as well as collections of networks defined on the same node set.

Related Generative Models

Our model is an extension of the Random Dot Product Model (RDPM) introduced by Kraetzel et al. [4] and further developed by Scheinerman and Young [10]. The RDPM is a latent space model, with pairwise connection probabilities defined by the dot products of the associated vectors. Scheinerman and Young showed that, for a broad class of initial distributions, the RDPM generates networks that have properties commonly seen in social networks, such as short average path length and high clustering, [10]. Later, Scheinerman and Tucker gave an efficient algorithm for estimating the latent vectors from a given network [7]. O'Connor et al. have recently adapted a logistic version of the RDPM for community detection [5].

Poisson versions of the stochastic block model have previously been used to simplify probabilistic computations. Recently, several other generative models have been developed for weighted networks [1, 6, 8]. Many of these methods can be realized as special cases of our model, by limiting the dimension of the latent space or restricting to discrete distributions.

Our Model (WRDPM)

In order to generalize the RDPM for weighted networks we allow the edges to be drawn from an arbitrary parametrized probability distribution instead of a Bernoulli trial. We call our model the Weighted Random Dot Product Model (WRDPM). In order to accommodate more complex distributions, we incorporate several latent vectors for each node, one for each parameter. Our generative process proceeds as follows:

1) Begin by selecting the number of desired nodes n.
2) Select a parametrized probability distribution $P : \mathbb{R}^k \to \mathbb{R}$ for the edge weights.
3) For each parameter of P, select a dimension d_i and distribution W_i defined over $\mathbb{R}^{d_i}$.
4) For each node, $1 \leq j \leq n$, select k vectors (one from each parameter space), $W_i^j \in \mathbb{R}^{d_i}$, according to distribution W_i.
5) Finally, place an edge between each pair of nodes (ℓ, j) with weight drawn from: $P(\langle W_1^\ell, W_1^j\rangle, \langle W_2^\ell, W_2^j\rangle, \dots, \langle W_k^\ell, W_k^j\rangle)$.

This process gives rise to an undirected weighted network with no self–loops. An equivalent generalization can be given for the directed dot product networks presented in [10]. Throughout this poster we will be concerned with the case where P is chosen to be a distribution over the natural numbers, usually the Poisson distribution.

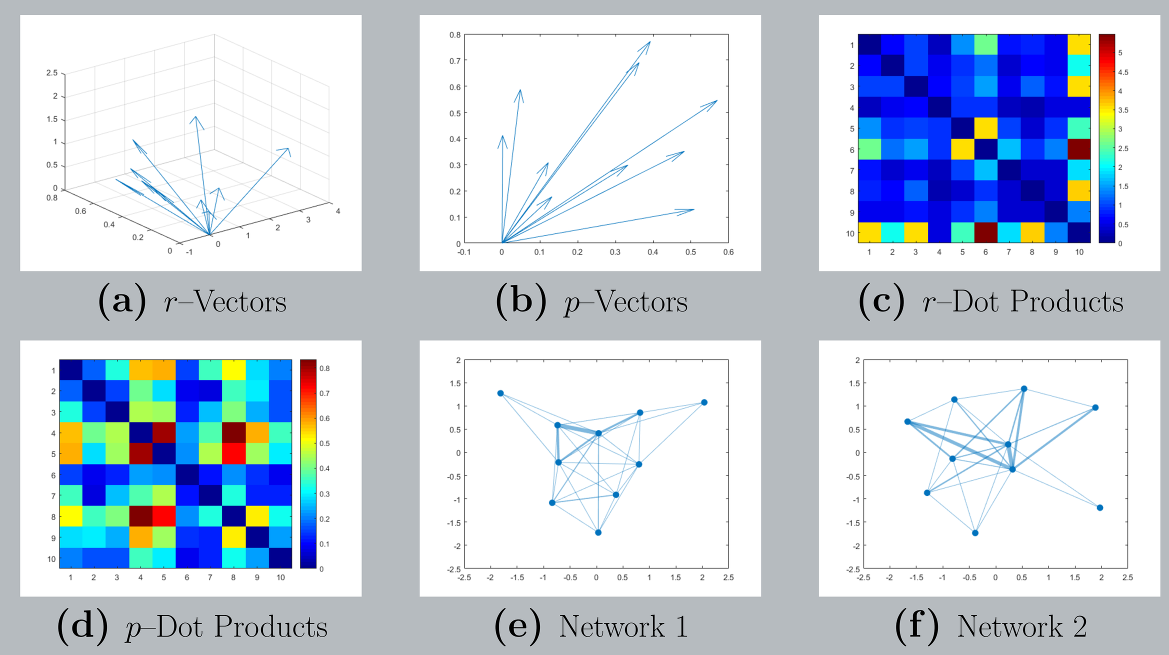
Special Cases and Variations

As in the case of the RDPM, restrictions of this model provide natural generalizations of other commonly studied simple network generative processes.

- When W_i is a distribution over a finite set of vectors in $\mathbb{R}^{d_i}$ we have a generalized stochastic block model as in [1].
- Further restricting W_i to a single vector describes a generalized Erdős–Rényi model.
- Selecting $d_i = 1$ gives a model where each node is associated to a single strength parameter generalizing the approach presented in [6].
- Conversely, restricting the distributions W_i to $S^{d_i-1} \in \mathbb{R}^d$ gives a model where the connection strengths only depend on the angle between vectors, which serve as a proxy for similarity and community membership.

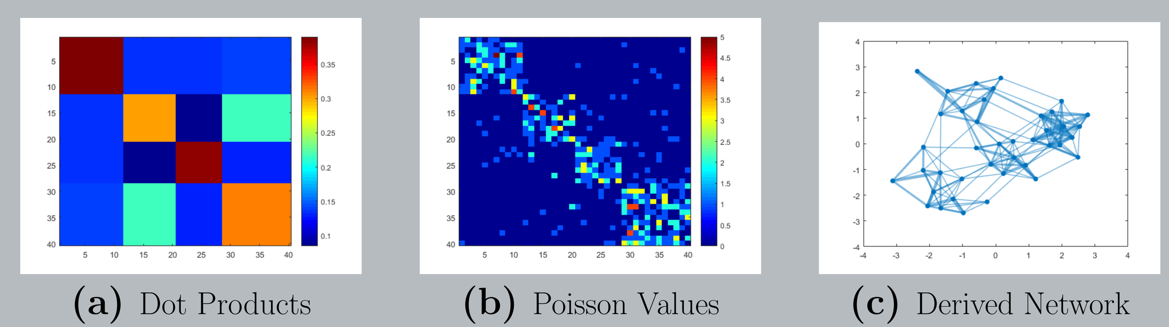
Generative Examples

In order to demonstrate the WRDPM process, we display the intermediate steps of two constructions. In Figure 1, we given a full WRDPM with P chosen to be the negative binomial distribution, while in Figure 2 we construct a Poisson stochastic block WRDPM.



(a) r–Vectors **(b)** p–Vectors **(c)** r–Dot Products **(d)** p–Dot Products **(e)** Network 1 **(f)** Network 2

Figure 1: Construction process of two negative binomial WRDPM networks parametrized by p and r. We choose $d_r = 3$, W_r 3–variate half–normal, $d_p = 2$, and W_p uniform on $[0,1]^2$. The dot products in (c) and (d) parametrize the edge weights in (e) and (f), e.g. $\langle W_r^1, W_r^2\rangle = .668$ and $\langle W_p^1, W_p^2\rangle = .179$ and so the edge weight $A_{1,2}$ is drawn from NegativeBinomial(.668, .179).



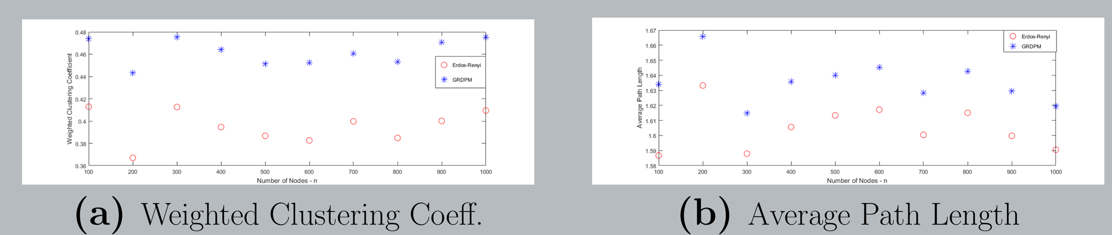
(a) Dot Products **(b)** Poisson Values **(c)** Derived Network

Figure 2: Construction process of a Poisson stochastic block WRDPM with $d = 4$ and W chosen to be the uniform distribution over $\{[.6,.1,.1,.1],[.1,.5,.05,.2],[.1,.05,.6,.1],[.1,.2,.1,.5]\}$.

Small World Properties

In [10] the RDPM is shown to exhibit clustering and small network diameter for a large class of probability distributions, as the number of nodes goes to infinity. For the Poisson WRDPM with parameter λ, we can extend these results to distributions W_λ over $\mathbb{R}^d$ where W_λ has compact support and $\mathbb{P}(\langle W_\lambda^i, W_\lambda^j\rangle > 0) = 1$, generalizing the inner product condition necessary for the RDPM case. The key idea is to use $1 - e^{-\langle W_\lambda^i, W_\lambda^j\rangle}$ as an edge existence probability. Similarly, we can use $\sum_{\ell=k}^{\infty} \frac{\langle W_\lambda^i, W_\lambda^j\rangle^\ell e^{-\langle W_\lambda^i, W_\lambda^j\rangle}}{\ell!}$ to generalize the metrics to a weighted version considering only edges of weight at least k.

We can also evaluate these small world properties empirically by generating synthetic networks with the WRDPM and comparing them to aErdős–Rényi WRDPM networks of the same parameter. The results of such a computation are shown in the figure below:



(a) Weighted Clustering Coeff. **(b)** Average Path Length

Figure 3: Comparison of weighted clustering coefficient and average path length between synthetic networks and Erdős–Rényi networks as n varies. The weight distribution is Poisson with $d = 2$ and W taken to be the uniform distribution over $[0,1]^2$. The Erdős–Rényi WRDPM is constructed to match the expected degree of the original WRDPM.

Why WRDPM?

1) **Generality:** Since P can be any parametrized probability distribution, the WRDPM can be used to model networks derived from a wide variety of real–world data.
2) **Geometry:** As a latent space model, the WRDPM provides an embedding of the network into Euclidean space, allowing us to use tools from linear algebra to analyze our networks.
3) **Interpretability:** Using the dot product to parametrize the network distinguishes the WRDPM from other latent space models where distance is the standard measure. This approach allows us to understand the embedding in terms of the magnitude of each vector, which captures the corresponding node's propensity to communicate, and the direction of each vector, which captures the standard latent space notion of node similarity [7, 10].

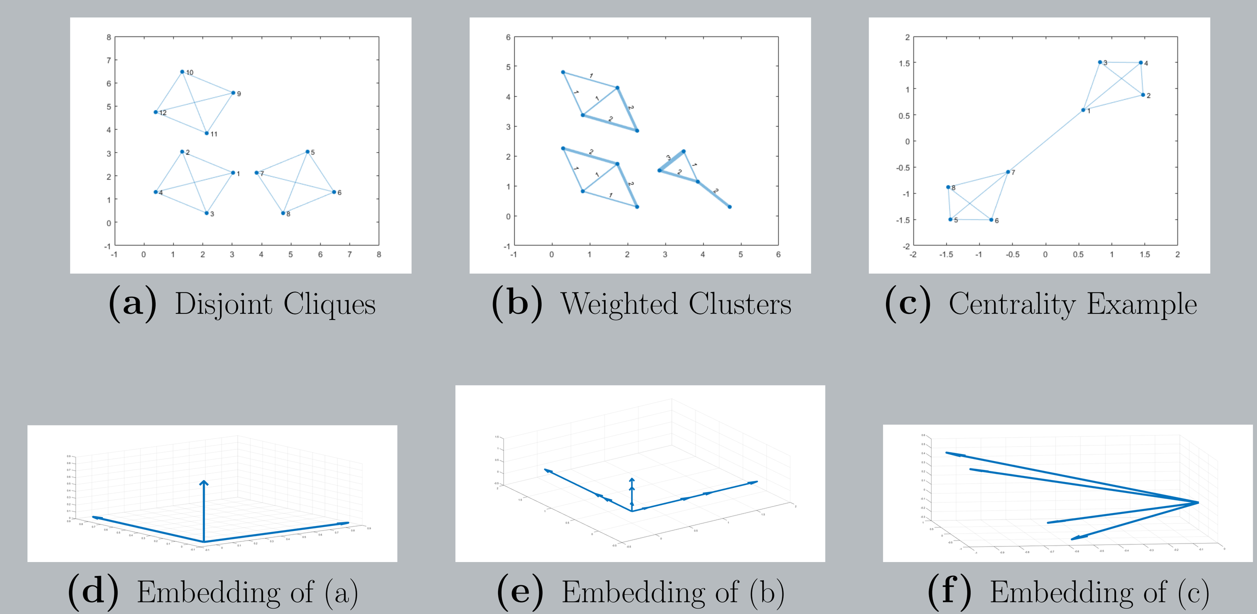
Inference Methods

The inverse problem for the Poisson WRDPM can be solved using a generalization of the iterative algorithm presented in [7]. That is, given a weighted network with adjacency matrix A, we construct a collection of vectors $\{X_i\}_{i=1}^n \subset \mathbb{R}^d$ so that $\langle X_i, X_j\rangle \approx A_{i,j}$ for all $i \neq j$, using a matrix factorization technique. Geometric methods can then be used to study the vectors in this lower dimensional representation in order to learn about the network. For example, a version of the angular clustering procedure in [7] can be used for community detection on the embedding.

Communities and Centrality

The community structure of a multi–network has a strong connection to the geometry of the associated embedding, as the embeddings of networks with well–defined communities will separate into nearly orthogonal components. This is intuitive in light of the interpretation of the vector angle as a similarity measure [7, 10]. The dimension of the embedding plays a key role in community detection, since it determines the number of available orthogonal subspaces.

The magnitudes of the vectors learned from the WRDPM embedding capture a version of centrality that is both related to the degree of the node, or its propensity to form connections [4, 7, 10], as well as the betweeness of the node with respect to community structure. Since the communities are nearly orthgonal, nodes that share edges between communities must have a higher magnitude in order to effectively approximate the network structure. The figure below explores these ideas on some toy examples. In Figure 4(f), the two nodes that connect the communities are sent to the two upper vectors of length 2, while the six nodes with only intra–clique links are sent to the lower vectors of length 1.



(a) Disjoint Cliques **(b)** Weighted Clusters **(c)** Centrality Example **(d)** Embedding of (a) **(e)** Embedding of (b) **(f)** Embedding of (c)

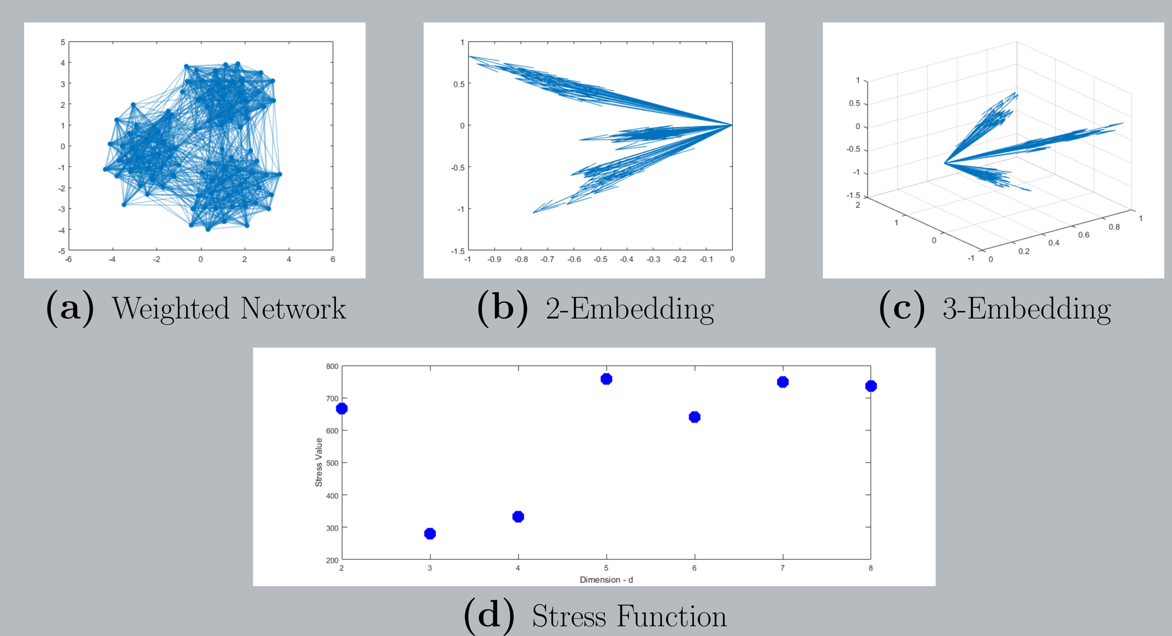
Figure 4: Toy examples illustrating the basic properties of the WRDPM embeddings. Figures (a) and (d) show the orthogonal embedding of disjoint cliques, Figures (b) and (e) show the centrality aspect of magnitude for disjoint embeddings, and Figures (c) and (f) show the effect of an inter–community links on magnitude.

Dimension Selection

Considering the trivial case of ℓ disconnected communities, each with z_ℓ nodes, we observe that the sum of intra–community dot products is $\sum_{i=1}^{\ell} \binom{z_i}{2}$ and the sum of inter–community dot products is zero. This leads us to define a stress function of the form:

$$s(d) = \sum_{i=1}^{d} \binom{z_i}{2} - \mathrm{s}_{\mathrm{intra}}(d) + \mathrm{s}_{\mathrm{inter}}(d),$$

where $\mathrm{s}_{\mathrm{intra}}(d)$ is the sum of the dot products of all intra–community pairs and $\mathrm{s}_{\mathrm{inter}}(d)$ is the sum of the dot products of all inter–community pairs. The dimension d that minimizes this value and its associated clusters, are then appropriate candidates for partitioning the multi–network.

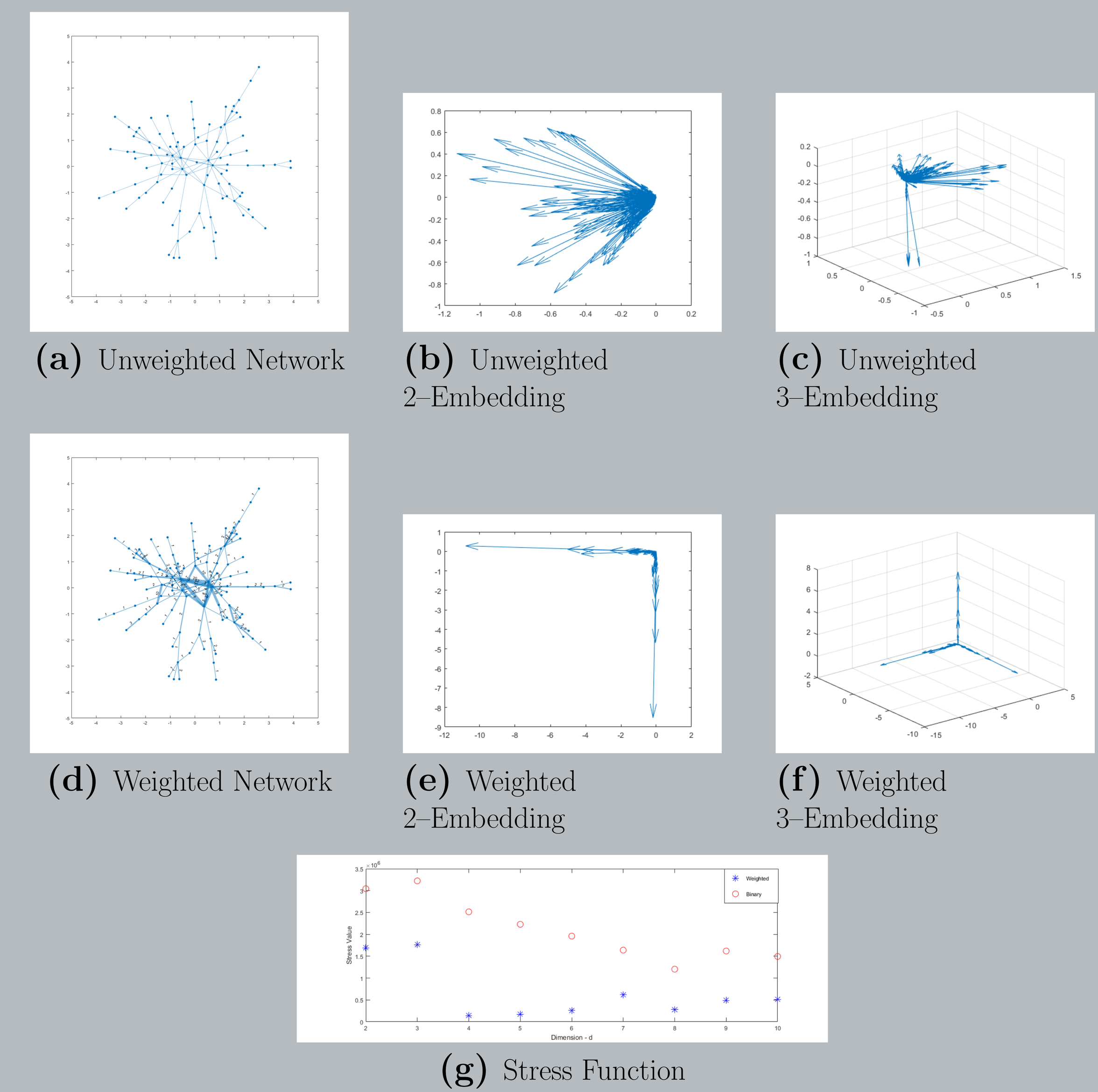


(a) Weighted Network **(b)** 2-Embedding **(c)** 3-Embedding **(d)** Stress Function

Figure 5: Comparison of WRDPM embeddings of a weighted network (a) as the dimension of the embedding varies. The minimum value occurs at $d = 3$, matching the correct structure.

Application: Coauthorship Networks

Scientific collaboration networks are often studied as a proxy for the professional interaction networks of researchers. In the most common formulation of these networks, the nodes are scientists and two scientists are connected by an edge if they have written a paper together. However, these interactions also have a natural multi–network structure, where the number of edges between two scientists is computed as a (weighted) sum of the papers coauthored by them. We consider the large connected component of a collaboration network from the field of computational geometry [2], with 7,343 authors and 11,898 publications, where the edges are weighted by the number of co–publications. To compare to the RDPM we also consider the unweighted underlying collaboration network.



(a) Unweighted Network **(b)** Unweighted 2–Embedding **(c)** Unweighted 3–Embedding **(d)** Weighted Network **(e)** Weighted 2–Embedding **(f)** Weighted 3–Embedding **(g)** Stress Function

Figure 6: Comparison of WRDPM embeddings to binarized embeddings of a weighted coauthorship network. The weighted embeddings correspond to much lower stress values.

Future Work

The WRDPM is a very general model and the material presented here is a preliminary outline of its properties. We intend to extend this research in several ways:

- Considering non–Euclidean embeddings motivated by information geometry
- Computing spectral bounds for specific choices of P and W
- Developing methods for time series derived networks
- Proving expected bounds for specific families of distributions
- Constructing factorization algorithms for specific classes of networks

References

[1] C. Aicher, A.Jacobs, and A. Clauset: *Learning Latent Block Structure in Weighted Networks*, Journal of Complex Networks, 3(2), (2015), 221–248.

[2] V. Batagelj and A. Mrvar: *Pajek datasets*, (2006), http://vlado.fmf.uni-lj.si/pub/networks/data/.

[3] B. Karrer and M. Newman: *Stochastic blockmodels and community structure in networks*, Physical Review E, 83(1), (2011), 1-11.

[4] M. Kraetzel, C. Nickel, and E. Scheinerman: *Random Dot Product Networks: A model for social networks*, Preliminary Manuscript, (2005).

[5] L. O'Connor, M. Médard, and S. Feizi: *Clustering over Logistic Random Dot Product Graphs*, ArXiv: 1510.00850, (2015).

[6] J. Ranola, S. Ahn, M. Sehl, D. Smith, and K. Lange: *A Poisson Model for random multigraphs*, Bioinformatics, 26, (2010), 2004–2011.

[7] E. Scheinerman and K. Tucker: *Modeling graphs using dot product representations*, Computational Statistics, 25, (2010), 1–16.

[8] T. Shafie: *A Multigraph Approach to Social Network Analysis*, Social Structure, 16, (2015), 1–21.

[9] A. Thomas and J. Blitzstein: *Valued Ties Tell Fewer Lies: Why Not to Dichotomize Network Edges With Thresholds*, ArXiv: 1101.0788, (2011), 1–36.

[10] S. Young and E. Scheinerman: *Directed Random Dot Product Graphs*, Internet Math, 5, (2008), 91–112.