論 文 の 内 容 の 要 旨

論文題目　Unsupervised Structural Pattern Analyses for the Management of Network Traffic and Configurations
（ネットワークのトラフィックおよび構成のマネジメントにおける 教師なし構造パターン分析）

氏　　名　肥村　洋輔

Computer networking, particularly the Internet, has become an essential platform for human life and many aspects of industries. Such the successful growth of the computer networking, instead, insists upon the importance of its efficient management such as traffic management and configuration management. On the other hand, one of recent expectations towards efficient network management stems from the recent fashion of statistical data analytics such as Artificial Intelligence (AI), whereas one difficulty in applying statistical data analyses is derived from its low-level numerical nature differing from actual domain-specific data set. An insight is that the widespread of computer networking has been contributing to data of inter-similar objects (e.g., a number of hosts and/or a number of multiple virtual networks over the shared infrastructure), which can be compared each other to uncover knowledge (e.g., typical patterns and atypical anomalous patterns); this indicates the potential usefulness of unsupervised analyses (e.g., cluster analysis), which extract similar patterns inside data itself without established knowledge database pre-defined by human experts. However, orthodox unsupervised approaches based on extracting and clustering numerical values (i.e., feature vectors) face the difficulty in interpretation of resulting outputs, which is important in management domain in order to take appropriate actions against the outputs. In this research, we study unsupervised approaches with structural patterns (e.g., graph structure), which is more interpretable (e.g., by visualizations) than only using numerical features.

In Chapter 2, we present structural pattern analysis on network traffic data. This analysis is demonstrated with the traffic data obtained in a measurement point in the Internet. We particularly focus on end-host profiling by analyzing network traffic, which is a major stake in traffic engineering. The use of graphlet for end-host traffic analysis is efficient for interpreting host behaviors, which essentially consists of a visual representation as a graph. However, graphlet analyses face the issues of choosing between supervised and unsupervised approaches. The former can analyze a priori defined behaviors but is blind to undefined classes, while the latter can discover new behavioral patterns at the cost of difficult a posteriori interpretation. This work aims at bridging the gap between the two. First, to handle unknown classes, unsupervised clustering is originally revisited by extracting a set of graphlet-inspired attributes for each host. Second, to recover interpretability for each resulting cluster, a synoptic graphlet, defined as a visual graphlet obtained by mapping from a cluster, is newly developed. Comparisons against supervised graphlet-based, port-based, and payload-based classifiers with two datasets demonstrate the effectiveness of the unsupervised clustering of graphlets and the relevance of the a posteriori interpretation through synoptic graphlets. This development is further complemented by studying evolutionary tree of synoptic graphlets, which quantifies the growth of graphlets when increasing the number of inspected packets per host.

In Chapter 3, we present structural pattern analysis on network configuration data. This analysis is demonstrated with the configuration data in a multi-tenant datacenter network, where multiple customer (tenant) networks are virtualized over a single shared physical infrastructure. The use of multi-tenancy is cost-effective but poses significant costs on manual configuration. Such tasks would be alleviated with configuration templates, whereas a crucial difficulty stems from creating appropriate (i.e., reusable) ones. In this work, we propose a graph-based method of mining configurations of existing tenants to extract their recurrent patterns that would be used as reusable templates for upcoming tenants. The effectiveness of the proposed method is demonstrated with actual configuration files obtained from a business datacenter network.

In Chapter 4, we also present future perspectives for structural pattern analysis for the management of actual field domain. Contrary to the previous chapters (Chapters 2 and 3) that deal with traffic and configuration data in the computer networking domain, this chapter discusses the possibility of extending those analytical approaches to the field domains (e.g., building), which have become more important as more field devices become connected. We consider that the spirit of using data analytics for the management of end-hosts (by traffic analysis) and virtual network configurations (by configuration analysis) should have commonality to some extent for the management of field domains such as to manage end-devices (e.g., sensors and actuators) and their inter-link configurations. However, analytics on data obtained in field domains is relatively difficult due to the differences in the characteristics of those domains (e.g., open systems for field domains, rather than closed systems in computer networking domain), particularly leading to the difficulty in interpretation of analytical results. In addition, characteristics of data in computer networking fields are rather aggregated view of discrete and relative information (i.e., aggregated cardinality), which would be suitable for the discrete structural model (e.g., graphlet), whereas field data additionally suggests the importance of individual point-by-point time-series view with non-discrete and absolute values. We discuss the need for trial-and-error approach of analyzing field data (e.g., per-room temperature in a building), interpretation of which is gradually complemented by introducing multiple types of contexts (e.g., switch states of air-conditioning appliances in rooms) and even other data sources (e.g., external temperature outside the building) that are considered as structurally related. This approach is demonstrated with an unsupervised pattern analysis over an actual dataset about Electrical Heat Pump (EHP) equipped in a number of rooms in a building.