Network Reconstruction from Dynamic Data

K.P. Unnikrishnan[†], Naren Ramakrishnan[*], P.S. Sastry[+], and Ramasamy Uthurusamy[$]
[†] General Motors R&D Center, Warren, MI 48090, USA
[*] Department of Computer Science, Virginia Tech, Blacksburg VA 24061, USA
[+] Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India
[$] General Motors IS& S, Detroit, MI 48207, USA

ABSTRACT

Over the past decade, many powerful data mining techniques have been developed to analyze temporal and sequential data. The time is now fertile for addressing problems of larger scope under the purview of temporal data mining. The fourth SIGKDD workshop on temporal data mining focused on the question: *What can we infer about the structure of a complex dynamical system from observed temporal data?* The goals of the workshop were to critically evaluate the need in this area by bringing together leading researchers from industry and academia, and to identify promising technologies and methodologies for doing the same. We provide a brief summary of the workshop proceedings and ideas arising out of the discussions.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications - Data Mining; I.2.6 [Artificial Intelligence]: Learning

General Terms: Algorithms.

Keywords: temporal data mining, network reconstruction.

1. INTRODUCTION

Established in 2001, the SIGKDD workshop series on temporal data mining (TDM) is aimed at inferring patterns from large databases that contain either explicit or implicit temporal information. Over the past decade, many powerful data mining techniques have been developed to analyze temporal and sequential data. The time is now fertile for addressing problems of larger scope under the purview of temporal data mining. The fourth SIGKDD workshop on temporal data mining hence focused on the question: *What can we infer about the structure of a complex dynamical system from observed temporal data?* This topic of reconstructing system dynamics from sequential data traces is an important one in many areas:

- Neuroscience: Determining functional connectivity in neuronal systems form multi-electrode data;
- Genetics: Inferring gene regulatory networks from time-series of gene expression measurements;
- Epidemiology: Disease spread modeling from people movement data;
- Chemical Engineering: Chemical process and pathway reconstruction from concentration measurements;
- Manufacturing: Root-cause diagnostic inference from plant-floor data; and
- Automotive: Prognostics and fault diagnostics from vehicle data.

In all these applications, the aim is to construct the underlying system model (reflecting connectivity, hierarchy, and/or strength of influences) from observed time-indexed discrete symbol sequences (and, sometimes, continuous-valued measurements). In many of the areas mentioned above, there are isolated pieces of work (see for example, [1], [2]) beginning to appear. A special area of interest to the organizers is neuroscience, where this approach can help discover neural codes and facilitate the creation of brain-computer interfaces [3].

The fourth SIGKDD (half-day) workshop on temporal data mining served as a forum to discuss network reconstruction as a concerted theme, critically evaluate the need in this area by bringing together leading researchers from industry and academia, and identify technologies and methodologies that worked (and didn't) in specific application contexts. Invited speakers included Vijay Nair (University of Michigan), Bud Mishra (New York University), C. Lee Giles (Penn State University), and Vinod Sharma (Indian Institute of Science). In addition, the organizers contributed two papers as background work. The workshop also featured the release of a challenge dataset from computational neuroscience that embodied multiple facets of network reconstruction.

2. NETWORK RECONSTRUCTION IN MANY GUISES

As evidenced by the invited talks, network reconstruction re-surfaces in multiple contexts: network tomography (Nair and Sharma), social networks (Giles), and bioinformatics (Mishra).

2.1 Network Tomography

There are interesting challenges in collecting and analyzing data from computer and communication networks for the purpose of assessing and monitoring quality of service characteristics. The talk by Nair—'Computer and communications networks: assessing and monitoring quality of service'—gave an overview of the field of network tomography, with particular mention of two classes of network tomography problems and related research on network monitoring. The basic idea is to send test probes from source nodes to destination nodes, observe end-to-end latency and

loss characteristics, and use this information to infer individual link characteristics. In passive network tomography, we require 'buy-in' from the network nodes, whereas in active network tomography, we use unicast or multicast schemes to estimate internal network characteristics. Most of the research in network tomography assumes that the logical topology is known but newer work focuses on restricted classes of topologies and solves identifiability problems in the chosen contexts. These and other ideas were covered in detail in Nair's talk. Following this introduction, the talk by Sharma—' Estimating traffic intensities in a communication network via active network tomography'—further developed network tomography problems and analyzed optimal (in space and time) power transmission policies, and characterized many tradeoffs underlying inference in communication system design.

2.2 Social Networks

The increasing amount of communication between individuals in e-formats (e.g. email, instant messaging and the web) has motivated computational research in social network analysis (SNA). Previous work in SNA has emphasized the social network (SN) topology measured by communication frequencies while ignoring the semantic information in SNs. The talk by Giles—'Probabilistic models for discovering temporal semantic social networks'—proposed two generative Bayesian models for semantic community discovery in SNs, combining probabilistic modeling with community detection in SNs. To simulate the generative models, an EnFGibbs sampling algorithm was proposed to address the efficiency and performance problems of traditional methods. Experimental studies on Enron email corpus showed that this approach successfully detected the communities of individuals and in addition provides semantic topic descriptions of these communities.

2.3 Bioinformatics

The talk by Mishra—'Remembrance of experiments past: analyzing time course datasets to discover complex temporal invariants'—focused on reconstructing networks that explain gene expression signatures underlying temporal datasets. Current microarray data analysis techniques draw the biologist's attention to targeted sets of genes but do not otherwise present global and dynamic perspectives (e.g., invariants) inferred collectively over a dataset. Such perspectives are important in order to obtain a process-level understanding of the underlying cellular machinery; especially how cells react, respond, and recover from environmental changes. Mishra described GOALIE (Gene-Ontology for Algorithmic Logic and Invariant Extractor), a novel computational approach and software system that uncovers formal temporal logic models of biological processes from time course microarray datasets. GOALIE 'redescribes' data into the vocabulary of biological processes and then pieces together these redescriptions into a Kripke-structure model, where possible worlds encode transcriptional states and are connected to future possible worlds. Such a model then supports various query, inference, and comparative assessment tasks, besides providing descriptive process-level summaries.

3. CONTRIBUTED PAPERS

The paper by Fernandez et al.—'Reconstructing Partial Orders from Linear Extensions'—posed network reconstruction as a problem of recovering partial order descriptions of sequential traces (linear extensions). In addition to giving complexity bounds for specific classes of problems, the paper presented a general framework to pose and study various inference tasks, and algorithmic results for mining restricted classes of posets. The paper by Patnaik et al.—'Discovering Network Patterns in Microelectrode Array Data'—presented analysis techniques that can unearth interesting regularities involving combinations of neurons from multi-electrode array (MEA) data. MEA recording is a relatively new experimental technique in neurobiology for studying simultaneous activity of groups of neurons. Patnaik et al show, through simulations, that by combining discovery of different types of episodes with suitable temporal constraints, one can discover the network structures and connectivity patterns of the neurons constituting the network.

4. CHALLENGE DATASET

The synthetic dataset released at the workshop was intended to resemble spike sorted, simultaneously recorded, multineuronal data. Five different acyclic connectivity patterns were 'planted' in the dataset using a data generation model based on frequent episode discovery. Essentially, we begin with 26 neurons in the network, labeled A-Z. The neurons are first randomly interconnected with the weight of the connection uniformly distributed in $[-1, 1]$. After that, one or more patterns are introduced by modifying the appropriate interconnection weights. Interestingly, the talk by Bud Mishra presented a successful mining of this dataset (based on the GOALIE approach) to reconstruct the planted patterns! We hope that this dataset will help seed further research in network reconstruction.

5. DISCUSSION

Network reconstruction from dynamic data is a fertile data mining problem; there are many important application areas besides those studied at the workshop, such as chemical reaction modeling and epidemiology. It is anticipated that the proceedings of this workshop will help seed further consolidation of network reconstruction research and contribute to a core body of algorithms, software, and datasets.

6. REFERENCES

[1] A. Arkin, P. Shen, and J. Ross. A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements. *Science*, Vol. 277(5330):pages 1275–1279, Aug 1997.

[2] P.E. Barbano, M. Spivak, J. Feng, M. Antoniotti, and B. Mishra. A Coherent Framework for Multiresolution Analysis of Biological Networks with "Memory": Ras Pathway, Cell Cycle, and Immune System. *PNAS*, Vol. 102(18):pages 6245–6250, May 2005.

[3] G. Santhanam, S.I. Ryi, B.M. Yu, A. Afshar, and K.V. Shenoy. A High Performance Brain Computer Interface. *Nature*, Vol. 442:pages 195–198, 2006.