# Structured Regression on Multilayer Networks*

Athanasia Polychronopoulou† Zoran Obradovic†

**Abstract**

Until recently, research in social network analysis was focused on single layer networks, where only one type of links among nodes is considered. This approach does not consider the variety of interactions that exist among nodes, resulting in the loss of a large amount of information. In the last few years there is an advanced interest in multilayer networks analysis, where multiple types of nodal connections are considered jointly. In most approaches however the contributions of the various interactions are averaged, resulting again in the loss of information. In this work we present a structured regression model for node attribute prediction in multilayer networks. Our Gaussian Conditional Random Fields model is designed to maximize the information gained from the use of data with multiple layers of graphical structure. Our model accommodates graphs with layers that share the same set of nodes allowing for missing nodes and unobserved connections. At the same time it models the evolution of such networks over time without requiring the addition of a new layer. We present evidence that this model outperforms the traditionally used one and that it offers predictive accuracy that increases as the number of layers used grows, on both synthetic data and challenging real world applications such as predicting citation count and sepsis hospitalization admission rate at all hospitals in California.

## 1 Intoduction

Social network analysis, as emerged from modern sociology, uses graph theory to characterize social structures. This idea of traditional sociograms, has inspired the use of graphs for the representation of different kinds of structures in different disciplines. Graphs are broadly used for the representation of biological networks [1], scientific collaboration networks [2] or even disease networks [3]. As a result, networks with a variety of properties for nodes and links have been created, from networks with multiple kinds of nodes [4], to networks with time dependent existence of links [5].

Up until recently, research in social network analysis was focused on single layer networks, where only one kind of node interactions is considered. However, in natural systems the entities interact in multiple ways. These systems can be represented by a set of graphs over the same vertices, with the edges of each graph capturing a different type of interaction. The common practice for analyzing such systems is their reduction to a network where the vertices are connected by only a single set of links, resulting in the loss of the information carried by the heterogeneity of their interactions.

In the last few years a variety of studies have been published using terminologies such as "multirelational" [6] or "multiplex" networks [7]. Researchers have tried to develop a framework to study and describe these kinds of systems in a comprehensive fashion [8]. Alongside these efforts, the problem of clustering vertices based on multiple graphs is studied, in both unsupervised and semi-supervised settings [9]. Other areas that have also been investigated, are graph layer reduction [10] and link prediction [6].

In this paper we propose a Gaussian Conditional Random Fields model (GCRF) that is specifically designed for use with multilayer systems. Thereby, we present a regression method for structured prediction of node attribute values. Several GCRF models have so far been successfully applied for regression problems in different domains [11, 12]. However in most cases either graphs of only one layer were used or the integration of information from several layers was computed by averaging. The GCRF model proposed in this work aims to maximize the information gained from data with multiple layers of graphical structure. Our model considers the possible correlations among layers and integrates the information in a more educated way. This way instead of averaging the information from the heterogeneous connections, our model aggregates it, offering a more precise prediction of the node attributes.

In this study we will use the more general term "multilayer network" as introduced in [8] to describe systems where the nodes are connected via multiple types of links. An example of such a network is shown in Figure 1, where three types of links are considered simultaneously among a constant set of nodes and over

*This research was supported in part by DARPA grant FA9550-12-1-0406 negotiated by AFOSR, NSF BIGDATA grant 14476570 and ONR grant N00014-15-1-2729. Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality, provided part of the data used in this study.

†Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, USA,
Email: n.polychr, zoran.obradovic@temple.edu

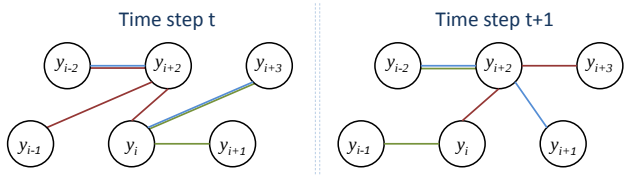


Figure 1: A multilayer network with three layers, each represented by a different color of links

multiple time steps. The objective of the structured regression task would be the prediction of the response variable $y$ for all nodes at the future time step. In our approach we assume that different layers share the same set of nodes. Nevertheless, the absence of nodes from any layer does not affect the accuracy or methodology of our approach. Furthermore notice that, as in [13], our model also allows us to include the evolution of such networks over time. In the GCRF framework we are able to include several timesteps in our analysis and the correlations of nodes among those timesteps without having to include an additional layer, as the evolution over time can be included by a simple join of the matrices that describe the different connections among nodes in different timesteps. In the experimental sections we will present evidence that this model outperforms the traditionally used model on both synthetic and real world data. Furthermore we use synthetic data to gain a detailed perspective of the various properties of the model, and we show that this model offers predictive accuracy that increases as the number of layers used grows.

## 2 Structured Regression on Networks by Conditional Random Fields

**Problem Definition:** The objective would be the prediction of a real valued N-dimensional vector of possible output $\mathbf{y} = (y_1, ..., y_N)$, describing a nodal attribute. For this regression task we follow a structured learning approach, where we are given all the input nodal attributes $\mathbf{x} = (x_1, ..., x_N)$ and the dependencies between the outputs $y$, represented by a set of graphs, each describing one of the multiple types of connections among the nodes.

In the continuous case, Conditional Random Fields (CRF) compute the probability $P(\mathbf{y}|\mathbf{x})$ of the real-valued possible output $\mathbf{y} = (y_1, ..., y_N)$ given the input $\mathbf{x} = (x_1, ..., x_N)$ by an equation of the form [14, 15]:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z} exp(-\sum_{i=1}^{N}\sum_{k=1}^{K}\alpha_k(y_i - R_k(\boldsymbol{x}))^2 - \sum_{j\sim i}\beta(\boldsymbol{\psi}, \boldsymbol{x})(y_i - y_j)^2) \quad (2.1)$$

where $\alpha$ and $\psi$ are parameters and $j \sim i$ denotes that the two nodes $i$ and $j$ are neighboring nodes. $Z$ is a normalization constant, which is an integral over $y$ of the term in the exponent. In the first term, $R_k(\boldsymbol{x})$ is the $k^{th}$ baseline predictor and $\alpha_k$ is the corresponding weight. In general the baseline predictor can use the entire input $\mathbf{x}$ to predict the value of $y_i$ but does not use the correlations among the outputs, so that we will be referring to $R_k(\boldsymbol{x})$ as unstructured predictor. We can introduce an arbitrary number of unstructured predictors and their relevance will be learned from data, as more relevant predictors will be given greater weights. Also, the quadratic function is easy to interpret as it dictates a value of $y_i$ close to $R_k(\boldsymbol{x})$. In the second term, the function $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ describes the way that the outputs $y_i$ and $y_j$ are correlated. The quadratic form of the potential forces the values of $y_i$ and $y_j$ to be more similar as the value of function $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ increases. The $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ function that is commonly used is:

$$\beta(\boldsymbol{\psi}, \boldsymbol{x})^{(i,j)} = \sum_{l=1}^{L}\psi_l S_{ij}^{(l)}(\boldsymbol{x}) \quad (2.2)$$

where $S_{ij}^{(l)}(\boldsymbol{x})$ represents the similarity between nodes $i$ and $j$, as is defined for the graph layer $G^{(l)}$. The value of $S_{ij}^{(l)}(\boldsymbol{x})$ is zero if the nodes $i$ and $j$ are not connected in this specific graph layer. Similar to the case of the association potential, $\psi_l$ is the corresponding weight that during training determines the relevance of the similarity matrix.

In general regression problems, both learning and inference can be difficult, due to the integration over $y$ in the normalizing constant $Z$. However, our approach has been shown to allow efficient learning and inference [11] as it corresponds to a multivariate Gaussian distribution. It is easy to notice in (2.1) that the exponent of the probability distribution $P(\mathbf{y}|\mathbf{x})$ is a quadratic function in terms of y, and that it can be written as:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})) \quad (2.3)$$

We will therefore refer to the model of (2.1) as Gaussian Conditional Random Fields (GCRF). The learning task is to choose the parameters $\alpha$ and $\psi$ to maximize the conditional log-likelihood of the training data. Following the technique used in [12] that applies exponential transformation on $\alpha$ and $\psi$ parameters to guarantee that they are positive, we can ensure the feasibility of the learning task [11, 16] and at the same time convert the learning task to an unconstrained optimization problem and use standard gradient descent to solve it.

The inference task is to find the outputs $y$ for a given set of observations $\mathbf{x}$ and estimated parameters $\alpha$ and $\psi$ such that the conditional probability $P(\mathbf{y}|\mathbf{x})$ is maximized. In the case of GCRF, since the model

is Gaussian, the prediction for $y$ will simply be the expected value of the distribution, which is equal to the mean $\mu(\boldsymbol{x})$. To obtain explicit expressions for $\mu(\boldsymbol{x})$ and $\Sigma(\boldsymbol{x})$, we match equations (2.1) and (2.3) under the assumption that the matrix $\beta^{(i,j)}(\boldsymbol{\psi}, \boldsymbol{x})$ is symmetric. In order to analytically write the equations, let us first define the $N \times N$ matrix $\boldsymbol{Q}$ with elements:

$$Q_{ij} = \begin{cases} \sum_{k=1}^{K} \alpha_k + \sum_{k \neq j} \beta^{(i,k)}(\boldsymbol{\psi}, \boldsymbol{x}), & i = j \\ -\beta^{(i,j)}(\boldsymbol{\psi}, \boldsymbol{x}) & i \neq j \end{cases} \tag{2.4}$$

Then, the mean and the inverse of the covariance matrix of $P(\mathbf{y}|\mathbf{x})$ can be calculated as $\mu = \Sigma b$ and $\Sigma^{-1} = 2Q$, where $b_i = 2\sum_{k=1}^{K} \alpha_k R_k(\boldsymbol{x})$

## 3 GCRF for Regression on Multilayer Networks

The GCRF framework described above can be used to model applications of multilayer networks, as it allows for the integration of knowledge from multiple layers of connections among nodes, represented by the similarity matrices. However the integration of all this information is so far being done in a direct and simplistic manner using the function given by (2.2). This function is averaging the contributions of various layers and does not take into account any possible correlations among them. In order to create a model that incorporates all the information from the multilayer network so that the information gain and the accuracy are maximized, we propose the following function:

$$\beta(\boldsymbol{\psi}, \boldsymbol{x})^{(i,j)} = \sum_{\lambda=1}^{L} \psi_\lambda (\sum_{l} \psi_l S_{ij}^{l}(\boldsymbol{x}))^\lambda \tag{3.5}$$

where $\lambda$ is a variable that takes all the integer values between one and $L$, the number of layers of the graph. In this function, notice that the first term (for $\lambda = 1$) corresponds to the summation that is currently applied in the GCRF framework. For higher values of $\lambda$ more complicated, nonlinear terms are created. Thereby this function is composed of terms that contain all the possible products of the similarity matrices of the graph's various layers. More specifically we can see that for $\lambda = 2$ we get terms that contain all the possible combinations of two matrices, or the second order correlations of layers. Similarly, for $\lambda = 3$ we get terms that contain all the possible combinations of three matrices, or the third order correlations of layers. This way we take into account all possible correlations among the layers of the graph, allowing the parameters $\psi_\lambda$, learned during training, to determine the importance of the correlations of order $\lambda$. We will refer to the GCRF model that utilizes this $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ function as Power Function GCRF (PF-GCRF).

In general GCRF models have been shown to be a very powerful tool in structured regression [11, 12], which led many researchers to focus recently on optimizing the running time of GCRF in order to enable its applications on large graphs [20]. AS PF-GCRF considers all possible correlations among the layers of the multilayer network, a time complexity issue is automatically introduced. However, the main focus of this paper is not to provide a fast model for structured regression but to improve the accuracy performance of GCRF and to introduce evidence supporting the significance of the various layers information aggregation.

## 4 Baseline Regression Models

In the following experimental sections, we compare the performance of PF-GCRF with a variety of baseline predictors. First, we use the unstructured predictor, used as a standalone predictor to get a lower baseline. Then we also compare to the original GCRF, as proposed by [11], that utilizes the $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ function in (2.2).

None of the above baseline predictors however, take into account correlations among the layers of the graph and therefore we introduce one more predictor to act as a baseline, inspired by the work of [17]. The original model of the $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ function in (2.2), models the data under the assumption that two nodes are expected to have similar values of $y$ if the similarities state on average that they are similar. However, this assumption is not always ideal as it is significantly affected by links that may be unobserved in a specific layer, or layers that are less accurate. Therefore this new baseline predictor is designed in such a way that allows us to directly build propositions such that two nodes should have similar values of $y$ if at least one of the similarities states that they are similar, or that two nodes have similar values of $y$ only if all the similarities state that they are similar. More specifically, these propositions can be generated by a function $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ that resembles a probabilistic "logical gate". Inspired by the use of logical gates, the Noisy-OR function, represents a non-deterministic disjunctive relation between an effect and its possible causes and has been used in artificial intelligence [17,18]. In a similar spirit we can also use a Noisy-AND function to finally create a prior that picks up for the strongest (OR) or the universal (AND) support among the various information sources. The following equation represents the simplest form of combining the probabilistic OR and AND gates:

$$\beta(\boldsymbol{\psi}, \boldsymbol{x})^{(i,j)} = \sum_{l=1}^{L} \psi_l S_{ij}^{(l)}(\boldsymbol{x}) + \psi_{L+1}(\prod S_{ij}^{l}(\boldsymbol{x})) + \psi_{L+2}(1 - [\prod (1 - S_{ij}^{l}(\boldsymbol{x})]) \tag{4.6}$$

In this approach the first term of the $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ function

corresponds to the summation that is currently applied in the GCRF framework. The other two terms aim to add some flexibility to the model so that the propositions mentioned above can be taken into account. Again the parameters $\psi$ are learned from training data so that the importance of the various layers and also the two probabilistic propositions are adjusted to the given application. We expect that the importance of each of the propositions may be application specific. For example in multilayer networks where the nodes are Internet users, their friendship in facebook is usually enough for them to be assumed as interacting users, and the absence of a strong link between them in the youtube layer does not demote their connection. In this case the use of OR gate is expected to benefit GCRF. The combination of those terms is expected to guarantee application independence and at the same time add an additional degree of freedom to the model. We will refer to the GCRF model that utilizes this $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ function as Logical Gates GCRF (LG-GCRF).

## 5 Experiments on Synthetic Networks

As a first step, we perform experiments on synthetic data as they enable us to avoid the various peculiarities of real-world data and focus on analyzing the properties and performance of our proposed model. We report our findings on two major experiments. The first is applied on data generated using Erdős - Rényi graphs [19] and node attributes generated from a normal distribution. The second is applied on data sampled from the GCRF model, treated as a generative model. In the following sections we present the detailed data generation process followed by our results.

### 5.1 Experiments on Multilayer Random Graphs

**5.1.1 Exploring the information gain from multilayer graphs** The goal of this set of experiments is to investigate the way that PF-GCRF incorporates the information carried by the layers of the graph and compare the information gain of this model with the baselines. Therefore we design a set of experiments, where we study the change in the prediction accuracy as we utilize more and more layers of the graph. Then we can show that PF-GCRF combines the information of the layers in an accumulative and not averaging way.

**Generation of $y$ and $R_k(x)$ values:** The first step of our data generation process is the creation of an Erdős - Rényi graph with 300 nodes. We then assign each node with an output value $y$ chosen randomly from a standard normal distribution. For the creation of the unstructured predictor $R_k(\boldsymbol{x})$ we use the values of $y$ and we add noise, sampled from a normal distribution (N(0, 2/3)).

**Generation of 8 layers:** In order to create a multilayer graph structure we assume that the links included in the various layers of data, all originate from a true and unobserved network ( [17] has used a similar assumption). In this context all the similarity matrices of our data are instantiations of the true and unobserved similarity matrix. To this end, we first create the weights of the edges of our true network, as $w_{ij} = e^{(-(y_i - y_j))}$. Then we create the layers by sampling a percentage of the edges of the true network and adding noise to their weights.

More specifically, we create the values of $y$ and $R_k(x)$ along with five different layers using the methodology described above. In those first five layers we add random noise to the true edges that is sampled from N(0, 0.1), N(0, 0.2), N(0, 0.5), N(0, 1), N(0, 1) correspondingly and randomly remove a percentage of links (20% in the first layer and 40% in the rest). Then we create layer 6 with random values sampled from uniform distribution, layer 7 with random values sampled from standard normal distribution and layer 8 with a completely uninformative similarity measure where all the nonzero links are equal to one.

**Evaluation of layers:** The eight similarity matrices have significantly different informative strengths. Considering the noise that has been added during the formation of the layers, it is clear that the first layer, is the more informative one. The similarity matrices of layers four and five offer the least amount of information, while layers six and seven are completely random. Finally, the eighth layer is completely uninformative.

**GCRF train and test data:** We use the previously described synthetic graph as the training data set for all the models. The evaluation is done by repeating the process to create a new completely independent test data set.

**Experimental Setup:** In order to investigate the ability of the GCRF models to incorporate information from various data sources, we design an experiment where we start from using only one layer of the graph and then we repeat the process adding one layer of information at a time. Given that we have created eight layers, both training and testing of the models is done eight times, so that every time we use one more similarity matrix. As the first of the layers has the more informative similarity matrix, we start by including this layer and then we add one by one the less informative layers. This way we can investigate not only the way that the models incorporate the information offered but also their robustness when uninformative similarity matrices are included. We report the resulting accuracy, at the top

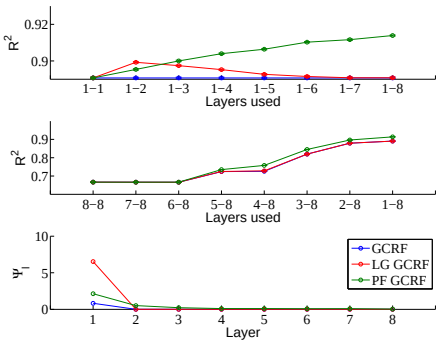


Figure 2: Top: $R^2$ of the three models as the number of layers used changes. At first only layer 1 is included (1-1) and then all layers between layer 1 and $\ell$ (1-$\ell$) are included for $\ell$=2,.., 8. Middle: $R^2$ of the three models: At first only layer 8 is included (8-8) and then all layers between layer $\ell$ and 8 ($\ell$-8) are included for $\ell$=7,.., 1. Bottom: Parameters $\psi_l$ learned by the three models when all eight layers are included

part of Figure 2.

Next, we investigate the behavior of the models when there is not such an overpowering layer, as the first one, included from the beginning. Therefore we run the same experiment but this time start by including only the last layer, which is the least informative, and then we add one layer at a time, this time in decreasing order. The results are reported in the middle part of Figure 2.

Notice that we repeat these experiments ten times, so that the reported $R^2$ values are average values over the 10 repetitions. The variance of the $R^2$ values in those ten repetitions has a maximum value of 0.0004 and although it is not distinguishable it also included in Figure 2.

**Results:** The value of our results can be seen in Figure 2, where the change in the accuracy of the models by the inclusion of more layers is plotted. In the top part of the Figure we can see that, while the accuracy of the original GCRF model remains practically constant, as a result of the averaging properties of the corresponding $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ function, the accuracy of the other two models changes. The $R^2$ of LG-GCRF originally increases, however as less informative similarity matrices are added the accuracy drops. On the other hand, the accuracy of PF-GCRF continues to increase. The intuition behind that is that although the newly added similarities are not as informative, their addition allows one more term to the corresponding $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ function, which gives the model one more degree of freedom, and it therefore captures more correlations and performs better.

In the middle part of Figure 2 we can see that for the first three setups, where only the uninformative layers eight, seven and six are included, all three models have a constant accuracy. However, starting from the inclusion of layer five the accuracy starts increasing with the PF-GCRF outperforming the other two models. Notice that, in this middle part of Figure 2, the lines for GCRF and LG-GCRF overlap.

Finally, one way to look further into the reasons why PF-GCRF performs best is by looking at the learned parameters. In the bottom part of Figure 2 we can see the parameters $\psi_l$ learned by the three models and for each of the layers $l$, in the case where all eight layers of the graph are included. We notice that GCRF and LG-GCRF learn a parameter for the first layer that is significantly higher than all the other ones. This way the models determine that the first layer is the only informative one, as the rest of the parameters have values that are significantly lower. On the contrary, PF-GCRF assigns higher values to the parameters, and through that, higher importance to layers two and three. As the contribution of the second and third layers cannot be comparable to the contributions of the rest, we can state that the parameters learned from PF-GCRF for the first three layers represent more realistically the informative power of the corresponding similarity matrices.

**5.1.2 Additional Experiments on Random Graphs** The previously described set of experiments utilized an Erdős - Rényi graph. To further characterize the generalization of our results, we conducted a set of experiments using synthetic graphs, obtained by different graph generation method [22–26]. Although the detailed results are omitted we have confirmed that the improvement of PF-GCRF noticed in the previous section is evident regardless of the graph generation process used during synthetic data creation.

We have also verified that the PF-GCRF model proposed here, performs best regardless of the accuracy and value of the unstructured predictor that is utilized. However, the results also indicate that, as the quality of the unstructured predictor drops and the room for improvement in the prediction grows, so does the benefit from using PF-GCRF. The reason is that, when the accuracy of the unstructured predictor drops, the importance of the information carried by the graphical layers grows, so that the way that this information is combined becomes more important.

**5.2 Experiments on Multilayer Networks Generated by GCRF** In the previous set of experiments the values of $y$ were generated from a Gaussian distribution, favoring the application of Gaussian models. In this set of experiments we evaluate the accuracy of PF-GCRF when data that hold different intrinsic structures are used. To this end we use the GCRF model as

a generative model to create synthetic data that hold a specific structure defined by the $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ function.

**Generation of values for $\boldsymbol{R(x)}$ and layers:** We first create an Erdős - Rényi graph with 300 nodes, and then similarly to previous sections we assign to each node a value sampled from a standard normal distribution. This value now corresponds to the unstructured predictor $R_k(\boldsymbol{x})$. We use this value and equation: $w_{ij} = e^{(-(R_i - R_j))}$ to create the edges of the true and unobserved network and then we construct the various layers by adding noise to this value sampled from N(0, 0.2) and randomly removing a percentage of links (40%). We remove the high correlation between the values of $R(x)$ and the similarity matrices by adding extra noise to $R(x)$, sampled from N(0, 0.5).

**Generation of $\boldsymbol{y}$:** To create data with various internal structures we modify the $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ function used. As seen in Table 1 we chose a variety of functions $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ that produced data with different properties. More specifically $\beta^{AND}$ and $\beta^{OR}$ create data with enhanced probabilistic AND and OR functionality correspondingly. Function $\beta^{AV}$ produces data for which the contribution of the various layers is averaged, while $\beta^{AV}_{AND}$ and $\beta^{AV}_{OR}$ create data that average the contributions taking into account the probabilistic AND and OR functionality. More complex data are creates by the last three $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ functions, $\beta^{QS}$, $\beta^{TH}$ and $\beta^{Q}$, that construct data with enhanced correlations among the various layers. Finally in order to create the data we choose a set of parameters $\alpha$ and $\psi$, such that the graph structure is significantly more important than the unstructured predictor ($\psi >> \alpha$) and we run the GCRF model that utilizes those $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ functions to produce the output $y$. As in previous section we run the experiment ten times and we report mean accuracy on Table 1 for all the models and all the synthetic data.

**Results:** From the results of Table 1, we can see that PF-GCRF model, outperforms the original GCRF model in most of the cases. The benefit in the accuracy depends, as expected, on the details of the data structure. For example in the case of $\beta^{AND}$ it is clear that LG-GCRF outperforms the other two models. For $\beta^{OR}$ LG-GCRF outperforms the original GCRF, as expected, while PF-GCRF is also able to capture the data structure and significantly increase the accuracy. The function $\beta^{AV}$ is actually producing data that match exactly the function of the original GCRF and therefore this model is expected to outperform the rest. Furthermore, $\beta^{AV}_{AND}$ and $\beta^{AV}_{OR}$ functions are again designed to create data whose structure benefits the LG-GCRF and this result is reflected on Table 1. In the last three cases, where the data hold higher order correlation, only PF-GCRF is able to capture them and the difference in the model's accuracy is more significant.

Table 1: Accuracy, in terms of $R^2$ for experiments with synthetic networks that were generated using GCRF as a generative model with data of 4 layers and the $\beta(\boldsymbol{\psi}, \boldsymbol{x})$ function shown here

| | R square (E-2) | | |
|---|---|---|---|
| $\beta(\boldsymbol{\psi}, \boldsymbol{x})^{(i,j)}$ | GCRF | LG GCRF | PF GCRF |
| $\beta^{AND} = \prod_l \psi_l S^l_{ij}(\boldsymbol{x}))$ | 72.22 | **99.13** | 77.29 |
| $\beta^{OR} = e^{(1-\prod_l \psi_l (1-S^l_{ij}(\boldsymbol{x})))}$ | 85.61 | 86.60 | **93.50** |
| $\beta^{AV} = \frac{1}{L}\sum_{l=1}^{L} \psi_l S^{(l)}_{ij}(\boldsymbol{x})$ | **98.47** | 98.35 | 98.31 |
| $\beta^{AV}_{AND} = \beta^{AV} \prod_l S^l_{ij}(\boldsymbol{x})$ | 85.12 | **99.65** | **89.54** |
| $\beta^{AV}_{OR} = \beta^{AV}(1-[\prod(1-S^l_{ij}(\boldsymbol{x}))]$ | **98.58** | 98.24 | 98.15 |
| $\beta^{SQ} = (\sum_{l=1}^{L} \psi_l S^{(l)}_{ij}(\boldsymbol{x}))^2$ | 77.04 | 76.92 | **78.53** |
| $\beta^{TH} = (\sum_{l=1}^{L} \psi_l S^{(l)}_{ij}(\boldsymbol{x}))^3$ | 47.71 | 44.54 | **51.68** |
| $\beta^{Q} = (\sum_{l=1}^{L} \psi_l S^{(l)}_{ij}(\boldsymbol{x}))^4$ | 74.71 | 73.39 | **76.14** |

## 6 Experiments on Real Applications

In this section we study the performance of the GCRF models on two real world applications. We present two sets of experiments, using data from different domains. The first experiment uses data of a citation network while the second one utilizes a graph inspired from health analytics, where the nodes represent hospitals.

**6.1 Citation Count Prediction** We use PF-GCRF to predict the citation count of research papers for high energy physics. We utilize a citation network, constructed from the data of 2003 KDD Cup competition [27], using the publications of high energy physics, theory track. Thus we create a bibliographic network that consists of 29,955 papers and 352,807 citations among them, spanning over 11 years. We build a network that is well established and is not very sparse, by focusing on the 800 most-cited papers, written before year 2000. We will track their citation counts starting at year 2000, having this way a 43 month period of observations [21].

For the creation of the multilayer graph, we use those 800 most-cited papers as nodes and we construct the layers of the networks based on the paper's citation history. The first layer corresponds to the historical similarity of the two papers and is based on the Euclidean distance between the two paper's citation counts over a lag of 3 timesteps. More specifically for two papers $i$ and $j$ the historical similarity is written as $S^{hist}_{ij} = exp(-\frac{d^2_{ij}}{Z_h})$ where $d_{ij}$ is the Euclidean Distance

Table 2: Accuracy of citation count prediction in terms of $R^2$

| | Group 1 | Group 2 |
|---|---|---|
| Unstructured | $0.608 \pm 0.004$ | $0.348 \pm 0.016$ |
| GCRF | $0.665 \pm 0.005$ | $0.506 \pm 0.019$ |
| LG GCRF | $0.676 \pm 0.005$ | $0.510 \pm 0.019$ |
| PF GCRF | $0.684 \pm 0.003$ | $0.504 \pm 0.020$ |

of the two paper's citation counts and $Z_h$ a normalization constant that represents the average Euclidean Distance between two papers in the dataset.

The second layer corresponds to the Co-Citer similarity, which is based on the count of papers that cited both papers divided by each paper's individual citation count at a particular timestep, so that:

$$S_{ij}^{cs} = \frac{2 \times CoCitations^{(i,j)}}{Citations^{(i)} \times Citations^{(j)}}$$

Using this multilayer graph, our regression problem is the prediction of the citation count, $y_t$, of each of the papers and for each of the available timesteps, $t$. We train the GCRF models using the citation count $y_{t-1}$ as output variable, the unstructured predictor for $y_{t-1}$, which actually corresponds to the citation count $y_{t-2}$ and the similarity matrices built from the previous timestep $t-2$. Then we use the trained model to predict the citation count $y_t$ from the previous timestep citation count $y_{t-1}$ and the similarity matrices built from data of $t-1$. The use of the previous timesteps for training reduces the timesteps of data that are available for predicting $y$ to 41. The average accuracy of the three GCRF models and the unstructured predictor, over those 41 timesteps is shown in Table 2.

The results of Table 2 are separated in two groups. The first group (Group 1) corresponds to timesteps where the prediction of citation count by the unstructured predictor gave an $R^2$ higher than or equal to 0.5 and the second group (Group 2) corresponds to timesteps with $R^2$ of the unstructured predictor lower than 0.5. The reason for this very high variance of $R^2$ is that some months of the data are inherently more difficult to predict than others. We split the results into those two groups to study the performance of our models in both cases. We can see from the results of Table 2 that for Group 1, both LG-GCRF and PF-GCRF perform better than the original GCRF model. For group 2, the mean accuracy of the PF-GCRF is slightly lower than that of the other two models. However, if we take into account the values for the variance of $R^2$, that are also reported on Table 2 we can see that there is no true difference between the final values of $R^2$ for GCRF and PF-GCRF. A more detailed view of the results is offered on Figure 3 where we can see that in Group 1 PF-GCRF constantly outperforms GCRF with an improvement of

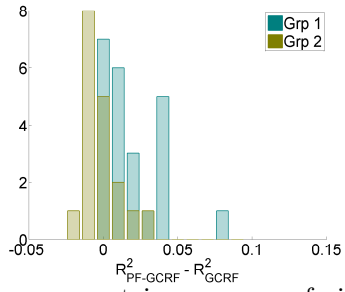


Figure 3: Improvement in accuracy of citation count prediction, in terms of $R^2$, offered by PF-GCRF over GCRF for the two groups of papers

4% or even 8%. In Group 2 PF-GCRF and GCRF show a comparable performance with a difference in accuracy being within the variance of the model's prediction accuracy. Concluding we can see that PF-GCRF in most cases offers an improvement in the prediction accuracy while in the rest of the cases the accuracy remains the same, within the margin of error, as the accuracy of the original GCRF. The same statement is also true for LG-GCRF, although the improvement is lower.

**6.2 Predicting Number of Sepsis Hospitalizations** In the second real world application we utilize the multilayer nature of PF-GCRF to predict the next month's number of sepsis patients per hospital. The data source is the California, State Inpatient Database (SID) [28]. SID is an archive that stores US hospital inpatient stays, is provided by the Agency for Healthcare Research and Quality and is included in the Healthcare Cost and Utilization Project (HCUP). The SID includes inpatient discharge records from community hospitals in the state (California), tracking all hospital admissions at the individual level. We used all data between 2007 and 2011, and all the hospitalizations that are related to sepsis, which is one of the leading causes of in-hospital mortality.

Using SID data we build a set of multilayer graphs of 231 nodes representing California hospitals, connected with various link types, one for each month of data. The output $y$ of the prediction task is the normalized number of patients diagnosed with sepsis per month and hospital. The unstructured predictor will again be the value of $y$ at the previous month.

We create the layers taking into account that there are many ways in which two hospitals can be connected. For example, two hospitals can be regarded similar if they treat patients with similar characteristics, such as age or ethnicity. Furthermore two hospitals are similar based on their specialization on treating specific diagnosis or applying specific procedures. Such hospital attributes can be typically seen as distributions, as the distribution of patients age for each

hospital, and hence we use the Jensen-Shannon divergence to calculate the similarity between specific attribute distributions for each pair of hospitals. Jensen-Shannon divergence is a symmetrized and smoothed version of the KullbackLeibler divergence $D(P \parallel Q)$ and for a pair of distributions $P$ and $Q$ is defined as $\text{JSD}(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M)$, where $M = \frac{1}{2}(P + Q)$. As $P$ and $Q$ distributions we use the distributions of a variety of attributes provided by the SID, such as the distributions of patient's age and race, distribution of admission source, length of stay, number of patient's chronic diagnosis, numerical codes describing the primary payers and also distribution of percentages of females and patients who died. We also used the distribution of patient's location based on a six-category urban-rural classification and patient's county of residence based on a four category urban-rural designation. Finally we included the specialization similarity of two hospitals, as this is calculated by the frequency distributions of diagnoses treated in each hospital. More precisely we used the Clinical Classifications Software (CCS) codes, provided by SID for each diagnosis, and we calculated the Jensen-Shannon similarity of their distribution for each pair of hospitals. All the similarity matrices are constructed for each month of available data as their values change over time.

For each of the 60 available months of data, we predict the normalized number of patients diagnosed with sepsis at each hospital, $y_t$, using $y_{t-1}$ as unstructured predictor and the similarity matrices constructed from the data of $t-1$. The training of the GCRF models is done using the value of $y_{t-1}$ as target variable, the value of $y_{t-2}$ for unstructured predictor and the similarity matrices at $t-2$. The use of values for $t-2$ reduces the timesteps available for prediction to 58.

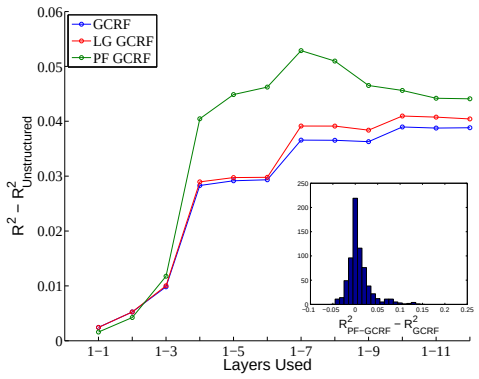


Figure 4: Gain in accuracy over the unstructured predictor, for all three GCRF models, when layers 1 until $\ell$ are utilized for $\ell = 2, ...,12$. The insert shows distribution of the value of $R^2_{PF-GCRF} - R^2_{GCRF}$ for each experiment.

For each timestep, we start our experiment including only one of the layers of the graph. Then we repeat the experiment including more layers adding them one by one. This way we actually create 696 different experiments (58 timesteps $\times$ 12 total layers). We report our results on Figure 4, by plotting the improvement of prediction offered by each of the models over the unstructured predictor. Notice that these results are averaged over the available timesteps of prediction. We can see that PF-GCRF outperforms the rest of the models, while at the same time it offers increasing accuracy as the number of layers included in the model increases. The first two data points of the plot, indicate that GCRF and LG-GCRF both outperform PF-GCRF when less than three layers are included and the reason behind this is the increased complexity of the model. The figure in the insert represents the distribution of the value of $R^2_{PF-GCRF} - R^2_{GCRF}$ for each experiment. We can see that, although in many cases the improvement in accuracy that PF-GCRF offers is small, there is also a large number of cases where the improvement is close or even larger than 10%.

## 7 Conclusion

In this paper we propose PF-GCRF, a Gaussian Conditional Random Fields model that is specifically designed for use with multilayer networks. This model takes into account the correlations among the layers of such networks, and it therefore accumulates the information received from each of the layers, instead of averaging.

We have shown that our model outperforms the traditionally used method using two different real world datasets, a citation and a hospital network. Using the hospital network we showed that PF-GCRF accumulates the information from the graph's layers by showing an increase in the prediction accuracy as the number of layers used increases. We have additionally used artificial data with multiple types of internal structures to further investigate the properties of our model.

We have shown that PF-GCRF provides an accuracy improvement over GCRF that can even be larger than 10%, on real-world high impact datasets, where even a small improvement is highly appreciated. However, including multiple layers into the model intuitively introduces a time complexity issue, leading the user to a trade off between accuracy performance and running time. PF-GCRF can be enhanced by the development of a fast PF-GCRF approach, following the currently active research in the scalability of CRF approaches. In any case PF-GCRF is to the best of our knowledge the only regression model that has been shown to aggregate information from multiple layers of graphs and as such can also be utilized for the evaluation of a layer's

informative power.

Using PF-GCRF we have confirmed the significance of the structural properties of the data and the effect that their incorporation in a predictive model can have on the final accuracy. Thereby we have shown that the educated use of multilayer networks in predictive problems can significantly improve performance.

## References


[1] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[2] M. E. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[3] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007.

[4] P. Holme, F. Liljeros, C. R. Edling, and B. J. Kim, "Network bipartivity," *Physical Review E*, vol. 68, no. 5, p. 056107, 2003.

[5] P. Holme and J. Saramäki, "Temporal networks," *Physics reports*, vol. 519, no. 3, pp. 97–125, 2012.

[6] Y. Yang, N. V. Chawla, Y. Sun, and J. Han, "Predicting links in multi-relational and heterogeneous networks." in *ICDM*, vol. 12, 2012, pp. 755–764.

[7] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *science*, vol. 328, no. 5980, pp. 876–878, 2010.

[8] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of Complex Networks*, vol. 2, no. 3, pp. 203–271, 2014.

[9] W. Tang, Z. Lu, and I. S. Dhillon, "Clustering with multiple graphs," in *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*. IEEE, 2009, pp. 1016–1021.

[10] M. De Domenico, V. Nicosia, A. Arenas, and V. Latora, "Structural reducibility of multilayer networks," *Nature communications*, vol. 6, 2015.

[11] V. Radosavljevic, S. Vucetic, and Z. Obradovic, "Continuous conditional random fields for regression in remote sensing." in *ECAI*, 2010, pp. 809–814.

[12] T. Qin, T.-Y. Liu, X.-D. Zhang, D.-S. Wang, and H. Li, "Global ranking using continuous conditional random fields," in *Advances in neural information processing systems*, 2009, pp. 1281–1288.

[13] P. Kazienko, E. Kukla, K. Musial, T. Kajdanowicz, P. Bródka, and J. Gaworecki, "A generic model for a multidimensional temporal social network," in *e-Technologies and Networks for Development*. Springer, 2011, pp. 1–14.

[14] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[15] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," *Introduction to statistical relational learning*, pp. 93–128, 2006.

[16] V. Radosavljevic, S. Vucetic, and Z. Obradovic, "Neural gaussian conditional random fields," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 614–629.

[17] P. Praveen and H. Fröhlich, "Boosting probabilistic graphical model inference by incorporating prior knowledge from multiple sources," *PloS one*, vol. 8, no. 6, p. e67410, 2013.

[18] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

[19] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hungar. Acad. Sci*, vol. 5, pp. 17–61, 1960.

[20] K. Ristovski, V. Radosavljevic, S. Vucetic and Z. Obradovic "Continuous Conditional Random Fields for Efficient Regression in Large Fully Connected Graphs." *AAAI*, 2013

[21] A. Uversky, D. Ramljak, V. Radosavljević, K. Ristovski, and Z. Obradović, "Panning for gold: using variograms to select useful connections in a temporal multigraph setting," *Social Network Analysis and Mining*, vol. 4, no. 1, pp. 1–13, 2014.

[22] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[23] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 177–187.

[24] S. N. Dorogovtsev and J. F. Mendes, "Evolution of networks," *Advances in physics*, vol. 51, no. 4, pp. 1079–1187, 2002.

[25] M. T. Gastner and M. E. Newman, "Shape and efficiency in spatial distribution networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2006, no. 01, p. P01015, 2006.

[26] G. Bounova and O. de Weck, "Overview of metrics and their correlation patterns for multiple-metric topology analysis on heterogeneous graph ensembles," *Physical Review E*, vol. 85, no. 1, p. 016117, 2012.

[27] J. Manjunatha, K. Sivaramakrishnan, R. K. Pandey, and M. N. Murthy, "Citation prediction using time series approach kdd cup 2003 (task 1)," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 152–153, 2003.

[28] HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2007-2011. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp