Ad Hoc Networks 000 (2016) 1–20

Contents lists available at ScienceDirect

Ad Hoc Networks

journal homepage: www.elsevier.com/locate/adhoc

# Big data and semantics management system for computer networks

Bassem Mokhtar[a,*], Mohamed Eltoweissy[b,1]

[a] Department of Electrical Engineering, Faculty of Engineering, Alexandria University, Egypt
[b] Department of Computer and Information Sciences, Virginia Military Institute, USA

ARTICLE INFO

*Article history:*
Received 16 March 2016
Accepted 24 June 2016
Available online xxx

*Keywords:*
Network management
Big data
Bio-inspired design
Semantics reasoning
Pattern learning
Hybrid intelligence

ABSTRACT

We define "Big Networks" as those that generate big data and can benefit from big data management in their operations. Examples of big networks include the current Internet and the emerging Internet of things and social networks. The ever-increasing scale, complexity and heterogeneity of the Internet make it harder to discover emergent and anomalous behavior in the network traffic. We hypothesize that endowing the otherwise semantically-oblivious Internet with "memory" management mimicking the human memory functionalities would help advance the Internet capability to learn, conceptualize and effectively and efficiently store traffic data and behavior, and to more accurately predict future events. Inspired by the functionalities of human memory, we proposed a distributed network memory management system, termed NetMem, to efficiently store Internet data and extract and utilize traffic semantics in matching and prediction processes. In particular, we explore Hidden Markov Models (HMM), Latent Dirichlet Allocation (LDA), and simple statistical analysis-based techniques for semantic reasoning in NetMem. Additionally, we propose a hybrid intelligence technique for semantic reasoning integrating LDA and HMM to extract network semantics based on learning patterns and features with syntax and semantic dependencies. We also utilize locality sensitive hashing for reducing dimensionality. Our simulation study using real network traffic demonstrates the benefits of NetMem and highlights the advantages and limitations of the aforementioned techniques.



## 1. Introduction

Due to semantically-oblivious protocol operations, the current Internet cannot effectively or efficiently cope with the explosion in services with different requirements, number of users, resource heterogeneity, and widely varied user, application and system dynamics [1]. This is leading to increasing complexity in Internet management and operations, thus multiplying the challenges to achieve better security, performance and QoS satisfaction. The current Internet largely lacks capabilities to extract network semantics to efficiently build runtime accessible dynamic behavior models of Internet elements at different levels of granularity to pervasively observe, analyze, predict and act upon network dynamics. For example, a network host might know the role of TCP, however, it might not know the behavior of TCP in mobile ad hoc networks. We refer to the limited utilization of Internet traffic semantics in networking operations as the "Internet Semantic Gap".

The current and future internets (e.g., Internet of things [2]) support a massive number of Internet elements with extensive amounts of data. Fortunately these data generally exhibit multi-dimensional patterns (e.g., patterns with dimensions such as time, space, and users) that can be learned to extract network semantics [3]. These semantics can help in learning normal and anomalous behavior of the different elements (e.g., services, protocols, etc.) in the Internet and in building behavior models for those elements accordingly. Recognizing and maintaining semantics as accessible behavior models related to various Internet elements will aid in possessing intelligence thus helping elements in predicting future events (e.g., QoS degradation and attacks) that might occur and affect performance of networking operations. Furthermore, learning behavior of those elements will enhance their self-∗ properties such as awareness with unfamiliar services and also advance reasoning about their behavior. For instance, a router can classify a new running service in a network as a specific type of TCP-based file transfer service when it finds similarity between behavior of that new service and that of an already known service.

There is a need to endow Internet operations and running services and applications with intelligence to mitigate the Internet semantic gap. In the literature, Internet (or network) intelligence (referred to here as InetIntel) is defined as the capability of Internet elements to understand network semantics to be able to make effective decisions and use resources efficiently [4]. InetIntel has to support Internet elements with the capability for learning normal

* Corresponding author.
*E-mail address:* bmokhtar@alexu.edu.eg (B. Mokhtar).
[1] The author is also affiliated with the ECE department at Virginia Tech and University of Arizona, USA

http://dx.doi.org/10.1016/j.adhoc.2016.06.013
1570-8705/

and dynamic/emergent behavior of various elements and in turn building dynamic behavior models of those elements.

InetIntel can be achieved via employing intelligence techniques to efficiently reason about semantics from tons of Internet traffic raw data and provide runtime accessible valuable information at different levels of granularity. InetIntel should be achieved in a way that will not negatively impact Internet robustness or scalability. InetIntel systems may use either monolithic or hybrid intelligence techniques (HIT). Each implemented technique has its mechanisms for learning data patterns, extracting features and reasoning about data semantics. The Internet has tremendous and ever-growing scale. It is noisy and dynamic with dissimilar communicating networks and heterogeneous entities, running diverse services and resources. Accordingly, generated and transmitted Internet data have special characteristics such as massive volume with high- and multi-dimensionality that might negatively affect the performance of intelligence techniques. HIT are being investigated to better mitigate that challenge [5]. For example, in [6], comparison among various InetIntel techniques for intrusion detection systems showed HIT's superiority in achieving higher detection accuracy. Some works (e.g., [7,8]) proposed HIT for InetIntel using neural network (NN) with evolutionary algorithms. But, they did not provide solutions for extracting semantics from high data volume with large number of attributes mitigating NN's challenges in their complicated design and long processing time at large scale problems.

In [9–11], we proposed a preliminary design of a network memory system, termed NetMem, to support smarter networking and InetIntel. NetMem design is inspired by the functionalities of human memory [12], which maintains conceptual models that describe associative concepts according to learned multi-dimensional patterns of data captured from the outside world through human sensory system. Those models are updated continually and used for learning novel things and predicting future events achieving human intelligence. Analogy with human memory's functionalities, NetMem has a memory system structure comprising short-term memory (StM) and long-term memory (LtM). StM maintains highly dynamic network data or data semantics with lower levels of abstraction for short-time, while LtM keeps for long-time slower varying semantics with higher levels of abstraction. Maintained data in NetMem can be retrieved at runtime and on-demand to be used in matching and prediction processes within the various networking operations. From a system's perspective, NetMem can be viewed as an overlay network of distributed "memory" agents, called NMemAgents, located at multiple levels targeting different levels of data abstraction and scalable operation.

In our preliminary design [9–11], NetMem utilized monolithic intelligence techniques, e.g., Hidden Markov Models (HMM) [13], for learning multi-dimensional data patterns to reason about network data semantics and to build conceptual models accordingly. HMM are computationally efficient and well-suited to handle new raw data and extract inter-related network concepts at different levels of granularity. We classified concepts using the separation of network concerns (application, communication and resource concern) as presented in [14]. Concepts are represented using functional, behavioral, and structural (FBS) engineering design framework [15]. For example, NetMem can learn data semantics concerning TCP and UDP protocols (i.e., communication concerns) within file transfer services (i.e., an application concern) in wireless contexts (i.e., a communication concern). Those semantics can be used to construct a conceptual model, which describe: a) functions (i.e., functional aspect) of TCP and UDP protocols; b) the normal and abnormal behavior (i.e., behavioral aspect) of those protocols and the different behavior classes that can emerge (e.g., different attacks under the abnormal behavior class); c) relations (i.e., structural aspect) among concept classes such as overlapped TCP abnormal behavior classes that share common features.

Network data characteristics include massive volume, high- and multi-dimensionality, dynamicity and complexity (variety in representation models and languages). In [16], authors proposed a generative model based on Latent Dirichlet Allocation (LDA) [17] and HMM for learning words with short-range syntax and long-range semantics dependencies. Consequently, this aids in forming richer ontology with more associated semantic topics and classes. We surmise that there are similarities between characteristics (e.g., huge volume, high dimensionality, and complexity) of datasets in networks and language modeling. In this paper, we explore different reasoning models for NetMem to learn patterns and extract semantics from big network data both with and without data dimensionality reduction via locality-sensitive hashing (LSH) algorithm [18]. We show capabilities of each model in extracting features and learning semantics based on real captured data by snort [19] and data attributes discovered and classified using associative rule learning (ARL) [20] and fuzzy membership functions (FMF) [21]. Extracted semantics are represented and classified as concept classes using the separation of network concerns (application, communication and resource (ACR) concern) as presented in [14]. Concept classes are represented using functional, behavioral, and structural (FBS) engineering design framework [15].

Our study for building reasoning models includes HMM, LDA and simple statistical analysis models. Additionally, we present a HIT for NetMem integrating LDA and HMM for combining the advantages of both algorithms in efficiently learning patterns of big network data with high- and multi-dimensionality. We provide a semantic reasoning model for NetMem using the proposed HIT in order to have highly abstracted and associated network traffic semantics on different levels of granularity and related to various network concerns. We are motivated in our HIT design by the capability of LDA to discover latent and classified high-level features from multi-dimensional network data patterns with long-range semantics dependencies. Those classified features will be sequenced to enable semantic reasoning via HMM with higher accuracy. HMM are structured architectures that are able to predict sequences of semantic topics (related to different network concerns) based on input sequences of extracted network features by LDA. Depending on input sequences or pattern of highly-discriminative network data features, HMM with forward and backward algorithms can learn semantics efficiently showing their FBS aspects.

Some related work (e.g., [7,8,22,23]) adopted monolithic and hybrid techniques for enhancing networking operations such as intrusion detection and efficient routing. However, those works were application-specific and they did not provide customizable application-agnostic way for building ontology of concept classes. Also, unlike [24,25], NetMem provides scalable memory storage for network data with various levels of abstraction for data semantics matching and behavior predictions processes.

The main contributions in this paper are:

- Design of a human-inspired memory management system for big networks with efficient processes for extraction of classified high-level features and reasoning about rich semantics;
- Hybrid intelligence technique for efficient and effective reasoning about network semantics; and
- Comparative study of semantic reasoning techniques suitable for big networks, with guidelines highlighting their advantages and limitations.

The remainder of this paper is organized as follows. Section 2 presents our proposed methodology for semantics extraction from big data. Section 3 presents an overview of our human-memory inspired network memory system. Sections 4, 5 and 6 describe three different techniques for semantics reasoning. Section 7 describes our proposed HIT for semantics management. Section 8 highlights the differences between the proposed HIT and

Please cite this article as: B. Mokhtar, M. Eltoweissy, Big data and semantics management system for computer networks, Ad Hoc Networks (2016), http://dx.doi.org/10.1016/j.adhoc.2016.06.013