NETWORK DATA MODELING VIA GRAMMATICAL STRUCTURES

By

Sahin Cem Geyik

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: COMPUTER SCIENCE

Approved by the
Examining Committee:

_____
Professor Boleslaw K. Szymanski, Thesis Adviser

_____
Professor Christopher D. Carothers, Member

_____
Assistant Professor Sanmay Das, Member

_____
Associate Professor Koushik Kar, Member

_____
Professor Alex Pentland, Member

Rensselaer Polytechnic Institute
Troy, New York

June 2012
(For Graduation August 2012)

ABSTRACT

Data modeling in computer networks means finding a compact and faithful representation of the data on which network applications work. This data can range from the information that is being transferred in a wireless network to mobility information of nodes in a mobile network, or to the observations of the entity behaviors recorded by the network, etc. In any kind of application, data modeling holds an important place if high efficiency is a requirement.

In this thesis, we examine the use of grammatical structures, with a high emphasis on probabilistic context free grammars (PCFG), as the modeling framework for such data. Informally, PCFGs are regular context free grammars where production rules are assigned a probability value, representing how likely these rules are to be used when generating a sentence from this grammar. Utilization of PCFGs includes initially deriving the grammatical structure from the real world traces, and later applying the structure for the necessary purposes of the application. These purposes include saving disk space in space-constrained systems (such as sensor networks), classification and recognition of observed events, facilitating manual inspection (i.e. improving interpretability), etc. The subject matter presented in this thesis contains both the grammar construction process as well as different domains to which the grammatical structure can be applied.

The contributions of this PhD thesis can be summarized as follows:

- An automated Probabilistic Context Free Grammar (PCFG) construction algorithm, which lowers the time complexity of the previous approaches. We also evaluate (both theoretically and empirically) how much of an advantage we provide over the previous approaches complexity-wise, and the comparison of the goodness of grammars constructed with multiple methods.

- Application of the PCFG learning and later processing approach to different domains of Network Data Modeling. These domains include:

  - Event Recognition in Sensor Networks,
  - Mobility Modeling and Synthetic Trace Generation for Mobile Networks,
  - Behavior Modeling in Social Networks, and,
  - Learning of Service Composition Rules for Service Oriented Architecture in Sensor Networks.

- We also contribute to metric-based service composition in sensor networks, as well as to application of switch options in pervasive sensor applications.