
AFRL-AFOSR-VA-TR-2019-0098

# Prospective Analysis of Large and Complex Partially Observed Temporal Social Networks

**Zoran Obradovic**
**TEMPLE UNIVERSITY-OF THE COMMONWEALTH SYSTEM OF HIGHER EDUCA**

**08/14/2018**
**Final Report**



Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From - To)* |
|---|---|---|
| 14-04-2019 | Final Performance | 01 Aug 2012 to 31 Jul 2017 |

**4. TITLE AND SUBTITLE**
Prospective Analysis of Large and Complex Partially Observed Temporal Social Networks

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**
FA9550-12-1-0406

**5c. PROGRAM ELEMENT NUMBER**
61102F

**6. AUTHOR(S)**
Zoran Obradovic

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
TEMPLE UNIVERSITY-OF THE COMMONWEALTH SYSTEM OF HIGHER EDUCA
1801 N BROAD ST
PHILADELPHIA, PA 191226003 US

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
AF Office of Scientific Research
875 N. Randolph St. Room 3112
Arlington, VA 22203

**10. SPONSOR/MONITOR'S ACRONYM(S)**
AFRL/AFOSR RTA2

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**
AFRL-AFOSR-VA-TR-2019-0098

**12. DISTRIBUTION/AVAILABILITY STATEMENT**
A DISTRIBUTION UNLIMITED: PB Public Release

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**
The objective of our study was to explore in detail the hypothesis that a unified approach is feasible for modeling large-scale temporal social networks that:
provide predictions and uncertainty estimates for the network structure and the properties of nodes,
capture simultaneously positive and negative influences among nodes,
consider simultaneously node properties of multiple types and of different temporal dynamics,
address multiple kinds of interactions,
allow learning in partially observed temporal graphs including predicting properties for emerging nodes.
The objective of the study is completely achieved, and the results are published in 52 articles listed at the end of the report.

**15. SUBJECT TERMS**
graphs, networks, social

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | NGUYEN, TRISTAN |
| Unclassified | Unclassified | Unclassified | UU | | 19b. TELEPHONE NUMBER *(Include area code)* 703-696-7796 |

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

*Prospective Analysis of Large and Complex Partially Observed Temporal Social Networks*
Final Report for DARPA GRAPHS FA9550-12-1-0406

**PI: Zoran Obradovic** (Temple University)

## Project aims and tasks overview

Our DARPA GRAPHS project developed and validated effective predictive modeling technology to achieve the following aims and tasks:

Aim 1: GCRFs for Modeling Complex Evolving Social Networks
- Task 1: Simultaneously modeling node attributes, link properties, and multi-modal graph structure
- Task 2: Learning attribute importance and positive and negative influence/correlation
- Task 3: Inference and uncertainty analysis of GCRF
- Task 4: Prediction of node states and graph links
- Task 5: Modeling dynamical attributes and dynamical influence/correlation
- Task 6: Modeling temporal networks based on partially observed data
- Task 7: Uncertainty propagation in GCRF
- Task 8: GCRF for directed graphs
- Task 9: Continuous Conditional Dependency Network for structured regression

Aim 2: Convex Optimization for Learning Large GCRFs
- Task 10: Basic formulation and solution approaches
- Task 11: Sparsity-inducing regularization
- Task 12: Fast sparse Gaussian Markov Random Fields learning based on Cholesky factorization
- Task 13: Structured regression on multi-scale networks

Aim 3: Non-Stationary and Time Evolving Correlation Analysis in Social Networks
- Task 14: Regime-switching models
- Task 15: Graph-constrained stationary covariance processes
- Task 16: Ensemble-based structured regression

The results of our research are published in 52 articles listed at the end of the report. Here we summarize the main achievements.

## Aim 1: GCRFs for Modeling Complex Evolving Networks

Modeling complex phenomena through instances that are highly structured and interdependent is a challenging task which is different from traditional machine learning approaches. Many applications have such properties and they are usually modeled as graphical structure models. Structured models for regression on evolving networks are less developed than classification problems, leaving numerous applications even more difficult to solve. In our DARPA GRAPHS project we developed representationally powerful and computationally efficient methods to facilitate modeling complex evolving networks. The new methods were applied to:

- **(a)** Prediction of long-term precipitation across the US at the resolution on the levels of individual stations.
- **(b)** Citation prediction for scientific papers and patents, observed over time.
- **(c)** Characterizing player engagement in an online multiplayer game, based on behavior and social interactions.
- **(d)** Disease co-occurrence modeling in a network of hospitals across US over time.
- **(e)** Prediction of attributes in evolving social networks defined over friendships.

Our approach to modeling evolving attributed networks is based on a **Gaussian Conditional Random Field (GCRF)** model developed at Prof. Z. Obradovic's lab at Temple University. GCRF is a probabilistic exponential model that captures both the network structure of variables of interest (**y**) and attribute values of the nodes (**x**). It is a model over a general graph structure (not only chains or trees), and can represent the structure as both the function of time, space or any other user defined structure. It models the structured regression problem as estimation of a joint continuous distribution over all nodes

$(\mathbf{y}), P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \exp(\sum_{i=1}^{N} A(\boldsymbol{\alpha}, y_i, \mathbf{x}) + \sum_{j \sim i} I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x}))$, where the dependence on input measurements is modeled by the "association potential" $A(\boldsymbol{\alpha}, y_i, \mathbf{x}) = \sum_{k=1}^{K} \alpha_k f_k(y_i, \mathbf{x})$, and the structure between outputs is modeled by the "interaction potential": $I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x}) = \sum_{l=1}^{L} \beta_l g_l(y_i, y_j, \mathbf{x})$. The association potential is used to incorporate domain unstructured models ($R_k$): $f_k(y_i, \mathbf{x}) = -(y_i - R_k(\mathbf{x}))^2, k = 1,...K$, and the interaction potential represents multi-modal graph structure that includes multiple layers of node inter-dependence: $g_l(y_i, y_j, \mathbf{x}) = -S_{ij}^{(l)}(\mathbf{x})(y_i - y_j)^2$. For such choice of feature functions, the distribution can be expressed in the Gaussian form: $P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{(2\pi)^{N/2} |\boldsymbol{\Sigma}|^{1/2}} \exp(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}))$, which makes the inference and learning of the model more feasible. The inference problem then finds the mode of the Gaussian distribution: $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}}(P(\mathbf{y} \mid \mathbf{x}))$ and the learning of the parameters (α, β) is done by convex optimization of the log likelihood: $\log P = -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) - \frac{1}{2} \log|\boldsymbol{\Sigma}|$

In our project numerous challenges with applying the GCRF model on large-scale complex real-world problems were addressed and solutions were developed for in the tasks related to Aim 1.

*Large-scale learning and inference of GCRF models*

To extend the GCRF models to handle very large datasets, we developed approaches based on partitioning a large network followed by distributed GCRF modeling, with an alternative of approximating GCRF on a single computer by a faster method (Figure 1). The results of the two approaches are summarized in this section.

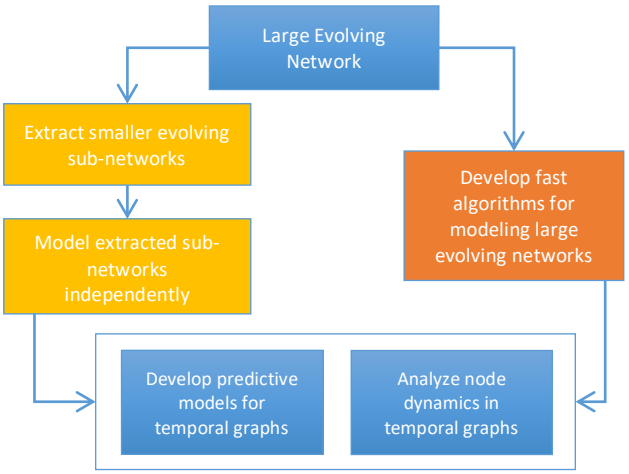


Figure 1. Two approaches for enabling the GCRF model for large-scale applications

***Discovery of evolving communities in big weighted networks***: We have developed a local minimization approach for detection of community structure in evolving weighted networks, which scales well to large-scale problems. Our EGC (Evolving Graph Crawling) method directly optimizes the edge-cut (total weight of edges coming out of the communities) by local search and is able to shrink as well as expand a temporally observed community. In this process instead of relying on a single node, the entire community structure found at previous times is used to seed search for the next temporal adjustment. The algorithm computational time depends on the community size rather that the number of nodes in the graph. It is applicable to both undirected and directed graphs. The results obtained on synthetic and real networks with millions of nodes provide evidence that the proposed method is faster and more accurate that the state-of-the-art alternatives when applied to identification of communities in large weighted evolving networks (Figure 2). Also, the algorithm is applicable to large static networks, where it is comparable in accuracy to the alternatives.

***Learning GCRFs on discovered communities***: After partitioning graphs into communities ("GGP"), we showed that training GCRF models to the obtained partitions independently and in parallel could be performed several orders of magnitudes faster than training GCRF on the entire graph ("No partitioning") (Figure 3), without significant loss in accuracy.

Experiments conducted on synthetic and two real-world evolving networks with up to 750 000 nodes resulted in improved accuracy, since community-specific models were well specialized for each sub-graph. Another benefit of modeling each community independently is easy parallelization, which we utilized in a software implementation that uses the MPI parallelization platform.

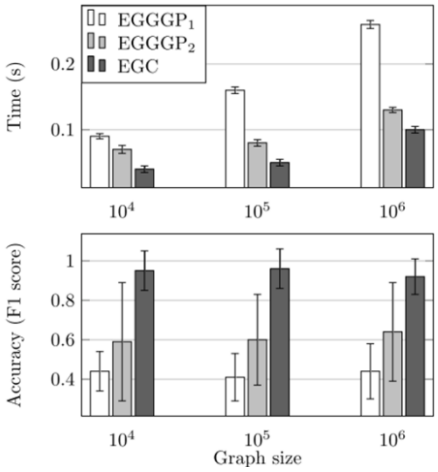


Figure 2. Speed and accuracy of finding communities for our EGC method, comparing to two state-of-the-art method variations

***Developing fast approximations of GCRFs for efficient learning and inference***: In the second approach for extending the GCRF model to large-scale applications, we developed a fast algorithm for approximate GCRF inference and learning, since the original GCRF takes $O(N^3)$ time and $O(N^2)$ space for training on densely connected graphs. The first step in addressing this problem is to substitute the joint distribution $P(\mathbf{y}|X)$ with the distribution $Q(\mathbf{y}|X)$ that can be expressed as a product of independent marginals $Q(\mathbf{y}|X) = \prod_{i=1}^{N} Q_i(y_i|X)$. Under mean field theory, the best approximation of P is a distribution Q which minimizes Kullback-Leiber (KL) divergence, which can also be represented as a Gaussian distribution, whose mean and variance are:

$$\mu_i = \frac{\sum_{k=1}^{K} \alpha_k R_k(X) + 2\sum_{l=1}^{L} \beta_l \sum_{j \neq i} k_l(\mathbf{p}_i^{(l)}, \mathbf{p}_j^{(l)})\mu_j}{\sum_{k=1}^{K} \alpha_k + 2\sum_{l=1}^{L} \beta_l \sum_{j \neq i} k_l(\mathbf{p}_i^{(l)}, \mathbf{p}_j^{(l)})}$$

$$\sigma_i^2 = \frac{1}{2(\sum_{k=1}^{K} \alpha_k + 2\sum_{l=1}^{L} \beta_l \sum_{j \neq i} k_l(\mathbf{p}_i^{(l)}, \mathbf{p}_j^{(l)}))}$$

In other words, the predictions ($\mu$) and their uncertainties ($\sigma^2$) can be solved directly and iteratively for a fixed number of iterations $I << N$, which reduces the computation complexity to $O(I\,N^2)$. Since the most intensive part in this calculation is the summation over kernel ($k_l$) matrix, by using the Gaussian kernel function, we can express this computation as a convolution of $\mu$ vector with the Gaussian kernel.

$$(G \otimes S)(\mathbf{p}_i^{(l)}) = \sum_j k_l(\mathbf{p}_i^{(l)}, \mathbf{p}_j^{(l)})\mathbf{s}_i$$

Based on this idea we applied to our problem a recently developed signal processing technique for approximate kernel convolution using mapping to permutohedral lattice. This reduces the computational complexity to O(N) for densely connected graphs, with memory requirements O(N). This approximation that we call FF-GCRF results in a huge computational speed up (Figure 4), and it allows modeling dense graphs with millions of nodes, on a single machine, with very little loss in accuracy, as shown in Table 1 for a fully connected temporal graph of N nodes corresponding to the large-scale high impact remote-sensing application of predicting aerosol concentrations over the entire continental USA.

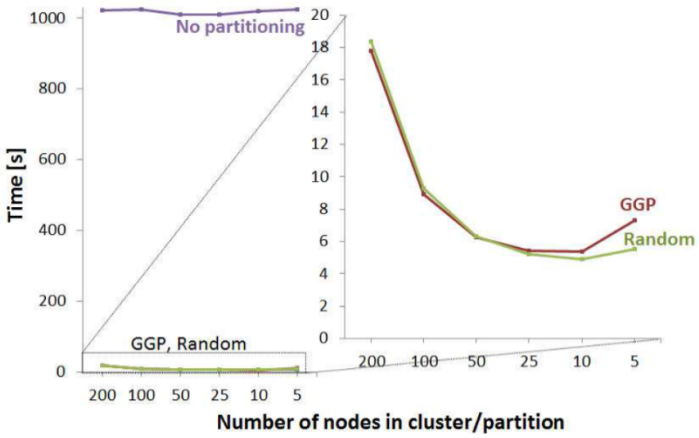


Figure 3. Speed of training GCRF on communities (GGP), random communities, and no communities (full network)

| Method | Sample size | |
|---|---|---|
| | N=3000 | N=20000 |
| **FF-GCRF** | **0.117±0.001** | **0.115±0.005** |
| **GCRF** | **0.116±0.001** | N/A |
| **C005** | 0.126±0.010 | 0.129±0.010 |

Prediction error (RMSE)

| Method | Sample size | | | |
|---|---|---|---|---|
| | N=3000 | | N=20000 | |
| | Learning | Inference | Learning | Inference |
| **FF-GCRF** | **2s** | **0.009s** | **250s** | **2s** |
| **GCRF** | **87s** | **0.630s** | **N/A** | **N/A** |

Execution time (seconds)

Table 1. Accuracy and speed of the FF-GCRF on the aerosol prediction problem, compared to GCRF and the knowledge-based C005 method

We have also demonstrated the benefit of this approach on denoising images represented as fully connected graphs, where FF-GCRF was faster and more accurate than some of the state-of-the-art image processing algorithms.

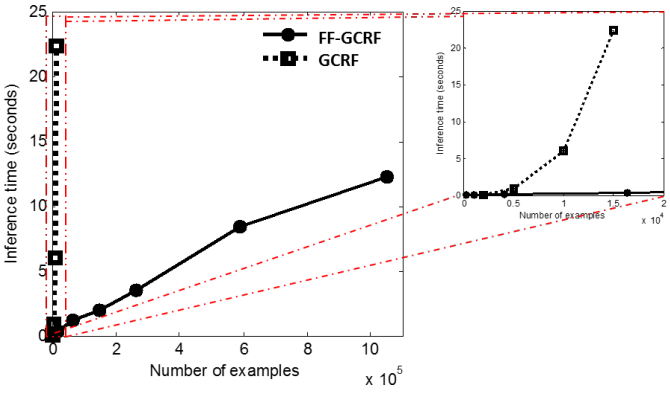


Figure 4. Inference speed of FF-GCRF compared to the original GCRF. Part of the figure is showed zoomed.

***Extension of the fast GCRF approximation method for general graphs***: The limitation of our FF-GCRF method is that it requires the use of a Gaussian kernel, which limits applications to problems that can represent the structure as features in Euclidean space. To enable the use of this method on general networks, we have utilized Landmark Multi-Dimensional Scaling to map an arbitrary graph structure to a Euclidean space, where Gaussian kernel can be used. This method also has linear time complexity with respect to the number of nodes in the graph. The two-stage approach is currently validated by an application to a big evolving network containing millions of nodes describing a massive online multi-player game, where our objective was to model player involvement time in a social network setting, and use features that describe the user behavior.

## *Learning partially observed graphs*

We explored two ways of handling missing data in the structured datasets. Our approach based on imputation of the missing values of the connected nodes, and an alternative based on extending the GCRF model to naturally use missing (or unlabeled) data are summarized in this section.

### *Data imputation in evolving networks*

We developed a method that can simultaneously fill in missing links and missing attribute values of graph nodes in an evolving network. Our data imputation method is extending our temporal exponential random graph model, which we call ITERGM, and it uses an EM algorithm over two Markov Chain Monte Carlo inferences, for modeling links and attributes temporally. For sampling, our data imputation uses our etERGM model that was developed for modeling evolving graphs. In our approach the attribute prediction is modeled as: $P(x^t|x^{t-1},N^t,\gamma)=\frac{1}{Z(x^{t-1},N,\gamma)}\exp\{\gamma'\psi(x^t,x^{t-1},N^t)\}\mathbb{N}(x^t|V_0,\Sigma_0)$, where we use temporal feature functions that capture the expected change between time steps, and we also regularize the values with normal priors. The link prediction step is modeled as: $P(N^t|N^{t-1},\boldsymbol{x}^t,\boldsymbol{\theta})=\frac{1}{Z(x^{t-1},\boldsymbol{x}^t,\boldsymbol{\theta})}\exp\{\boldsymbol{\theta}'\psi(N^t,N^{t-1},\boldsymbol{x}^t)\}$. Two models are optimized iteratively, until convergence. We compared our model to available alternative methods on several well-known real-world applications, including predicting delinquency (Figure 5) and alcohol consumption of teenagers. The obtained results provide evidence that the algorithm has good performance even when a large fraction (up to 60%) of data points are missing. It is also stable with little variance after 10,000 bootstrapping experiments, and has a linear scalability in a number of time steps, and quadratic in number of nodes in the graph.

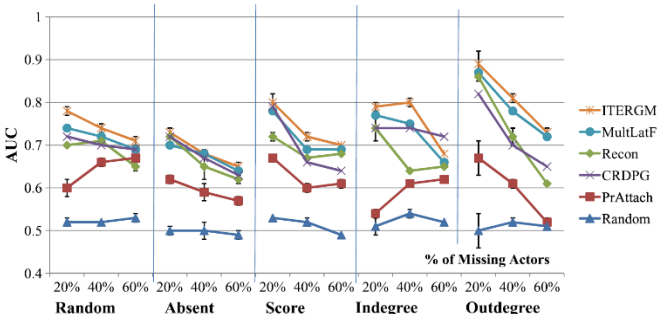


Figure 5. Accuracy of recovering missing values by imputation using different methods, for different amounts of missing data, and for various missingness mechanisms.

### *Extension of the GCRF model for handling missing labels*

Since the GCRF model originally assumes all data for training is given, we extended the training algorithm to make use of the instances with missing data, which is especially beneficial in structured models, where ignoring such instances can damage the structure in the data. Our approach consists in marginalizing the overall joint distribution over the data with missing values in the training set. In general settings, this could pose a bigger problem, especially because in the regression problem we need to integrate instead of add over these data points. However, since our model is multivariate Gaussian, we can effectively reduce this operation to matrix operations, which preserve the distribution as Gaussian. Our experiments provide evidence that there is a significant benefit to this approach, even when large percent of labels is missing (regression results for up to 80% data missing completely at random in a graph of 1600 nodes observed over 5 time steps are shown at Figure 6 where benefits of marginalization based m-GCRF model are evident).

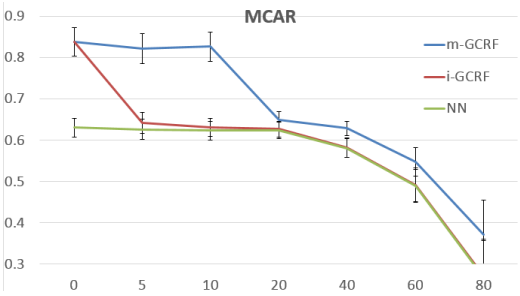


Figure 6. Accuracy ($R^2$) for GCRF model that integrates over missing labels (m-GCRF), ignores missing labels (i-GCRF) and the baseline unstructured method (Neural Networks). Accuracy is show for various percentages of missing data.

*Identifying useful similarity measures as a graph structure*

In many situations we can choose among several ways to represent graphs, based on how we code the graph from input data (e.g. spatial proximity could be calculated using different distance functions). We developed a method for pre-emptively evaluating any given similarity measure, before using it inside the GCRF model. Following the intuition that the similarity is "good" if higher similarity values result in closer node state values, we used the variograms of the similarities to inspect the candidate similarities. We show that carefully selecting appropriate similarities this way not only saves effort (by discarding irrelevant similarities), but also can increase the accuracy of the original model (Table 2). We evaluated this approach on several co-citation temporal networks of up to 350,000 nodes, where we show which similarity functions (out of 40 considered) should be used for graph construction in that domain.

| MLR | KNN | GCRF + author nonTemporal (bad) | GCRF + author*coCiter (good) | GCRF + coCiter (good) |
|---|---|---|---|---|
| 0.671 ± 0.004 | 0.582 ± 0.004 | 0.673 ± 0.004 | 0.691 ± 0.004 | 0.715 ± 0.002 |

Table 2. Accuracy ($R^2$) for GCRF model that uses "good" VS "bad" similarity measure, or no similarities at all (MLR and KNN)

*Prediction of weighted links of a temporal graph*

Predicting weighted links in a network is also a challenging task, especially for temporal networks. Our goal was to predict link values in the current time step, given the history of the graph in previous T steps. We built on our etERGM model, and extended it to handle evolving weighted graphs. The model has the following form: $P(W^t|W^{t-1},\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{W^{t-1}},\boldsymbol{\theta})} exp\{\boldsymbol{\theta}'\psi(W^t,W^{t-1})\}\mathbb{N}(W^t|W^{1:T})$, where we use sufficient statistics ψ that capture temporal aspects such as stability and variance, and we use history of the graph in previous timesteps 1:T as a prior in a distribution. We applied our method on the problem of disease co-occurrence, using the admission data at over 4000 hospitals from 2007-2009, and evaluated it with other methods (Table 3).

| Stratum | Average | Last Step | Adamic Adar | Common Neighbor | WETRGM |
|---|---|---|---|---|---|
| #1 | 8.92 | 7.59 | 505.19 | 3811.25 | **7.18** |
| #2 | 6.07 | 5.80 | 427.51 | 3075.64 | **5.28** |
| #3 | 10.70 | 12.36 | 186.37 | 1399.33 | **10.13** |
| #4 | **11.44** | 13.54 | 234.52 | 1764.92 | 12.18 |
| #5 | 9.55 | 11.10 | 204.34 | 1519.11 | **9.08** |

Table 3. Accuracy (MAE) of various models for predicting disease occurrence in the 12th month in each stratum, which is defined based on region, hospital size, ownership and hospital type

*Node degree prediction for evolving social networks*

Node degree in graphs can be very useful information, showing the importance of nodes in a (social) network that is evolving. Our problem was to predict the degree of nodes in the future state, given the snapshots of the social network at previous time steps. To tackle this problem, we built a number of features that utilize the network topology and describe the local structure of the graph at each node. These features are then used to model the influence of a node's neighborhood on its future degree. Beside the concept of regular equivalence of nodes that is based on their attributes, we also considered the structural equivalence of nodes that is determined by the shared neighborhood of a pair of nodes. Using these features, we built a GCRF model for predicting future node degree. Evaluating this approach on the sample data from Facebook with about 6000 users, we show which network local topology features are useful for such a task (Figure 7). We are considering the possibility of applying this approach to a health-related social network, using the data from a "Patients like me" social network.

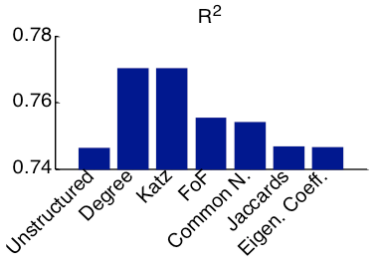


Figure 7. Accuracy ($R^2$) of GCRF using various similarity measures that capture the topology of the neighborhood

*Uncertainty propagation in GCRF*

Prof. Z. Obradovic's team developed an efficient uncertainty propagation model for structured regression that extends GCRF. The model corrects estimated uncertainty by relying on modeling noisy inputs when predicting. In each iteration of this approach, after the prediction is performed for one step, the model updates the input distribution by considering the second moment of multi-dimensional Taylor expansion that approximates the distribution of the target variable expressed by marginalizing noisy inputs, which accounts for the uncertainty resulting from the previous predictions. To the best of our knowledge, this is the first approach developed for structured regression models. In challenging weather-related and healthcare-related applications when predicting up to 96 steps ahead in a network, our extended model greatly improved the uncertainty estimates compared to linear iterative regression and nonlinear iterative Gaussian process models.

*GCRF for directed graphs*

For many real-world applications, structured regression is commonly used for predicting output variables that have some internal structure. GCRFs are a widely used type of structured regression model that incorporate the outputs of unstructured predictors and the correlation between objects in order to achieve higher accuracy. However, applications of this model are limited to objects that are symmetrically correlated, while interaction between objects is asymmetric in many cases. Prof. Z. Obradovic and collaborators developed a new model, called Directed Gaussian conditional random fields (DirGCRF), which extends GCRF to allow modeling asymmetric relationships (e.g. friendship, influence, love, solidarity, etc.). The DirGCRF models the response variable as a function of both the outputs of unstructured predictors and the asymmetric structure. The effectiveness of the proposed model is characterized on six types of synthetic datasets and four real-world applications where DirGCRF was consistently more accurate than the standard GCRF model and baseline unstructured models.

*Continuous Conditional Dependency Network for Structured Regression*

Structured regression on graphs aims to predict response variables from multiple nodes by discovering and exploiting the dependency structure among response variables. This problem is challenging since dependencies among response variables are always unknown, and the associated prior knowledge is non-symmetric. In previous studies, various promising solutions were proposed to improve structured regression by utilizing symmetric prior knowledge, learning sparse dependency structure among response variables, or learning representations of attributes of multiple nodes. However, none of them are capable of efficiently learning dependency structure while incorporating non-symmetric prior knowledge. To achieve these objectives, Prof. Z. Obradovic and his team developed Continuous Conditional Dependency Network (CCDN) for structured regression. The intuitive idea behind this model is that each response variable is not only dependent on attributes from the same node, but also on response variables from all other nodes. This results in a joint modeling of local conditional probabilities. The parameter learning is formulated as a convex optimization problem and an effective sampling algorithm is proposed for inference. CCDN is flexible in absorbing non-symmetric prior knowledge. The performance of CCDN on multiple datasets provides evidence of its structure recoverability and superior effectiveness and efficiency as compared to state-of-the-art alternatives.

## Aim 2: Convex Optimization for Learning Large GCRFs

We developed efficient optimization methods for convex sparse unconstrained minimization problems of the following general form

$$\min_w \; F(w) = \lambda \|w\|_1 + L(w) \qquad (1)$$

where $L: \mathbb{R}^p \to \mathbb{R}$ is convex and twice differentiable and $\lambda > 0$ is the regularization parameter that controls the sparsity of the optimal $w$.

Problems of form (1) have been the focus of much research lately in the fields of signal processing and machine learning. This form encompasses a variety of machine learning models (especially, graphical models), in which feature selection is desirable, such as sparse logistic regression, sparse inverse covariance selection, Lasso. These settings often present common difficulties to optimization algorithms due to their large scale. During the past decade most optimization efforts aimed at these problems focused on development of efficient first-order methods. These methods enjoy low per-iteration complexity, but typically have low local convergence rates. Their performance is often hampered by small step sizes. Due to the very large size of these problems, second order methods are often not a practical alternative. In particular, constructing and storing a Hessian matrix, let alone inverting it, is prohibitively expensive for values of $p$ larger than 10,000, which often makes the use of the Hessian in large-scale problems impractical, regardless of the benefits of fast local convergence rate.

Nevertheless, several new methods were proposed recently for sparse optimization which make careful use of second order information. These methods exploit special properties of the sparse problems and the Hessian of $L(w)$ for specific applications and efficiently apply a coordinate descent method to inexactly solve subproblems arising on each iteration. In our work we further improved upon these ideas to obtain efficient general schemes, which apply beyond the special cases of sparse logistic regression and covariance selection, and in particular, extend to the optimization problems arising in Aim 1.

Many of the methods mentioned above share similar algorithmic features but lack general global convergence analysis. This is due to the use of inexact subproblem optimization and utilization of coordinate descent, which lacks convergence rates. We have proposed modifications to the algorithmic framework which allowed us to prove convergence results and also improve overall efficiency of the general approach.
Along this route we have the following three main results.

1. We have shown that if we replace the usual line search approach by a prox-parameter update mechanism, we can derive sublinear global convergence results for the above methods under mild assumptions on Hessian approximation matrices, which can include diagonal, quasi-Newton and limited memory quasi-Newton approximations. We also provide the convergence rate for the case of inexact subproblem optimization.
2. We have used probabilistic complexity bounds of randomized coordinate descent to show that a very simple stopping criterion can be employed to produce inexact (but accurate enough) solutions to subproblems at each iteration. This gives us the first complete global convergence rate result for the algorithmic schemes for practical proximal Newton-type methods.
3. Finally, we proposed and implemented (in Matlab and C++) an efficient *general purpose* algorithm that uses the same theoretical framework which we analyze, but which does not rely on the special structure of the Hessian, and yet in our tests compares favorably with the state-of-the-art, specialized methods such as QUIC and GLMNET. We replace the exact Hessian computation by the limited memory BFGS Hessian approximations (LBFGS) and exploit their special structure within a coordinate descent approach to solve the subproblems.

*Models based on Conditional Random Fields*

Given data over N locations (nodes) and T time periods, the learning task described in Aim 1 is to choose α and β to minimize the conditional negative log-likelihood plus the $L_1$ regularization to encourage sparse graphical model structure,

$$(\alpha^*, \beta^*) = \arg\min_{\alpha,\beta} f(\alpha,\beta) = \frac{1}{2} y^T X(\alpha,\beta) y - y^T b(\alpha) + \frac{1}{2} b^T(\alpha) X^{-1}(\alpha,\beta) b(\alpha) - \frac{1}{2} \log\det X(\alpha,\beta) + \sum_i \lambda_i |\beta_i|$$

Where $X$ denote the inverse covariance matrix dependent on parameters α and β as follows

$$X(\alpha,\beta) = 2\Big(\sum_i A_i \alpha_i + \sum_i M_i \beta_i + A_0\Big)$$

The gradient of $f(\alpha, \beta)$ can thus be computed as follows,

$$\begin{aligned}\frac{\partial f}{\partial \alpha_i} &= \frac{1}{2}\operatorname{tr}(y^T A_i y) - \hat{B}_i^T y + \hat{B}_i^T X(\alpha,\beta)^{-1} b \\ &\quad - \frac{1}{2} b^T X(\alpha,\beta)^{-1} A_i X(\alpha,\beta)^{-1} b - \frac{1}{2}\operatorname{tr}(X(\alpha,\beta)^{-1} A_i) \\ \frac{\partial f}{\partial \beta_i} &= \frac{1}{2}\operatorname{tr}(y^T M_i y) - \frac{1}{2} b^T X(\alpha,\beta)^{-1} M_i X(\alpha,\beta)^{-1} b - \frac{1}{2}\operatorname{tr}(X(\alpha,\beta)^{-1} M_i)\end{aligned} \tag{2}$$

*Overcoming Computational Bottleneck*

We note that computation of each partial derivative element in (2) is dominated by a vector-matrix-vector product which takes $O(N^2T^2)$ flops, assuming $X^{-1}$ and $X^{-1}b$ are pre-computed and stored as a separate vector; hence, to evaluate a full gradient vector in general takes $O(N^4T^2)$ flops, which amounts to over 10000 billion flops for N = 77 and T = 480, unless structure is taken into account. By exploiting the special structure of the models we consider in Aim 1 and using sparse matrix operations, we are able to reduce the work to $O(N^2T)$. Since there are $O(N^2)$ variables, obtaining the full gradient in $O(N^2T)$ is close to the best complexity we can hope for on a single-core computer.

We have the algorithm implemented in MATLAB using sparse algebra, however one gradient evaluation still takes hours with N = 77 and T = 480. We are investigating several approaches of reducing this complexity by further exploiting the special structure of the model.

*Fast Sparse Gaussian Markov Random Fields Learning Based on Cholesky Factorization*

Learning the sparse Gaussian Markov Random Field, or conversely, estimating the sparse inverse covariance matrix is an approach to uncover the underlying dependency structure in data. Most of the current methods solve the problem by optimizing the maximum likelihood objective with a Laplace prior L1 on entries of a precision matrix. Prof. Z. Obradovic and his team developed a novel objective with a regularization term which penalizes an approximate product of the Cholesky decomposed precision matrix. This new reparametrization of the penalty term allows efficient coordinate descent optimization, which in synergy with an active set approach results in a very fast and efficient method for learning the sparse inverse covariance matrix. We evaluated the speed and solution quality of the newly proposed SCHL method on problems consisting of up to 24,840 variables. Our approach was several times faster than three state-of-the-art approaches. We also demonstrated that SCHL can be used to discover interpretable networks, by applying it to a high impact problem from the health informatics domain

*Structured Regression on Multi-Scale Networks*

To support real-time decision making, an even *faster and exact method was developed by Prof. Obradovic and his team by calculating gradients* much faster than in previous GCRF. We developed the first such method for GCRF learning, which provides a theoretically well-structured foundation to extend the capacity of GCRF and to speed up the model without any approximation. This is achieved by removing matrix inversion when computing the first order derivatives and likelihood function for the parameters optimization. This resulted in an almost 100 times increase in computational speed on a temporal network learned from 35 million observations by restricting the model to a single network and formulating linear bounds on convexity with respect to the model's parameters. An additional important benefit of the new model is that it extends optimization bounds, allowing capture of negative influences of the unstructured predictors, while maintaining the positive semi-definiteness of the precision matrix. We have recently shown that by using multiscale networks the exact solution for graph-based regression can be achieved in logarithmic time with respect to the network size allowing quick learning for networks with millions of nodes and trillions of links. This is achieved by modeling a multiscale network as a Kronecker product of networks and computing the Laplacian of a Kronecker product.

## Aim 3: Non-Stationary and Time Evolving Correlation Analysis in Social Networks

*Evolving Relational Structure Between Data Streams*

There is growing interest in inferring network or relational structure from indirect observations rather than direct observations (e.g., interaction data). An objective of our work was inferring relational structures from time series data and emphasizing structures that can evolve. Here, we envision every time series as a node in a graph.

In our foundational work, we examine methods for describing collections of time series (Approaches 1 and 2 below) where conditioned on the relational structure, the data streams evolve independently. In subsequent work (Approaches 3 and 4), we instead considered cases where the data streams evolve together in a correlated manner. In these cases, we can view the correlation structure itself as the description of interaction between the data streams.

### *Approach 1: Clustering Related Time Series*

One description of relationships between nodes in a graph (e.g., time series) is an idea of *community structure*. We define a community as a collection of time series that each share similar dynamic behaviors. Conditioned on the community assignments of nodes and the parameters defining these communities, the time series evolve independently. Our goal was to discover both the *number of communities* and their *memberships*. In addition to interpretability of the inferred memberships, such community detection allows us to pool information amongst the time series within a community, leading to improved predictive performance in forecasting tasks.

We considered this problem motivated by an application of predicting rates of violent crimes in a large set of regions, in particular, all census tracts in Washington, D.C. We found that spatially *disjoint* regions exhibit correlated crime patterns. It is this indeterminate inter-region correlation structure (i.e., community structure) along with the low-count, discrete nature of counts of serious crimes that motivates our community-structured tool. In particular, we modeled the crime counts in each region using an integer-valued first order autoregressive process (INAR). To flexibly discover a clustering of these region-specific time series in terms of their underlying INAR processes, we took a Bayesian nonparametric approach based on the *Dirichlet process*. In our analysis of weekly reported violent crimes between 2001-2008, we showed that capturing the underlying community structure of these time series (i.e., clustering census tracts with similar dynamics) leads to a mean predictive accuracy in week-ahead forecasts of 97%. These forecasts significantly outperform standard methods while additionally providing useful tools such as prediction intervals that naturally arise from our Bayesian approach.

### *Approach 2: Modeling Complex Dynamics Shared Between Data Streams*

In contrast to the simple autoregressive (AR) dynamics of crime counts in Approach 1, many time series exhibit *non-stationarities*. In many cases, the non-stationarities can be attributed to switches amongst a set of regimes, each of which has a locally simple description of the dynamics, such as a regime-specific AR dynamic. That is, the global non-stationarity of each node is modeled via switches amongst a set of AR processes.

A number of open questions arise in employing such regime-switching models. First, *how many regimes are present*? Second, this model only accounts for the time series generated by a single node in the network. Instead, our goal is to *relate multiple nodes* (i.e, the collection of time series). One notion of "relatedness" is sharing dynamic regimes. We proposed a Bayesian nonparametric approach to the problem of discovering dynamical regimes or *behaviors* shared among the sequences and segmenting each time series into regions defined by a subset of these behaviors. As a concrete example, we considered time series produced by motion capture sensors on the joints of people performing exercise routines. An individual recording provides a multivariate time series that can be segmented into types of exercises (e.g., jumping jacks, arm-circles, and twists). From recordings from multiple individuals, our goal is to discover (1) the global set of exercise types and (2) a relational structure between the recordings based on the subset of behaviors exhibited by each individual. Our key modeling tool is the *beta process*, a Bayesian nonparametric prior that allows us to infer both the size of the behavior set and the sharing pattern from the data.

One of our key developments for this modeling formulation is an efficient, scalable Markov chain Monte Carlo (MCMC)-based inference procedure. In particular, our MCMC algorithm efficiently adds and removes behaviors via novel split-merge moves as well as data-driven birth and death proposals, avoiding the need to consider a truncated model. This procedure allowed us to scale up our computations to jointly examine hundreds of motion capture videos.

### *Approach 3: Adding Sparse, Switching Dependency Structures*

In Approaches 1 and 2, conditioned on the relational structure and associated parameters, the data streams evolve independently. In many applications, the notion of *relatedness* is instead the correlation structure and—crucially—how it evolves. We built upon Approach 2 by modeling *correlated dynamics between the data streams*. For scalability to a large numbers of nodes, we need to define a sparse dependency structure. For this, we use graphical models which encode statements of *conditional independence between nodes in the graph*. We additionally allow the correlations between nodes to change with time, in particular as a function of some underlying event state. That is, we encode a sparse and changing set of dependencies between the nodes using a Markov-switching Gaussian graphical model for the innovation process driving the node dynamics (i.e., a switching AR process).

As a motivating example, we consider an application of this methodology to intracranial electroencephalogram (iEEG) data, where each electrode or *channel* represents a node in the network. The switching AR model of Approach 2 is appropriate due to the nonstationary behavior of EEG signals; Approach 2 likewise allows the channels to have asynchronous switches between dynamic regimes. This network data also exhibits a changing dependency structure as the seizure event progresses. In particular, it is well-known that the correlations between EEG channels usually vary during the beginning, middle, and end of a seizure. These changes are captured by our evolving event state, and corresponding changes in the parameters of the Gaussian graphical model defining the channel correlations. We demonstrated that this model provides better parsing and out-of-sample predictions of iEEG data than the formulation of Approach 2 alone. We also showed that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures. Importantly, our graphical-model based approach allows us to scale to large electrode grids. Our computational burden for examining 96-electrode grids is equivalent to that of roughly 30 electrodes under a full covariance model. Differences would be even more significant with larger numbers of electrodes (i.e., nodes in the graph). In particular, assuming each of N nodes has M neighbors, our per-time step likelihood computations are $O(NM)$ in contrast to $O(N^2)$.

### *Approach 4: Gradually Evolving Correlations*

In Approach 3, we assumed a Markov-switching formulation for the correlation structure and scaled to large numbers of nodes using graphical models. In some applications, instead of having abrupt changes in correlation structure, we expect the correlations to be gradually evolving. We focused on developing a class of Bayesian nonparametric covariance process models, which allow an unknown p **x** p covariance matrix to change flexibly with time. The proposed model harnesses a *low-rank decomposition* of the covariance matrix and introduces a set of dictionary elements to model the low-rank evolution over time. In particular, we consider using Gaussian processes as a flexible dictionary element, describing a smooth, but nonlinear dynamic. To further aid in scaling to high dimensions, we assume that the low-rank evolution can be described as a weighted collection of an even smaller set of basis functions. Our proposed framework leads to a highly flexible, but computationally tractable formulation with simple conjugate posterior updates that can readily handle missing data. We derived theoretical properties of the proposed covariance process. We also considered an application to analyzing the changing correlations in activity in 183 different regions in the United States. In simulated data, we achieved scaling on the order of 5,000 nodes.

As an alternative approach, in contrast to the nonlinear dynamics above, we considered an *autoregressive covariance process*. That is, the AR process evolves on the cone of positive semidefinite (covariance) matrices. To accomplish this restricted multivariate AR evolution, we harnessed inverse Wishart (IW) theory to define a process that maintains inverse Wishart margins and, crucially, defines a nice form for the covariance transitions. We call the resulting model the IW-AR. We used the IW-AR to analyze multi-channel scalp EEG data recorded while a patient underwent an induced seizure.

### *Scaling Up Determinantal Point Processes*

In applications such as clustering (Approach 1) or segmentation across time (Approach 2 and 3), interpretability of the resulting clusters (i.e., communities or regime-specific dynamical models) can be improved if the clusters themselves are diverse from one another. In typical clustering approaches, there is nothing driving such repulsion between clusters (i.e., penalizing overlap). To address this, we examined a class of repulsive processes known as *determinantal point processes* (DPPs).

In machine learning, the focus of DPP-based models has been on diverse subset selection from a discrete and finite base set of N items. This discrete setting admits an efficient sampling algorithm based on the eigendecomposition of the defining kernel matrix resulting in $O(N^3)$ complexity. However, in certain applications, N may be so large that sampling from a DPP becomes computationally infeasible. For DPPs on continuous spaces, as commonly encountered in mixture modeling applications for clustering and segmentation, existing sampling results rely on rejection sampling leading to low

acceptance rates in all but the simplest of scenarios.

We addressed these issues by considering approximations to the DPP kernel that allow for efficient and closed-form sampling. In the large, finite N scenario, we proposed applying Nystrom approximation and derived error bounds for the approximated DPP in terms of variational distance (quite unlike the standard matrix norms previously established for kernel approximations). Random Fourier feature kernel approximations were also considered. Using such approximations, we then developed a closed-form method of sampling from DPPs on continuous and potentially multivariate spaces. We applied this continuous-DPP sampler to the task of repulsive mixture modeling and showed improved interpretability and classification performance for our inferred clustering.

## *Adaptive Skip-Train Ensemble-Based Structured Regression*

Forecasting in temporal networks is commonly done by employing a single model in order to predict the response for each node in one or multiple upcoming timesteps. Issues arise when the time for prediction is limited, and the computational and space complexities increase when learning from multiple previous timesteps. To overcome these limitations, one can train quickly an unstructured learner at the current time-step to perform a one-step ahead prediction. However, this can lower accuracy since unstructured models are not capable of capturing between-node correlations. Structured learners such as GCRFs may be more accurate, but they require more time for retraining at each timestep. Moreover, they consider the whole network structure while learning, without taking advantage of useful substructures within the network. Driven by these issues, Prof. Z. Obradovic's team proposed an adaptive ensemble-based model to automatically decide whether to skip the majority of unnecessary computations, which come from retraining at each time-step and make predictions in a timely and accurate manner (see Fig. 8). In order to achieve greater predictive performance, the proposed model incorporates multiple GCRFs into a single composite structured ensemble. Then, to capture the hidden network substructures, GCRFs are trained simultaneously on subnetworks generated by subsampling, which *decreases complexity* and *increases scalability*.

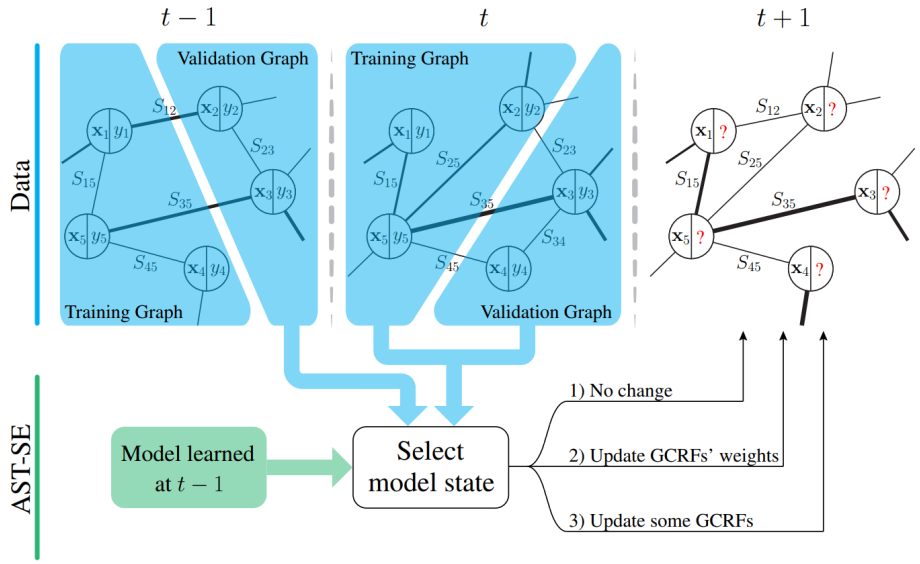


Fig. 8. Skip-training between consecutive timesteps.

Our Skip-Train structured model outperforms competitors, while learning in a more *efficient*, *scalable*, and *potentially even more accurate* manner. Our findings suggest that the model: (1) runs ~140 and ~4.5 times faster than GCRF and ensemble-based alternatives, respectively; (2) focuses on partial views of a network, and therefore, it is scalable as the network size expands; and (3) it is ~34-41% more accurate than alternatives.

## Publications List:

The results of Temple University DARPA GRAPHS project FA9550-12-1-0406 are published at the following 52 manuscripts:


1. Jordanski, M, Radovic, M., Milosevic, Z., Filipovic, N., Obradovic, Z. (2018) "Machine Learning Approach for Predicting Wall Shear Distribution for Abdominal Aortic Aneurysm and Carotid Bifurcation Models," *IEEE Journal of Biomedical and Health Informatics,* March 2018, Vol. 22, Issue 2, pp. 537-544.
2. Vujicic, T., Glass, J., Zhou, F., Obradovic, Z. (2017) "Gaussian Conditional Random Fields Extended for Directed Graphs," *Machine Learning,* October 2017, Vol 106, Issue 9-10, pp. 1271-1288.
3. Glass, J., Obradovic, Z. (2017) "Structured Regression on Multi-Scale Networks," *IEEE Intelligent Systems*, Vol. 32, Issue 2, Mar-April, 2017, pp. 23-30.
4. Stojkovic, I. Ghalwash, M., Obradovic, Z. (2017) "Ranking Based Multitask Learning of Scoring Functions," In: *Machine Learning and Knowledge Discovery in Databases. ECML PKDD* 2017, pp. 721-736, 2017**.**
5. Pavlovski, M., Zhou, F., Stojkovic, I., Kocarev, Lj., Obradovic, Z. (2017) "Adaptive Skip-Train Structured Regression for Temporal Networks," In: *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2017*, pp: 305-321, 2017.
6. Roychoudhury, S. Ghalwash, M., Obradovic, Z. (2017) "Cost Sensitive Time-Series Classification," In: *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2017*, pp**.** 495-511, 2017.
7. Stojkovic, I., Jelisavcic, V., Milutinovic, V., Obradovic, Z. (2017) "Fast Sparse Gaussian Markov Random Fields Learning Based on Cholesky Factorization," *Proc. 26th International Joint Conference on Artificial Intelligence (IJCAI)*, Melbourne, Australia, August 2017, pp. 2758-2764.
8. Han, C, Ghalwash, M., Obradovic, Z. (2017) "Continuous Conditional Dependent Network for Structured Regression," *Proc. 31st AAAI Conference on Artificial Intelligence (AAAI-17),* San Francisco, CA, February 2017, 1962-1968.
9. Albarakati, N., Obradovic, Z. (2017) "Disease-Based Clustering of Hospital Admission: Disease Network of Hospital Networks Approach," *Proc. 30th IEEE Int'l Symp. Computer-Based Medical Systems*, Thessaloniki, Greece, June, 2017.
10. Zhou, F., Ghalwash, M., Obradovic, Z. (2016) "A Fast Structured Regression for Large Networks," *Proc. 2016 IEEE International Conference on Big Data,* Washington, DC, Dec. 2016, pp. 108-115.
11. Mirowski, T., Roychoudhury, S., Zhou, F., Obradovic, Z. (2016) "Predicting Poll Trends using Twitter and Multivariate Time-series Classification," *Proc. 8th Int'l Conf. Social Informatics (SocInfo),* Seattle, WA, Nov. 2016, 273-289.
12. Stojanovic, J., Gligorijevic, Dj., Obradovic, Z. (2016) "Modeling Customer Engagement from Partial Observations," *Proc. 25th Int'l Conf. Information and Knowledge Management (CIKM-16),* Indianapolis, IN, Oct. 2016, pp. 1403-1412.
13. Gligorijevic, Dj., Stojanovic, J., Djuric, N., Radosavljevic, V., Grbovic, M., Kulathinal, R.J., Obradovic, Z. (2016) "Large-Scale Discovery of Disease-Disease and Disease-Gene Associations," *Scientific Reports, Nature Publishing Group,* 2016, Aug 31, 6:32404 doi 10.1038/srep32404. (PDF)
14. Feldman, K., Stiglic, G., Dasgupta, D., Kricheff, M., Obradovic, Z., Chawla, N. (2016) "Insights into Population Health Management Through Disease Diagnoses Networks," *Scientific Reports, Nature Publishing Groups,* 2016, July 27, 6:30465 doi: 10.1038/srep30465. (PDF)

15. Gligorijevic, Dj., Stojanovic, J., Obradovic, Z., (2016) "Disease Types Discovery from a Large Database of Inpatient Records: A Sepsis Study," *Methods,* Jul 28. S1046-2023(16)30232-8, doi: doi:10.1016/j.ymeth.2016.07.021. (PDF)
16. Stojanovic, J., Gligorijevic, Dj., Radosavljavic, V., Djuric, N., Grbovic, M., Obradovic, Z., (2016) "Modeling Healthcare Quality via Compact Representations of Electronic Health Records," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* Jul 14. Vol 14, Issue 3, pp. 545-554 doi:10.1109/TCBB.2016.2591523. (PDF)
17. Stojkovic, I., Jelisavcic, V., Milutinovic, V., Obradovic, Z. (2016) "Distance Based Modeling of Interactions in Structured Regression," *Proc. 25th International Joint Conference on Artificial Intelligence (IJCAI),* New York, NY, July 2016, pp. 2032 - 2038 (PDF)
18. Han, C, Zhang, S., Ghalwash, M., Vucetic, S, Obradovic, Z. (2016) "Joint Learning of Representation and Structure for Sparse Regression on Graphs," *Proc. 16th SIAM Int'l Conf. Data Mining (SDM),* 846 - 854 Miami, FL, May 2016. (PDF)
19. Polychronopoulou, A, Obradovic, Z. (2016) "Structured Regression on Multilayer Networks," *Proc. 16th SIAM Int'l Conf. Data Mining (SDM),* 612 - 620, Miami, FL, May 2016. (PDF)
20. Gligorijevic, Dj, Stojanovic, J., Obradovic, Z. (2016) "Uncertainty Propagation in Long-term Structured Regression on Evolving Networks," *Proc. Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16),* 1603-1610, Phoenix, AZ, February 2016. (PDF)
21. Glass, J., Ghalwash, M., Vukicevic, M., Obradovic, Z. (2016) "Extending the Modeling Capacity of Gaussian Conditional Random Fields while Learning Faster," *Proc. Thirtieth AAAI Conference on Artificial Intelligence, (AAAI-16),*1596 - 1602, 2016 Phoenix, AZ, February 2016. (PDF)
22. Dokic, T. Dehghanian, P., Chen, P.-C, Kezunovic, M., Medina-Cetina, Z., Stojanovic, J., Obradovic, Z. (2016) "Risk Assessment of a Transmission Line Insulation Breakdown due to Lightning and Severe Weather," *Proc. HICCS – Hawaii International Conference on System Science,* Kauai, Hawaii, January 2016.pp. 2488 - 2497 (PDF)
23. Stiglic, G., Brzan, P.P., Fijacko, N., Fei, W., Delibasic, B., Kalousis, A., Obradovic, Z. (2015) "Comprehensible Predictive Modeling Using Regularized Logistic Regression and Comorbidity Based Features," *PLOS ONE,Dec 8, 2015, DOI: 10.1371/journal.pone.0144439* (PDF)
24. Ma, Y.A., Chen, T. and Fox, E.B. (2015) "A Complete Recipe for Stochastic Gradient MCMC," submitted to Neural Information Processing Systems (NIPS), Montreal, Quebec December 2015, pp. 2917-2915.
25. Vukicevic, M., Radovanovic, S., Kovacevic, A., Sliglic, G., Obradovic, Z. (2015) "Improving hospital readmission prediction using domain knowledge based virtual examples," *Proc. the 10th Conf. on Knowledge Management in Organization*, Maribor, Slovenia, August, 2015, pp. 695-704. (PDF)
26. Tank, A. , Foti, N. and Fox, E.B. (2015) "Bayesian Structure Learning of Stationary Time Series," *Uncertainty in Artificial Intelligence (UAI),* Amsterdam, Netherlands, July 2015.
27. Radovanovic, S., Vukicevic, M., Kovacevic, A., Sliglic, G., Obradovic, Z. (2015) "Domain knowledge based hierarchical feature selection for 30-day hospital readmission prediction" *Proc. AIME 2015, the 15th Conference on Artificial Intelligence in Medicine*, Pavia, Italy, June, 2015, Artificial Intelligence in Medicine, Lecture Notes in Computer Sciences, vol. 9105, pp. 96-100. (PDF)
28. Tank, A., Foti, N. and Fox, E.B. (2015) "Streaming Variational Inference for Bayesian Nonparametric Mixture Models," *International Conference on Artificial Intelligence and Statistics (AISTATS),* San Diego, CA May 2015.
29. Ramljak, D., Davey, A., Uversky, A., Roychoudhury, S., Obradovic, Z. (2015) "Hospital Corners and Wrapping Patients in Markov Blankets," *4th Workshop on Data Mining for Medicine and Healthcare, 2015 SIAM International Conference on Data Mining,* Vancouver, Canada, April 30 - May 02, 2015 (PDF)
30. Gligorijevic, Dj., Stojanovic, J., Obradovic, Z. (2015) " Improving Confidence while Predicting Trends in Temporal Disease Networks," *4th Workshop on Data Mining for Medicine and*

*Healthcare, 2015 SIAM International Conference on Data Mining,* Vancouver, Canada, April 30 - May 02, 2015 (PDF)
31. Stojanovic, J., Jovanovic, M., Gligorijevic, Dj., Obradovic, Z. (2015) " Semi-supervised learning for structured regression on partially observed attributed graphs" *Proceedings of the 2015 SIAM International Conference on Data Mining (SDM 2015)* Vancouver, Canada, April 30 - May 02, 2015 (PDF)
32. Uversky, A., Ramljak, D., Radosavljevic, V., Ristovski, K., Obradovic, Z. (2014) "Panning for Gold - Using Variograms to Select Useful Connections in a Temporal Multigraph Setting," *Social Network Analysis and Mining*, 4:211, July 2014 (PDF)
33. Wulsin, D., Fox E.B., Litt B. (2014) "Modeling the complex dynamics and changing correlations of epileptic events," *Artificial Intelligence* 216, 55-75, 2014.
34. Bandeira A.S., Scheinberg K., Vicente L. N. (2014) "Convergence of trust-region methods based on probabilistic models," *SIAM Journal on Optimization*, 24(3), 1238-1264, 2014.
35. N. Foti, J. Xu, D. Laird, and E.B. Fox, "Stochastic Variational Inference for Hidden Markov Models," Neural Information Processing Systems (NIPS), Montreal, Quebec December 2014, pp. 3599-3607.
36. Tank, A., Foti, N. and Fox, E.B. (2014) "Streaming Variational Inference for Normalized Random Measure Mixture Models," *NIPS Workshop on Advances in Variational Inference*, Montreal, Quebec December 2014.
37. Polychronopoulou, A., Obradovic, Z. (2014) "Hospital Pricing Estimation by Gaussian Conditional Random Fields Based Regression on Graphs" *Proc. 2014 IEEE International Conference on Bioinformatics and Biomedicine,* Belfast, UK, Nov. 2014, pp. 564-567. (PDF)
38. Stiglic, G., Wang, F., Davey, A., and Obradovic, Z. (2014) "Readmission Classification Using Stacked Regularized Logistic Regression Models," *Proc. AMIA 2014 Annual Symposium*, Washington, DC, Nov. 2014, pp. 1072-81. (PDF)
39. Radosavljevic, V., Vucetic, S., Obradovic, Z. (2014) "Neural Gaussian Conditional Random Fields," *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases,* Nancy, France, September, 2014, pp. 614-629. (PDF)
40. Slivka, J., Nikolic, M., Ristovski, K., Radosavljevic, V., Obradovic, Z. (2014) "Distributed Gaussian Conditional Random Fields Based Regression for Large Evolving Graphs," *Proc. 14th SIAM Int'l Conf. Data Mining Workshop on Mining Networks and Graphs,* Philadelphia, April 2014. (PDF)
41. Scheinberg, K and Tang X., (2014) "Practical inexact proximal quasi-Newton method with global complexity analysis," Mathematical Programming, 3, 2014. pp 1-35.
42. Ouzienko, V., Obradovic, Z. (2013) "Imputation of Missing Links and Attributes in Longitudinal Social Surveys," *Machine Learning Journal*, vol. 95, no. 3, pp. 329-356. (PDF)
43. Stiglic, G., Davey, A., Obradovic, Z.,(2013) "Temporal Evaluation of Risk Factors for Acute Myocardial Infarction Readmissions," *Proc. 2013 IEEE International Conference on Healthcare Informatics, Workshop on Hospital Readmission Prediction and Clinical Risk Management*, Philadelphia 2013 (PDF)
44. Uversky, A., Ramljak, D., Radosavljevic, V., Ristovski, K., Obradovic, Z. (2013) "Which Links Should I Use? A Variogram Based Selection of Relationship Measures for Prediction of Node Attributes in Temporal Multigraphs," *Proc. 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* Niagara Falls, Canada, Aug. 2013. (PDF)
45. Ristovski, K., Radosavljevic, V., Vucetic, S., Obradovic, Z. (2013) "Continuous Conditional Random Fields for Efficient Regression in Large Fully Connected Graphs," *Proc. The Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI-13)*, Bellevue, Washington, July 2013, pp. 840-846. (PDF)
46. Affandi R.H., Fox E.B., Taskar, B., (2013) "Approximate inference in continuous determinantal processes," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

47. Affandi, R. H., Kulesza, A., Fox, E. B., Taskar, B., (2013) “Nystrom approximation for large-scale determinantal processes,” *Proc. Int’l Conf. Artificial Intelligence and Statistics*, 2013.
48. Aldor-Noiman, S., Brown, L. D., Fox E.B., Stine R.A., (2013) “Spatio-temporal low count processes with application to violent crime events,” arXiv:1304.5642, 2013.
49. Fox, E.B., Hughes M., Sudderth E.B., Jordan, M.I. (2013) “Joint modeling of multiple time series via the beta process with application to motion capture segmentation,” arXiv:1308.4747, 2013.
50. Wulsin D., Fox E.B., Litt B., (2013) “Parsing epileptic events using a Markov switching process model for correlated time series,” *Proc. Int’l Conf. Machine Learning*, pages 356-364, 2013.
51. El-Arini K., Xu M., Fox E.B., Guestrin C.E., (2013) “Representing Documents Through Their Readers,” *Proc. Int’l Conf. on Knowledge Discovery and Data Mining (KDD),* August 2013.
52. Mathew, G., Obradovic, Z. (2012) "Distributed Privacy Preserving Decision System for Predicting Hospitalization Risk in Hospitals," *Proc. 11th International Conference on Machine Learning and Applications: Machine Learning in Health Informatics Workshop*, Boca Raton, Florida, Dec. 2012. (PDF)