Which Links Should I Use? A Variogram-Based Selection of Relationship Measures for Prediction of Node Attributes in Temporal Multigraphs

Alexey Uversky, Dušan Ramljak, Vladan Radosavljević, Kosta Ristovski, and Zoran Obradović*
Center for Data Analytics and Biomedical Informatics,
Department of Computer and Information Sciences,
College of Science and Technology,
Temple University, Philadelphia, PA, USA
*Corresponding author's Email: zoran.obradovic@temple.edu

***Abstract*—When faced with the task of forming predictions for nodes in a social network, it can be quite difficult to decide which of the available connections among nodes should be used for the best results. This problem is further exacerbated when temporal information is available, prompting the question of whether this information should be aggregated or not, and if not, which portions of it should be used. With this challenge in mind, we propose a novel utilization of variograms for selecting potentially useful relationship types, whose merits are then evaluated using a Gaussian Conditional Random Field model for node attribute prediction of temporal social networks with a multigraph structure. Our flexible model allows for measuring many kinds of relationships between nodes in the network that evolve over time, as well as using those relationships to augment the outputs of various unstructured predictors to further improve performance. The experimental results exhibit the effectiveness of using particular relationships to boost performance of unstructured predictors, show that using other relationships could actually impede performance, and also indicate that while variograms alone are not necessarily sufficient to identify a useful relationship, they greatly help in removing obviously useless measures, and can be combined with intuition to identify the optimal relationships.**

Keywords: conditional random fields, temporal social networks, citation networks

Topics: Application of Social Network Mining, Link and Node Prediction in Social Networks, Temporal Analysis on Social Networks Topologies

I. Introduction

As researchers are becoming increasingly aware of the wealth of information that is available in various social networks and its potential uses, more and more of them turn to studying datasets built on top of these networks over more conventional datasets. Research in this topic can tackle an extremely wide array of problems, from evaluating delinquency and alcohol consumption in students based on their social interactions [1] to creating recommender systems for various online marketplaces based on customer interactions with other customers and products [2] to helping ad companies place relevant ads for its customers based on the interactions among said customers and their friends [3]. With such a wide range of potential research areas, however, come a host of challenges related to the underlying tasks: when multiple links exist between users, which ones should be chosen? If temporal information is available, should the data be aggregated to obtain a more generalized result or should each time step be treated separately for a more specific look into the behavior of the data?

The intent of this project is to address several such challenges, the most notable being the utilization of temporal information and determining the efficiency of using various kinds of links. With such challenges in mind, the ultimate goal is to examine how effective a state-of-the-art predictive model, namely Gaussian Conditional Random Fields (GCRF) which was first used for classification in computer vision [4] but now adapted for regression in a social network context, is at the task of predicting node attributes in temporal networks with a multigraph structure. While many significantly simpler models have shown to be quite effective at this task, they fail to utilize much of the information available in the networks they are applied to. The models that did use some of the relationships that could be extracted from the networks relied on intuition when selecting the kinds of relationships they use, and using the idea of variograms and GCRF we hope to establish a more concrete methodology for selecting relationships that are used for various models.

Applying the GCRF model to a bibliographic dataset to achieve the goal of relationship evaluation is a perfect fit, given both the nature of bibliographic networks and the strengths of the GCRF model. Bibliographic networks have a rich heterogeneous structure, including relationships among papers, authors, terms, and venues, which can be extracted from the network even if they are not explicitly defined. Once these relationships are extracted, they can be used within a GCRF framework, which examines both the relationships between input variables and output variables, as well as the relationships between output variables. Here we apply GCRF to a bibliographic network of theoretical high energy particle physics papers (HepTh)[1], treating past citation counts for papers as inputs, future citation counts as outputs, and using several different kinds of relationships (which we refer to as "similarity measures") among the papers as the different link types among nodes in a multigraph setting. We use a GCRF model because it simultaneously assigns weights to

[1] http://www.arxiv.org

both the unstructured predictors and similarity measures that comprise the model, which lets us easily gauge the usefulness of both components. Using this setup we are able to directly evaluate the performance of potentially useful relationships by examining how they affect the output of the GCRF model. We then compare how these findings relate to the perceived utility of each relationship that was first gleaned from its variogram plot.

The rest of the paper is organized as follows: in Section II we discuss some of the work that has already been done in the field of node and link prediction in a bibliographic network setting. Section III provides the formal description of the GCRF model we use, and defines how we use variograms. In Section IV we describe the dataset used in our experiments, as well as explain all of the similarity measures that we used and evaluated using variograms and the GCRF model. Section V provides an analysis of our experimental setup, our usage of variograms, the evaluation measures we use, and the results we obtained. Finally, Section VI consists of a summary of our findings, as well as a brief look into the future work we intend to do with this setup.

II. Related work

Citation prediction is by no means a new topic, and a fair amount of research has already been carried out in this area. One of the primary tasks of the KDD Cup 2003 competition was citation prediction for a bibliography of physics-related papers, and the method that received first place in that competition was able to outperform its competitors despite being extremely simple [5]. The method consisted of identifying papers that had similar patterns in citation histories to the paper in question, and the prediction was formed by averaging the values from these papers, completely disregarding network structure. Castillo et al. [6] focus on predicting citation counts by using primarily author-related information extracted from bibliographic networks, which again did not utilize all of the information that was available. Yan et al. [7] also studied the problem of citation count prediction, but used various features computed from a bibliographic network that were then used with Gaussian processes and a classification and regression tree model to perform the predictions. While this work did utilize much of the available network information, the authors focused on a rather coarse temporal granularity when forming predictions as they predicted citation counts in upcoming years, whereas we focus on monthly predictions.

Various other methods have also incorporated some degree of citation prediction, and were oriented towards link prediction and hence address the more challenging task of predicting who will cite whom rather than simply predicting how many citations a paper will get. Although this goal is different from ours, the methods that are used to attain it revolve around computing features for pairs of papers in order to discern potential citation relationships in the future, rather than focusing on features for individual papers. This is precisely the kind of information that we utilize in our GCRF model, so we mention a few related link prediction papers here. Shibata et al. [8] extract general network-based features for pairs of papers, and then use those features in a support vector machine model to determine if a citation relationship will evolve between a pair of papers. Bethard and Jurafsky [9] introduce a wide variety of paper-paper relationships, including those based on content, authors, and topics. Yu et al. [10] use a meta-path based set of features to compute the relationship strength between papers based on various meta-path counts that occur between the papers, which again utilize different aspects of the bibliographic network.

Finally, our model is an extension of the Conditional Random Field model first introduced in 2001 by Lafferty et al. [11], [12] that was more recently expanded to allow for faster learning and inference [4] and accommodate continuous values [13] and was used for regression of remote sensing data [14]. Our temporal social network model is a further extension to the Conditional Random Field used for regression, and we were able to tailor the bibliographic network data to allow the output to be regarded as a multivariate Gaussian distribution, which in turn allowed us to perform computations in a more computationally feasible way.

III. Model Description

A. Multigraph Networks

Here we briefly define the general structure of a multigraph network, as well as how we utilize this structure within the framework of our model.

Suppose that we are interested in a set of N nodes that we observe over a period of T timesteps. Each node contains a real-valued attribute that changes over time, and we observe that value for each node at each timestep. In addition, a node can be linked to any other node with a variety of connections, which can also change over time. Because there are multiple connections possible between a pair of nodes at the same timestep, the structure of this network can be considered as a multigraph. With this setup, we can formally define the problem we are interested in as the following:

Given a set of nodes N, a history of node attribute values for each node up to timestep t, and a history of connections among nodes up to timestep t, predict the node attribute values for all N nodes at timestep t + 1.

In the context of our experiments, N is the set of papers we focus on, T is a set of monthly snapshots, the node attribute values are the number of citations that each paper has received in that month, and the connections are the different links that exist among papers in the scope of a bibliographic network (which are defined in Section IV). Note that while the primary problem we address is predicting the citation counts for the set of papers at a future timestep, we are also interested in the secondary problem of evaluating the quality of the different types of connections that are available, in order to eliminate useless connections and thus improve both the accuracy and the runtime of our model.

B. Conditional Random Fields

Conditional Random Fields allow for modeling the conditional distribution of an output given an input based on various types of dependencies among outputs. It builds upon the traditional association potential function that is used in regression problems to associate input values with their appropriate output values. This association potential can be established by any number of association functions, which are represented in

our model by two unstructured predictors that use different methods for mapping inputs to outputs.

On top of utilizing various kinds of association functions Conditional Random Fields also utilize relationships among outputs, for which interaction potential functions are used. These functions can be either independent or dependent on the input, and can take on a variety of forms depending on the context in which they are used. For example, when performing citation prediction on a bibliographic network, we can model interaction potential as the shared number of authors for a pair of papers, or the number of similar terms used by a pair of authors.

A graphical representation of the kinds of relationships used by Conditional Random Fields is provided in Figure 1.

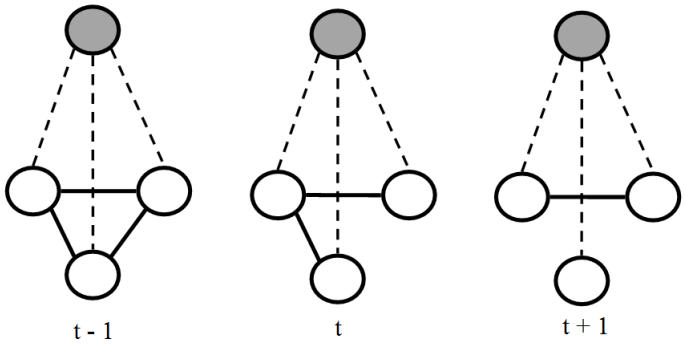

Fig. 1. Graphical representation of the relationships observed by a Conditional Random Field model over 3 timesteps. The gray nodes represent an input variable, the white nodes represent the corresponding output variables, the black links represent the interaction potential, and the dotted links represent the association potential. Note that the strength of the ties among both inputs and outputs varies over time, and the ties can disappear completely if the strength is zero at a particular timestep.

In order to model the conditional distribution of output vectors $y = (y_1 \dots y_N)$ on a set of input vectors $x = (x_1 \dots x_N)$, with association potential function $A(\alpha, y_i, x)$ where α is a K-dimensional set of parameters and interaction potential function $B(\beta, y_i, y_j, x)$ where β is a L-dimensional set of parameters, we represent the distribution as

$$P(y|x) = \frac{1}{Z(x,\alpha,\beta)} exp(\sum_{i=1}^{N} A(\alpha, y, x) + \sum_{j\sim i} I(\beta, y_i, y_j, x)) \quad (1)$$

where $j \sim i$ denotes the connected outputs y_i and y_j (connected with a black line at Figure 1) and where $Z(x, \alpha, \beta)$ is the normalization function defined as

$$Z(x,\alpha,\beta) = \int_y exp(\sum_{i=1}^{N} A(\alpha, y, x) + \sum_{j\sim i} I(\beta, y_i, y_j, x))dy \quad (2)$$

As already noted, A and I could be conveniently defined as linear combinations of a set of fixed features in terms of α and β

$$A(\alpha, y_i, x) = \sum_{k=1}^{K} \alpha_k f_k(y_i, x) \quad (3)$$

$$I(\beta, y_i, y_j, x) = \sum_{l=1}^{L} \beta_l g_l(y_i, y_j, x) \quad (4)$$

The use of features to define the model is convenient because it allows us to include arbitrary properties of observation-output pairs into the compatibility measure. In this way, any potentially relevant feature could be included in the model because parameter estimation automatically determines their actual relevance by feature weighting.

The learning task is to choose values of parameters α and β to maximize the conditional log-likelihood of the set of training examples (we assume that interactions among outputs are defined over the whole training set)

$$L(\alpha, \beta) = logP(y|x) \quad (5)$$

$$(\hat{\alpha}, \hat{\beta}) = \arg\max_{\alpha,\beta}(L(\alpha, \beta)) \quad (6)$$

The inference task is to find the outputs y for a given set of observations x and estimated parameters α and β such that the conditional probability $P(y|x)$ is maximized

$$\hat{y} = \arg\max_{y}(P(y|x)) \quad (7)$$

Conditional Random Fields were initially designed for classification problems, which was a significantly easier task since the normalizing function Z was a sum over a finite set of possibilities rather than an integral. This proves to be much more challenging for regression as Z must be an integrable function, which can be very difficult and computationally expensive to prove due to the complexity of the interaction and association potentials. To address this issue, $P(y|x)$ can be represented as a multivariate Gaussian distribution, which results in Gaussian Conditional Random Fields.

C. Gaussian Conditional Random Fields

The exponent portion E of the CRF model can be rewritten in Gaussian form as:

$$E = -\frac{1}{2}(y-\mu)^T Q(y-\mu) = -\frac{1}{2}y^T Qy + y^T Q\mu + const, \quad (8)$$

where Q and b are canonical parameters of Gaussian distribution defined below. By representing the quadratic terms of y in the association and interaction potentials as $y^T Q_1 y$ and $y^T Q_2 y$ respectively, and combining them, we obtain

$$Q = \Sigma^{-1} = 2(Q_1 + Q_2) \quad (9)$$

$$Q_{1ij} = \begin{cases} \sum_{k=1}^{K} \alpha_k, & i = j \\ 0, & i \neq j \end{cases} \quad (10)$$

$$Q_{2ij} = \begin{cases} \sum_{k=1}^{N} \sum_{l=1}^{L} \beta_l e_{ik}^{(l)} S_{ik}^{(l)}, & i = j \\ -\sum_{l=1}^{L} \beta_l e_{ij}^{(l)} S_{ij}^{(l)}, & i \neq j \end{cases} \quad (11)$$

where S_{ik} represents the similarity between outputs y_i and y_k, and $e_{ik}^{(l)} = 1$ if an edge exists between y_i and y_k under the particular interaction potential l, and $e_{ik}^{(l)} = 0$ otherwise.

To get μ, which is expressed as

$$\mu = Q^{-1}b \tag{12}$$

the linear terms in of the Gaussian form are matched with the linear terms in the exponent of the original form to get $\mu = \Sigma b$, where b is a vector with elements

$$b_i = 2\sum_{k=1}^{K} \alpha_k R_k(x) \tag{13}$$

Finally, using this new representation we see that

$$Z(\alpha, \beta, x) = (2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}} exp(const) \tag{14}$$

and hence

$$P(y|x) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} exp(-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)) \tag{15}$$

with Σ^{-1} and μ defined above. The α and β values used in this model are obtained from the association potential defined by the outputs of two unstructured predictors, and the interaction potential defined by various similarity measures introduced below, respectively.

1) Learning and inference: The learning task is to choose α and β to maximize the conditional log-likelihood, as defined in Equations 5 and 6. To have a feasible model with real valued outputs, Z must be integrable, which is ensured by the constraint that all elements of α and β are greater than 0. In this setting, learning is a constrained optimization problem. To convert it to the unconstrained optimization, we adopt a technique used in [13] that applies the exponential transformation of the parameters to guarantee that they are positive. All parameters are learned by the gradient-based optimization. To apply it, we need to find the gradient of the conditional likelihood:

$$\begin{aligned} \frac{\partial P}{\partial \alpha_k} =& -\frac{1}{2}(y-\mu)^T \frac{\partial Q}{\partial \alpha_k}(y-\mu) + \\ & (\frac{\partial b}{\partial \alpha_k} - \mu^T \frac{\partial Q}{\partial \alpha_k})(y-\mu) + \frac{1}{2} Tr(Q^{-1}\frac{\partial Q}{\partial \alpha_k}) \end{aligned} \tag{16}$$

$$\frac{\partial P}{\partial \beta_k} = -\frac{1}{2}(y+\mu)^T \frac{\partial Q}{\partial \beta_k}(y-\mu) + \frac{1}{2} Tr(Q^{-1}\frac{\partial Q}{\partial \beta_k}) \tag{17}$$

The inference task is to find the outputs y for a given input x, such that the conditional probability $P(y|x)$ is maximized. The GCRF model is Gaussian and, therefore, the maximum *a posteriori* estimate of y is obtained as the expected value μ of the GCRF distribution,

$$y_* = \arg\max_y (P(y|x)) = \mu \tag{18}$$

Uncertainty for each output can be taken as corresponding element from the diagonal of the covariance matrix.

D. Variograms

In order to establish an evaluation of the various similarity measures that we formally define in the next section, we plotted the variograms for each similarity measure to observe how it behaves in relation to the actual data. To obtain these variograms, we first decided on a subset of papers that we would be examining using the GCRF model, as well as the subset of time points that we would use, which ended up giving us 800 papers and 40 time points. We then computed the similarities among all pairs of papers from this set at each time point (for the similarities that change over time), and also computed the squared difference between the citation counts of the two papers at those time points. After collecting all of the similarities and variances over all time points, we binned the results into 10 roughly equal-sized bins, and plotted the similarity values versus the variance of each bin. By including the variance of the dataset on this plot and observing the behavior of the bins we were able to tell whether each similarity measure exhibits an appropriate behavior, namely, displaying a lower variance as the similarity value increases, and staying below the line of variance for the full dataset. These plots also allow us to discard unnecessary portions of the similarity measures, by setting similarity measure values whose corresponding variances are above the variance of the whole data to zero.

IV. Dataset and Feature Description

A. Data

The dataset we used for our experiments was the high energy physics theory bibliographic network which was extracted from arXiv for the 2003 KDD Cup competition. The network consists of 29,955 papers and 352,807 citations spanning over 11 years, and the dataset includes text versions of all papers that can be used to extract additional information about each paper. An XML version of this dataset, from Proximity HEP-Th database, which included most of the metadata available from the full texts of the papers was used to quickly extract the information that was used in our experiments. The Proximity HEP-Th database is based on data from the arXiv archive and the Stanford Linear Accelerator Center SPIRES-HEP database provided for the 2003 KDD Cup competition with additional preparation performed by the Knowledge Discovery Laboratory, University of Massachusetts Amherst. The citation pairs were used to construct a citation history matrix, which represents the number of citations that a particular paper has received at a particular time point.

B. Similarity Measures

The various types of information extracted from the XML file were used to compute a number of similarity features for a pair of papers, which were then used as the interaction potential functions within the GCRF framework to augment the outputs of the unstructured predictors. The following ten similarity measures were used in our experiments:

1) coCiter: Jaccard based similarity measure. Similarity between two papers A and B is expressed as:

$$Sim_{coCiter}(A,B) = \frac{2 \times \#of\ cocitations\ of\ A\ and\ B\ at\ t}{\#of\ cit.\ of\ A\ at\ t + \#of\ cit.\ of B\ at\ t} \tag{19}$$

2) history:

$$Sim_{history}(A,B) = \exp^{-\frac{d(A,B)^2}{k}} \quad (20)$$

where d(A,B) is Euclidean distance between the citation counts of papers A and B over a history of a particular length, and $k = \sum_t^T \sum_n^N \frac{d(A,B)^2}{N \times (N-1) \times T}$ where N and T are number of all papers and all timestamps respectively. We used a history of length 8 here.

3) termTFIDF non temporal: The rest of the similarity measures are all based on the classic TF-IDF term scoring from information retrieval, which were used in [9]:

$$tf(t,d) = |\{t' \in terms(d) : t' = t\} \quad (21)$$

$$idf(d) = 1 + log\frac{|D|}{|d \in D : t \in terms(d)| + 1} \quad (22)$$

$$score_{terms}(q,d) = \sum_{t \in q} tf(t,d)^{0.5} idf(d)^2 \quad (23)$$

This score increases when terms (t) are shared between the query (q) and the document (d), but terms that appear in many documents in the collection (D), such as the, are heavily discounted.

In order to use this score (which is asymmetric) as a similarity measure, we transform it into a symmetric measure in several ways.

$$Sim_{ttnt}(A,B) = \frac{score_{terms}(A,B) + score_{terms}(B,A)}{2} \quad (24)$$

4) termTFIDF non temporal cosine: In this case the same concept is used, but we compute the cosine similarity:

$$Sim_{tttntc}(A,B) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{tAB} w_{i,A} w_{i,B}}{\sqrt{\sum_{i=1}^{tA} w_{i,A}^2} \sqrt{\sum_{i=1}^{tB} w_{i,B}^2}} \quad (25)$$

where *tAB* is the set of terms common to papers A and B, *tA* is the set of terms in paper A, *tB* is the set of terms in paper B, and $w_{i,A} = tf(i,A)^{0.5} idf(A)^2$.

5) author non temporal: For authors the equations remain the same

$$Sim_{ant}(A,B) = \frac{score_{authors}(A,B) + score_{authors}(B,A)}{2} \quad (26)$$

$score_{authors}$ is calculated the same way as $score_{terms}$, except authors are considered as terms.

6) author non temporal cosine: In this case the equations are also the same as for terms, but we consider authors as terms instead:

$$Sim_{antc}(A,B) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{aAB} w_{i,A} w_{i,B}}{\sqrt{\sum_{i=1}^{aA} w_{i,A}^2} \sqrt{\sum_{i=1}^{aB} w_{i,B}^2}} \quad (27)$$

where *aAB* is the set of authors common to papers A and B, *aA* is the set of authors of paper A, *aB* is the set of authors of paper B, $w_{i,A} = tf(i,A)^{0.5} idf(A)^2$, and i is author rather than term.

7) author temporal: In the following similarity measures TFIDF scores of terms/author ids are calculated for the citers of papers that we compare, and averaged over the two papers we compare. The temporal aspect is satisfied because we aggregate all the citers up to the current time point:

$$aggtemp(A,B) = score_{terms}(authors(A), \underset{d \in citing(B)}{\text{concat}}(authors(d))) \quad (28)$$

$$Sim_{at}(A,B) = \frac{aggtemp(A,B) + aggtemp(B,A)}{2} \quad (29)$$

8) term temporal:

$$aggtemp(A,B) = score_{terms}(A, \underset{d \in citing(B)}{\text{concat}}(authors(d))) \quad (30)$$

$$Sim_{tt}(A,B) = \frac{aggtemp(A,B) + aggtemp(B,A)}{2} \quad (31)$$

9) author nontemporal v2: This score is exactly the same as the above author temporal score, but is calculated up to the first time point we consider.

10) term nontemporal v2: This score is exactly the same as the above term temporal score, but is calculated up to the first time point we consider.

C. Unstructured Predictors

To represent the association potential among the inputs and outputs of the GCRF model we used two simple unstructured predictors:

1) k-nearest-neighbor: A sliding window nearest neighbor predictor that forms predictions for a papers citation counts in a future time point by comparing its citation history to that of citation histories of other papers in the dataset, selecting the *k* papers that have the most similar history, and averaging their final corresponding citation counts to predict the count for the paper in question. After testing several configurations we selected a predictor using a window of size 8 and a *k* value of 9.

2) Multiple linear regression: A linear regression predictor whose coefficients were trained on the features of all papers up to the time point we were interested in, and then applying those coefficients on the features at the given time point to form the prediction. For this predictor we used a history of 3 time points as the feature set.

V. Experiments

A. Experimental Setup

In order to avoid the problem of sparsity, we first organized the data so that we were observing only papers that were written before year 2000, and tracking their citation counts starting at year 2000. We then filtered out papers that received less than 25 citations over the resulting 40 month period, leaving us with a matrix of 40 time steps and the citation counts for the 800 most-cited papers at each of those time points. Note that although we focused on the 40 time steps that took place after the year 2000 for training the GCRF model, we still had the citation counts for the previous time steps which we used when training the unstructured predictors.

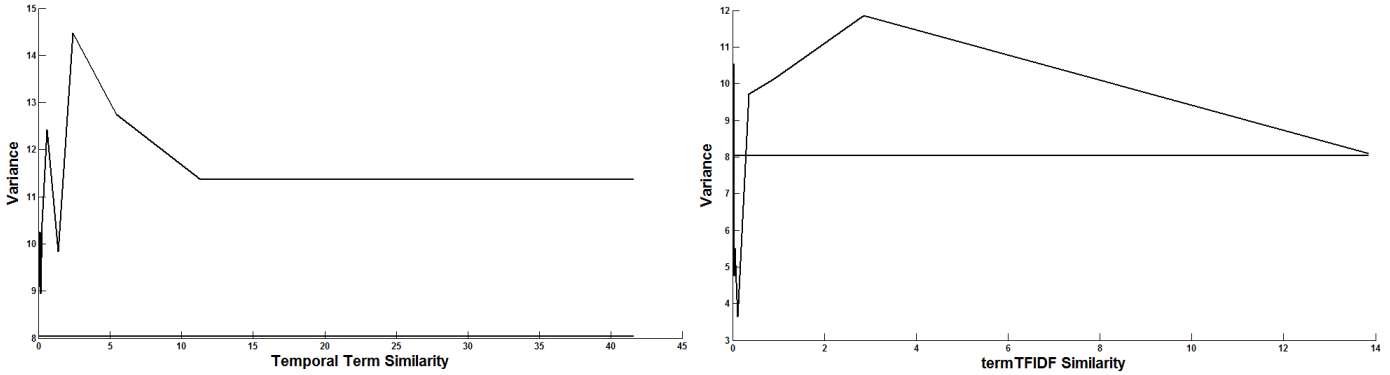

Fig. 2. Examples of variograms of bad similarity measures. The horizontal lines represent the overall variance of the dataset, indicating that regardless of the similarity strength, neither of these similarities are useful for identifying a relationship between the true values.

B. Variograms of Similarity Measures

After selecting the paper set that we would be focusing on, we extracted the appropriate information from the XML file to obtain the similarity measures for each pair of papers. We then examined each similarity measure via a variogram, and selected the most promising ones, of which there were nine, to be used within the GCRF model. Some of the variograms we examined are reproduced in Figures 2 and 3. Note here that we do not show all of the variograms we examined as all of the bad variograms had behaviors similar to the two bad ones presented in Figure 2, and the good ones looked very similar to the two good ones presented in Figure 3.

In total we examined over 40 variograms from the various combinations of the similarity measures defined above, and we discarded any individual or combined similarity whose variogram did not exhibit the behavior of the two "good" variograms shown in Figure 3, leaving us with nine potentially useful measures. Using these we then explored the overall performance of the model when training and testing on different intervals. Although the differences were fairly minor, we opted to use the intervals that had the highest overall performance, which consisted of training the GCRF model on time points 10-20, and testing on time points 21-30. In order to avoid unfair bias for the unstructured predictors that we used in our model, we used time points 1-9 to train them and obtain predictions for the 10-30 interval, thus ensuring that we weren't seeing additional data when forming those predictions. In the case of kNN, we treated the interval up to time point 9 as the only interval from which we could observe citation counts, and made predictions in the 10-30 interval using those values. In the case of Multiple Linear Regression we again only used features up to the 9th time point to learn the coefficients of the model, and used them to predict in the 10-30 interval.

Finally we conducted a series of experiments to explore the performance of the GCRF model using different similarity measures both independently and in combination, as well as comparing those results to those obtained from using GCRF with only one of the two unstructured predictors to help determine how much the similarities were actually helping.

C. Evaluation Measures

To gauge the performance of the GCRF model as well as the unstructured predictors by themselves, we consider two traditional evaluation measures for regression tasks:

1) R^2 coefficient of determination: a goodness-of-fit measure that displays how closely the output of the model matches the actual value of the data. A score of 0 indicates a very poor matching, while a score of 1 indicates a perfect match.

$$R^2 = 1 - \frac{\sum_i (y_i - f(x_i))^2}{\sum_i (y_i - y_{average})^2} \tag{32}$$

where $f(x_i)$ is the predicted value, y_i is the true value, and $y_{average}$ is the average of y values.

2) Root mean squared error: an overall measure of the difference between the predicted values and the actual values. Though it is sometimes difficult to interpret when the values being measured cover a wide range, in our case it is an appropriate evaluation metric as citation counts fall within a fairly narrow range.

$$RMSE = \sqrt{\frac{1}{n} \sum_n \left(f(x_i) - y_i\right)^2} \tag{33}$$

where $f(x_i)$ is the predicted value, y_i is the true value, and n is the number of samples.

D. Results

In order to accurately gauge the effectiveness of the different similarity measures we performed a number of experiments using the parameters defined in the previous sections this includes focusing on a set of 800 papers over the span of 40 months and using months 1-20 to train the model and months 21-30 to test it. Using these parameters we measured the performance of the unstructured predictors by themselves, the performance of the GCRF model using two of the bad similarity measures, using each of the good similarity measures individually, as well as a combination of the most promising of these similarities and a combination of the least promising of these similarities. We also observe the performance of GCRF using these similarity measures and only one of the

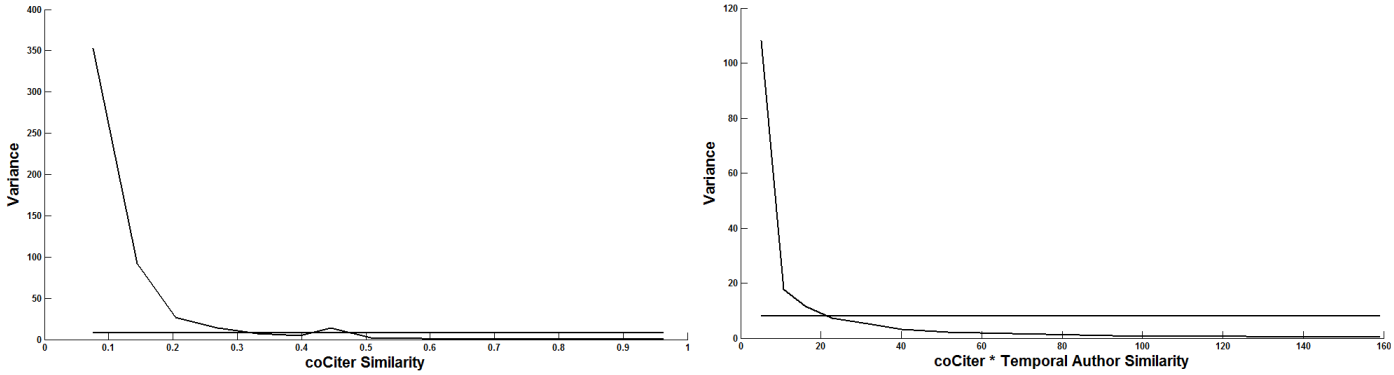

Fig. 3. Examples of variograms of good similarity measures. Note that after a certain similarity value, the respective variance drops below the overall variance of the data, indicating that higher similarity values help identify the relationship between outputs. These variograms also help identify the cutoff point for each similarity measure: for example, by setting any value of coCiter * Temporal Author Similarity below 20 to zero, we ensure that the similarity only identifies the meaningful relationships among the data points.

TABLE I. PERFORMANCE OF INDIVIDUAL UNSTRUCTURED PREDICTORS, AS WELL AS GCRFS THAT USED TWO BAD SIMILARITY MEASURES. NOTE THAT THERE WERE SIGNIFICANTLY MORE BAD SIMILARITY MEASURES, BUT AS THE RESULTS FROM USING THEM WITH GCRF WERE IDENTICAL TO THE RESULTS SHOWN HERE WE OMIT THEM.

Model	Average R^2
MLR	0.67
kNN	0.58
GCRF termTemporal	0.67
GCRF author	0.67

two unstructured predictors to better define the impact of the similarity measures on the individual predictors. The results are organized as follows: in Table I, the first two rows are the performance results of the individual unstructured predictors; the next two rows are the performance results of our GCRF model using both unstructured predictors and one bad similarity measure each. The remaining results in Tables II and III are trios of GCRF results using the specified similarity measure(s) and both unstructured predictors, just MLR, and just kNN, respectively. Although we computed the RMSE values and variances of all metrics for each set of experiments, the RMSE values displayed exactly the same results the R^2 values did, and the variances were small for all experiments, so we omit them from the results.

Analyzing these results yields several conclusions, all of which are consistent with either our variogram evaluations or the nature of the GCRF model. Firstly, by looking at Table I we can see that when GCRF uses both unstructured predictors, its performance is at least as good as the better of the two individual predictors, even when using a bad similarity measure. This behavior is to be expected as the model is able to form predictions using the information extracted from both predictors, and is able to offset the downfalls of one with the advantages of the other. We can also see that applying the similarity measures that were deemed bad as a result of their variograms offers no improvement in performance: in both cases GCRF performs identically to MLR, suggesting that those similarities did in fact offer no beneficial information to the model.

Observing the next set of results shown in Table II further

TABLE II. PERFORMANCE OF USING A SINGLE GOOD SIMILARITY MEASURE IN THE GCRF MODEL. HERE EACH COLUMN REPRESENTS USING A DIFFERENT UNSTRUCTURED PREDICTOR SETUP WITH THE GIVEN SIMILARITY MEASURE: BOTH MEANS GCRF IS TRAINED USING BOTH MLR AND KNN AND THE SIMILARITY MEASURE, MLR MEANS GCRF IS TRAINED USING ONLY MLR AS THE UNSTRUCTURED PREDICTOR, AND KNN MEANS ONLY KNN WAS USED AS THE UNSTRUCTURED PREDICTOR.

Similarity Measure	Average R^2 both	Average R^2 MLR	Average R^2 kNN
coCiter	**0.71**	**0.70**	**0.64**
authorCoCiter	0.69	0.68	0.59
authorCosineCoCiter	0.69	0.68	0.59
authorTemporalCoCiter	0.68	0.67	0.58
history	0.68	0.67	0.58
termTemporalHistory	0.68	0.67	0.58
authorCosineHistory	0.70	0.69	0.61
authorNewHistory	0.68	0.67	0.58
termNewHistory	0.68	0.67	0.58

showcases the benefits of having two unstructured predictors over one, but also shows the improvements offered by some of the good similarity measures. For example, we can see that the performance of GCRF using the coCiter similarity and only the kNN unstructured predictor is noticeably better than kNN by itself, suggesting that the coCiter similarity measure allows for more accurate predictions. On the other hand, several of the good similarity measures also offer little to no improvements in the performance of the model. This can be explained by one of two ideas: the similarities that are based on coCiter similarity such as authorTemporalCoCiter appeared to be good due to the effectiveness of the coCiter similarity alone, but their other components did not actually contribute anything positive to the model. The history-based similarities also displayed promising variograms, but when implemented within the GCRF did not really improve the performance. This can be explained by the fact that the unstructured predictors heavily rely on citation history, so using a similarity based on the same concept did not add any new information to the model. This behavior was not evident when computing variograms, since said variograms did not incorporate the citation histories as GCRF did (and hence the variograms of history-based similarities looked good but didn't actually help).

TABLE III. PERFORMANCE OF GCRF USING A COMBINATION OF THE BEST INDIVIDUAL SIMILARITY MEASURES, AND USING A COMBINATION OF THE WORST INDIVIDUAL SIMILARITY MEASURES.

Similarity Combinations	Average R^2 both	Average R^2 MLR	Average R^2 kNN
coCiter, authorCoCiter, authorCosineCoCiter, authorCosineHistory	0.70	0.70	0.63
authorTemporalCoCiter, termTemporalHistory, history authorNewHistory, termNewHistory	0.68	0.67	0.58

The final set of observations, shown in Table III, exhibit the results of using several similarity measures at once. When grouping 4 of the best-performing similarities together, the results of which are shown in the first row of Table III, we saw that the results were actually worse than the performance of the best similarity by itself (coCiter), and in fact took much longer to obtain. This is again consistent with the nature of the model, as two of the similarities were highly correlated with coCiter similarity, and the history-based similarity suffered from the problem discussed above. So when combining a single good measure with three others that offered little to no new information, we were needlessly complicating the model and hurting the performance of the best similarity. The final combination of the least-promising similarities further shows that good variograms by themselves are not sufficient to determine if the measure is actually useful, as all of the variograms in this combination looked promising. As a result, we can conclude that the nature of the model must also be taken into account when determining what similarity measures would offer the best performance.

VI. CONCLUSION AND FUTURE WORK

In this paper we examined the issue of selecting effective relationships from a large pool of potential connections among nodes in a social network. These relationships were first examined by computing their respective variograms and observing how the relationship strength relates to the actual values of the data. We then tested the benefits of these relationships when they were implemented in our Gaussian Conditional Random Field model that was adapted for node attribute prediction in temporal social networks with a multigraph structure. These experiments showed that while variograms offer a great amount of insight into the effectiveness of a relationship, they must be combined with insights based on the types of unstructured predictors that are used in the model to find the truly optimal relationships.

In future experiments we hope to further explore this behavior by constructing additional unstructured predictors that are less reliant on a common source of information (as our predictors were on citation history), as well as exploring additional similarity types, such as the meta-path-count-based similarities that were used in [10]. We will also apply our approach to additional datasets to see if the results are consistent in other networks.

This work is supported in part by DARPA Grant FA9550-12-1-0406 negotiated by AFOSR.

REFERENCES

[1] V. Ouzienko, Y. Guo, and Z. Obradovic, "A decoupled exponential random graph model for prediction of structure and attributes in temporal social networks," *Stat. Anal. Data Min.*, vol. 4, no. 5, pp. 470–486, Oct. 2011.

[2] F. E. Walter, S. Battiston, and F. Schweitzer, "A model of a trust-based recommendation system on a social network," *Autonomous Agents and Multi-Agent Systems*, vol. 16, no. 1, pp. 57–74, Feb. 2008.

[3] W.-S. Yang, J.-B. Dia, H.-C. Cheng, and H.-T. Lin, "Mining social networks for targeted advertising," in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences - Volume 06*, ser. HICSS '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 137.1–.

[4] C. Liu, E. H. Adelson, and W. T. Freeman, "Learning gaussian conditional random fields for low-level vision," in *In Proc. of CVPR*, 2007, p. 7.

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

[6] C. Castillo, D. Donato, and A. Gionis, "Estimating number of citations using author reputation," in *Proceedings of the 14th international conference on String processing and information retrieval*, ser. SPIRE'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 107–117.

[7] R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li, "To better stand on the shoulder of giants," in *JCDL*, 2012, pp. 51–60.

[8] N. Shibata, Y. Kajikawa, and I. Sakata, "Link prediction in citation networks," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 1, pp. 78–85, Jan. 2012.

[9] S. Bethard and D. Jurafsky, "Who should i cite: learning literature search models from citation behavior," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 609–618.

[10] X. Yu, Q. Gu, M. Zhou, and J. Han, "Citation prediction in heterogeneous bibliographic networks," in *SDM'12*, 2012, pp. 1119–1130.

[11] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[12] C. Sutton and A. Mccallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, Eds., 2007.

[13] T. Qin, T. Y. Liu, X. D. Zhang, D. S. Wang, and H. Li, "Global Ranking Using Continuous Conditional Random Fields," in *NIPS*, D. Koller, D. Schuurmans, Y. Bengio, L. Bottou, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. MIT Press, 2008, pp. 1281–1288.

[14] V. Radosavljevic, S. Vucetic, and Z. Obradovic, "Continuous conditional random fields for regression in remote sensing," in *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2010, pp. 809–814.