**Title:** Accurate Traffic Classification

**Authors:** Geza Szab6, Istvain Szab6, Dainiel Orincsay

**Location:** TrafficLab, Ericsson Research, Budapest, Hungary

1-4244-0992-6/07

2007 IEEE

**TRAFFIC CLASSIFICATION**

+ "The aim of traffic classification is to find out what types of applications are run by the end users, and what is the share of the traffic generated by the different applications in the total traffic mix, which one can observe at a given network interface."

+ Accuracy and completeness are metrics that characterize the performance of a traffic classification method.

**FLOWS**

+ "The communication between IP network nodes can be organized into flows..."

+ "A flow is a collection of IP packets sent from a given port at one IP address to a given port at another IP address using a specific protocol."

+ "A flow is identified by its five-tuple flow identifier: src IP addr, dst IP addr, src port, dst port, and protocol identifier."

**TRAFFIC CLASSIFICATION METHODS**

* **Classical traffic classification methods**: using port number & signature base
* **Connection pattern based method**: look at pattern generated by host, compare to behavior representing various activities/applications; relationship between the use of src and dst ports, relative cardinality of the sets of unique dst ports and IPs as well as the magnitude of these sets
* **Statistics based classification**: a statistical feature is grabbed and used to classify; this is combined with methods from AI, most common is the Bayesian analysis technique
* **Information theoretic approach**: relative uncertainty of five-tuple is calculated when one value is set to a fixed value, resulting to three values. These values are used a coordinates in a three-dimensional unit-cube.

**DRAWBACKS of the TRAFFIC CLASSIFICATION METHODS**

* Ports are dynamically allocated to applications that use the network. Also, traffic can be encrypted for security reasons, hence, making this method insufficient or vulnerable. [*For your thesis, indicate that the collected data are to be subjected to a simulation, hence the data are static.*]
* Some applications can be missed, or can produce false positives.
* It is difficult to identify specific patterns, thus accuracy suffers. Additionally, connection pattern based methods can only deal with broad classes of applications, not determine the specific application under the same group.
* Statistics based method that are used in combination with AI techniques such as Bayesian networks require an ample amount of previous data, labeled as “training data” in order to classify the rest of the trace.
* Information theoretic approach is not well-suited for flow level traffic classification. It just aids traffic classification. Similar to the previous method, it is limited to classifying broad groups, not specific applications.