Linguistic Summarization of Network Traffic Flows

Federico Montesino Pouzols, Angel Barriga, Diego R. Lopez and Santiago Sánchez-Solano

Abstract— **We address, by means of fuzzy linguistic summaries, two related problems: summarizing network flow statistics and making these statistics human-readable. Two complementary summarization methods are developed. First, a fixed set of protoforms of interest is defined, and the ones with a higher truth value are shown to the user as simple on-line summaries. This first method is suitable for real-time monitoring. Then, an association rules mining process is carried out in order to find hidden relations in flow records. Both approaches are implemented in a tool capable of real-time and off-line processing of network flow records. Experimental results for a number of heterogeneous NetFlow records show the usefulness of linguistic summaries to both network practitioners and users.**

I. Introduction

With an increasing diversity of technologies, applications and traffic patterns, the analysis of network traffic flows is becoming more and more complex and a full understanding of all the relevant facts is now far beyond the practical possibilities of network operators, managers and planners.

Current network measurement systems are becoming highly sophisticated and produce huge amounts of measurement data and statistics. High-precision network measurement in current backbones implies the generation of tens of GBs of data per hour. The major objective of network measurement systems is to provide an understanding of how networks perform. However, the gap between network measurement systems and user comprehension is increasing.

There are many visualization tools for network measurements (see [1] for an extensive list) which are mostly based on plots and charts to evaluate statistical properties of time series, scaling properties and protocol behavior. The visualization and reporting tools employed nowadays provide reports made of tens of plots, graphs and tables. Thus, it is not easy for experts to extract simple summaries. Additionally, the complexity of reporting and monitoring tools is holding back the adoption by end users.

Federico Montesino Pouzols and Santiago Sánchez-Solano are with the Microelectronics Institute of Seville, CSIC, Scientific Research Council, Avda. Reina Mercedes s/n. Edif. CICA. E-41012 Seville, Spain (phone: +34-955-056-666; fax: +34-955-056-686; email: {fedemp,santiago}@imse.cnm.es).

Angel Barriga Barros is with the Department of Electronics and Electromagnetism of the University of Seville, E-41012, Spain (phone: +34-955-056-666; fax: +34-955-056-686; email: barriga@us.es).

Diego R. Lopez is at RedIRIS, the Spanish National Research and Education Network, Edificio Bronce, Pza. Manuel Gómez Moreno s/n. Planta 2. E-28020 Madrid, Spain (phone: +34-912-127-625; fax: +34-915-568-864; email: diego.lopez@rediris.es).

This work has been supported in part by projects TEC2005-04359/MIC from the Spanish Ministry of Education and Science and project TIC2006-635 from the Andalusian regional Government.

Many methods for analyzing Internet measurement data have been developed throughout the years. However, most of them are quantitative, suited for specific data types, and designed for a particular purpose. There is a lack of general-purpose tools for qualitative exploration and analysis of Internet measurements, which is a first step needed for hypothesis-driven discovery, analysis and validation [2].

In this context, it is becoming more and more necessary to extract *concise* summaries that should be several orders of magnitude smaller than the original measurement dataset and should express how the network performs in ideally no more than a few lines of *human-readable* text [3]. In this paper, we address the problem of summarizing network statistics into brief reports and making them human-readable by means of fuzzy linguistic summaries.

Linguistic summaries via fuzzy logic have been shown to be a simple, efficient and human consistent data mining means. Linguistic summaries as introduced by Yager [4], [5] and further developed by Kacprzyk and Yager [6] and Kacprzyk and Zadrożny [7], are linguistically quantified propositions (as "Most traffic flows have an average packet size small") with a degree of truth.

Section II outlines network statistics based on the NetFlow technology. Section III defines linguistic summaries as considered in this work. Section IV defines linguistic summaries of network flow records and describes two complementary ways of implementing them. Finally, section V shows experimental results for a set of benchmark NetFlow records.

II. Network Flow Statistics: NetFlow

Most network operations centres currently collect statistics on the performance of their infrastructure. These statistics are mainly based on the concept of flow, defined as a unidirectional sequence of packets between given source and destination end-points.

NetFlow [8], introduced by Cisco Systems in 1996 as a technology for route caching, is nowadays a de facto standard for passive measurement and monitoring in the Internet. NetFlow based measurement is used for performance analysis, application and user monitoring, traffic engineering, capacity planning, billing, peering agreement and security.

From the viewpoint of a router, a flow is made of a sequence of IP packets sharing the same values for a set of properties within a time interval: source and destination IP address, source and destination transport level port, transport (layer 3) protocol type, type of service and incoming interface (see figure 1). Thus, a flow in the sense of NetFlow is unidirectional. NetFlow records consider flow properties at the link, network (IP) and transport layers, i.e., no specific information from the application layer is included.

978-1-4244-1819-0/08/$25.00

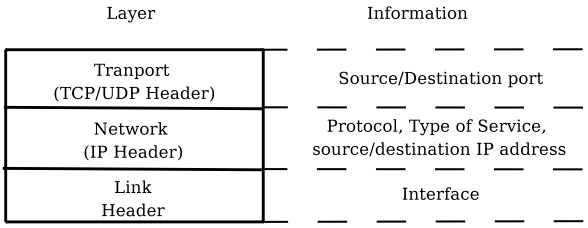

Fig. 1. Scheme of layers of information contained in Netflow traces

TABLE I
FLOW IDENTIFIERS AND SOME OF THE ATTRIBUTES CONSIDERED IN NETFLOW VERSIONS 5 AND LATER (IPFIX)

Attribute	Description
Source IP	IP layer address of sender
Destination IP	IP layer address of receiver
Packet count	Amount of packets transmitted
Byte count	Amount of bytes transmitted
Start time	Arrival time of first packet
End time	Arrival time of last packet
Input ifIndex	Index of input interface
Output ifIndex	Index of output interface
Type of service	IP header TOS field (Differentiated Services Code Point)
TCP Flags	Logical conjunction of activated flags
Protocol	Protocol code in the IP header
Next hop address	IP address of next router or host
Source AS number	Code of the source Autonomous System
Destination AS number	Code of the destination Autonomous System

When new flows are detected by a NetFlow capable router, a mapping between the flow and an outgoing interface is saved in memory. This way, next packets belonging to identified flows will not require to check routing tables, thus saving time and processing load.

This capability to identify flows can be applied to measure and characterize traffic traversing a router in real-time. Proper aggregation and summarization techniques allow for analyzing network performance.

The flow identifiers and some of the basic attributes considered in the most extended versions of NetFlow are listed in Table I. Additional attributes are available as extensions. Derived attributes are also defined from the measured attributes, such as the throughput or transfer rate (bytes/duration). Upon expiry of a flow, its statistics are accumulated and reported to a collector using the NetFlow protocol.

In order to standardize the NetFlow technology, the IP Flow Information Export (IPFIX) working group of the Internet Engineering Task Force (IETF) [9] is defining the IPFIX standard, which only differs from NetFlow in terminology and minor details as well as improvements to the flow transfer protocol but keeps the same principles, architecture, applicability and information model as NetFlow version 9.

Currently, the statistics reported by basic collector tools summarize flow records in the form of aggregation of counters and statistical descriptors such as percentiles. Analysis and visualization of flow collections is an active area of research [1] and many visualization techniques have been developed, particularly for topology analysis. However, the general summarization capabilities of available tools do not go beyond basic statistic descriptors, reports of top users and tables and plots of distribution functions, as for example the automatically generated Internet2 weekly reports [10] available online from http://netflow.internet2.edu/weekly/.

These tools are usually based on statistics and provide relevant reports. However, the reports can easily become human unreadable because of the huge amount of tables and graphs generated. Techniques and tools for extracting short yet meaningful reports are sought.

Simple aggregation capabilities have been introduced with recent versions of NetFlow as a simple method for summarization of flow records. It is possible for instance to request from a router flow records aggregated by autonomous system. This novel capability has been introduced as a response to the need of summarization mechanisms for preprocessing flow records Although these methods are powerful and reveal a great deal of useful information about how networks perform, the whole amount of available measurement data and most complex relations underlying them are still difficult to understand.

Recent versions of NetFlow also integrate sampling capabilities. With NetFlow sampling, only a percentage of traffic is accounted for measurement purposes. In order to restrain the load on network processors, a sampling of packets is performed typically on the 1-10% of the total traffic, while 90-99% of packets are not accounted for performance measurement purposes. Thus, some of flow statistics are affected by uncertainty. These sampling capabilities are extensively used in current measurement infrastructures.

III. LINGUISTIC SUMMARIES

Linguistic summaries as proposed by Yager [4] are a data mining technique for summarizing data collections using linguistically quantified propositions [11], such as "Most traffic flows are short lived". In this work, we consider the extended definition by Kacprzyk and Zadrożny [7], that leverages on the concept of protoform.

Linguistic summaries have a number of advantages when compared against classical statistical methods of summarization: they can summarize both numeric and non-numeric data, can provide many different summaries for specific purposes and have the ability to provide natural language summaries.

Linguistic summaries are obtained by means of a mining process on a usually large set of entities, by which a natural language expression summarizes essential facts about the set. In the sense of Yager [4], [5], a linguistic summary is defined as follows. Given:

- $\mathcal{D} = \{d_1, \ldots, d_N\}$, a set of entities that manifest some attributes, e.g., a set of traffic flows in a NetFlow collection.

- $\mathcal{A} = \{A_1, \ldots, A_M\}$ a set of attributes defined over the entities in the set $\mathcal{D}$, e.g., the set of attributes in a NetFlow collection, such as packet count, destination address, starting time, etc.

A basic linguistic summary is made of:

- A summarizer, $\mathcal{S}$, defined as a linguistic expression (or predicate) semantically represented by a fuzzy set, i.e., "short lived".
- A quantity in agreement or quantifier, $\mathcal{Q}$, defined as a linguistic quantifier that indicates the extent to which the entities satisfy the summary, e.g., "most".
- A measure of validity or quality of the summary. The basic validity criterion is the truth value of the summary, T, defined as a truth value of a linguistically quantified statement. The truth value can be computed using a number of methods, in particular Zadeh's fuzzy-logic-based calculus of linguistically quantified propositions [11] and Yager's OWA operators [12].

Fuzzy subsets are employed to represent the linguistic terms that specify a summarization $\mathcal{S}$ and a quantifier $\mathcal{Q}$. Thus, the truth value of both can be denoted by their respective membership functions, $\mu_{\mathcal{S}}(x)$ and $\mu_{\mathcal{F}}(x)$, being its universe of discourse that of one or more of the attributes in the set $\mathcal{A}$.

A summary ($\{\mathcal{S},\ \mathcal{Q}\}$) of a data set $\mathcal{D}$ with N elements from a measurement space $\mathcal{X}$ is usually written in generic form as "$\mathcal{Q}$ d's are $\mathcal{S}$", i.e., Q flows are S, as in the statement "most flows are long lived":

$$\{\mathcal{D}, \{\mathcal{Q}, \mathcal{S}\}\}, \text{ read as } \mathcal{Q}d_i \text{ are } \mathcal{S} \tag{1}$$

$\mathcal{S}$ is a then fuzzy subset of $\mathcal{D}$ and $\mathcal{Q}$ is a fuzzy set in the range $[0, 1]$. For instance, the membership function of the quantifier "most" can be defined as:

$$\mu_Q(x) = \begin{cases} 1, & \text{for } x \geq 0.85 \\ 2x - 0.7, & \text{for } 0.35 < x < 0.85 \\ 0, & \text{for } x \leq 0.35 \end{cases}$$

Then T is a truth value in $[0, 1]$ that can be computed from a summary as in equation 1 applying Zadeh's calculus:

$$T(\mathcal{D}, \{\mathcal{Q}, \mathcal{S}\}) = \mu_{\mathcal{Q}} \left(\frac{1}{n} \sum_{i=1}^{N} \mu_{\mathcal{S}}(d_i) \right)$$

The truth value of fuzzy linguistically quantified propositions is just a primary quality measure of summaries. Additional measures of the goodness of a linguistic summary, in terms of degree of interest, non-triviality or unexpectedness, are usually required in practice in order to select relevant summaries [6].

The kind of summarizer in equation 1 can be generalized to a compound summarizer form made of the conjunction of any number of linguistic expressions about the attributes of the entities in $\mathcal{D}$, as in "Most flows *are* long lived *and* have an average packet size small *and* are high throughput".

Extended linguistic summaries can be defined by adding a qualifier, $\mathcal{R}$, also a subset of $\mathcal{D}$, as "$\mathcal{Q}\,\mathcal{R}$ d's are $\mathcal{S}$", i.e., $\mathcal{Q}\,\mathcal{R}$ flows are $\mathcal{S}$, as in the statement "most flows at night are long lived":

$$\{\mathcal{D}, \{\mathcal{Q}, \mathcal{R}, \mathcal{S}\}\}, \text{ read as "} \mathcal{Q}\mathcal{R}d's \text{ are } \mathcal{S}'' \tag{2}$$

In the case of equation 2, the degree of truth of the summary can be determined by Zadeh's calculus as follows:

$$T(\mathcal{D}, \{\mathcal{Q}, \mathcal{R}, \mathcal{S}\} = \mu_{\mathcal{Q}} \left(\frac{\sum_{i=1}^{N} (\mu_{\mathcal{S}}(d_i) \wedge \mu_{\mathcal{R}}(di))}{\sum_{i=1}^{N} \mu_{\mathcal{R}}(d_i)} \right)$$

Extended linguistic summaries can be thought as fuzzy *if-then* rules where the antecedent is $\mathcal{R}$ and the consequent $\mathcal{S}$, stating that if $\mathcal{Q}$ entities (flows) satisfy $\mathcal{R}$ then they satisfy $\mathcal{S}$.

Linguistic summaries, whether extended or not, can be compound as well, as in "most high throughput flows are long lived and have a packet size medium". In this case, the universe of discourse of the summarizer is extended to that of a set of attributes.

Thus, linguistic summaries as considered here are essentially linguistically qualified propositions in the sense of Zadeh's calculus [11].

Protoforms of linguistic summaries are defined as abstracted prototypes and may form a hierarchy [13]. A classification of possible protoforms of linguistic summaries is developed in [7]. For instance, replacing $\mathcal{Q}$ with a concrete quantifier *Most* in equation 1, we obtain a particular kind of protoform: "Most flows are $\mathcal{S}$.

Another kind of protoform can be specified by fixing the attribute or attributes of interest for $\mathcal{S}$, as "$\mathcal{Q}$ flows are $\mathcal{S}^{A_c}$", where A_c is the attribute of interest. For instance, when one is interested in the duration of flows an appropriate protoform can be defined by restricting the summarizer to the linguistic labels defined for the duration attribute, "$\mathcal{Q}$ flows are $\mathcal{S}^{duration}$", where $\mathcal{S}^{duration}$ may take the form of any of the linguistic variables defined for the attribute duration.

IV. Definition of linguistic summaries of NetFlow Collections

We propose two methods for the linguistic summarization of NetFlow collections. Both are complementary to traditional methods of analysis and visualization of network flow statistics. To this end, a set of linguistic variables for flow attributes as well as a set of fuzzy quantifiers have been defined.

Some NetFlow attributes are modeled with crisp values (such as protocol, destination port and interface numbers), while some others are modeled using linguistic variables [14]. For the latter attributes, fuzzy sets are defined using domain specific terminology as shown in table II. For instance, "mice", "bulk" and "elephants" are terms usually employed in the Internet measurement literature to refer to recurrent kinds of traffic flows with regards to its place in flow size distributions [15], [10].

TABLE II
LINGUISTIC LABELS FOR SOME FLOW ATTRIBUTES

Attribute	Linguistic Labels
Duration	Short-lived, Long-lived
Average packet size	Small, Medium, Large, Jumbo
Throughput	Low, Medium, High
Bytes	Mice, Bulk, Elephants
Packets	Packet-Mouse, Packet-Bulk, Packet-elephant
Time (start, end)	Day, Night

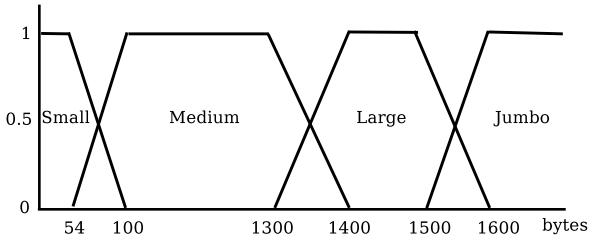

Fig. 2. Linguistic Labels for the attribute average packet size

TABLE III
BASIC PROTOFORMS FOR ON-LINE LINGUISTIC SUMMARIES OF NETFLOW

Concept	Summarizer, $\mathcal{S}$	Qualifier, $\mathcal{R}$
Throughput distribution	$A^{throughput}$	-
Duration distribution	$A^{duration}$	-
Transfer size distribution	A^{bytes}	-
Average packet size distribution	$A^{packetsize}$	-
Packets per flow distribution	$A^{packets}$	-
Throughput distribution qualified by transfer size	$A^{throughput}$	A^{bytes}
Bulk TCP transfer duration distribution	$A^{duration}$	TCP crisp filter and fuzzy qualifier "Bulk" over A^{bytes}

Derived metrics are also considered, as is the case of throughput, defined from the flow attributes as the ratio bytes/duration. As an example, the definition of the linguistic labels for the average packet size is shown in figure 2.

In practice, some attributes are crisp and its inclusion in summarizers and qualifiers of linguistic summaries can thus be modeled as filters that keep a subset of flows for certain crisp values or ranges of crisp values. For example, if the user is interested in summaries regarding only TCP flows, a first filtering step is carried out in order to account only those flows that correspond to TCP connections. This way, the summary "Most TCP flows are long-lived" differs from "Most flows are long-lived" in the set of flows to which they apply. Both are equivalent as for its evaluation as quantified proposition. Crisp attributes include the IP protocol field (IPv4, IPv6, ICMP, PIM, etc.), the transport layer protocol (TCP, UDP, SCTP, etc.) and transport port (HTTPS, SMTP, SSH, etc.) among others.

In principle, any quantifier could be considered as far as it is correctly defined. However, the concrete selection of quantifiers used to extract linguistic summaries has an impact on interpretability, informativeness and what kind of facts will be easier to find. The following sequence of quantifiers has been considered: "very few',' "few", "about 1/3", "about 1/2", "about 2/3", "most" and "almost all". The quantifiers "very few", "few", "most" and "almost all" denote different degrees of disparity [15] in the distribution of some property and are specially meant to find disparity conditions.

Once fuzzy quantifiers, qualifiers and summarizers are defined, linguistic summaries for flow records can be computed. When looking for the summaries that best describe flow records, two approaches can be considered: 1) the summarizer, the qualifier and the quantifier are given by the user, and 2) the three fuzzy sets are not fixed a priori and thus any possible combination must be considered. On the one hand, a tool that implements case a) would be of little value for users. On the other hand, an implementation of case b) would be very computationally intensive. However, applying the concept of protoforms [13], intermediate cases can be defined in between.

A. On-line Summarization of NetFlow Records

A first way of implementing linguistic summaries of NetFlow records is considered for on-line monitoring and generation of short reports. Since only one-pass algorithms are required to compute linguistic summaries, summaries within a bounded set can be generated in real-time.

For on-line summarization, a set of protoforms identified as conditions of interest are evaluated. Additionally, specific summaries specified as options to the tool are also evaluated.

In order to select a set of relevant protoforms our proposal combines ideas from reports found in traditional flow analysis and visualization tools, in particular from Internet2 weekly reports [10]. The basic set of protoforms considered for automatic on-line reports is shown in table III. Additional optional summaries have been defined for control, multicast and routing traffic.

B. Data Mining Summaries of NetFlow Records

Only a part of the potential of linguistic summaries is exploited using a set of fixed protoforms and protoforms specified by the user. If only known facts are considered when looking for informative summaries, more complex, unknown or unexpected relations are likely to be neglected. This issue can be addressed by means of automated data mining techniques. Particularly, hidden relations can be found in the form of fuzzy summaries using association rules mining algorithms.

Association rules are implications of the form $\mathcal{X} \rightarrow \mathcal{Y}$. With association rules mining algorithms, associations

TABLE IV
OVERALL COUNTERS OF THE ANALYZED NETWORK MEASUREMENTS

Name	Duration	Flows	Packets	Bytes
WIDE-F-1-Aug	15 min.	$2.84 \cdot 10^6$	$21.8 \cdot 10^6$	$15.3 \cdot 10^9$
CAIDA-OC48-0-Apr	1 hour	$18.1 \cdot 10^6$	$69.3 \cdot 10^6$	$35.2 \cdot 10^9$
CRAWDAD-Fall03	15 days	$5.05 \cdot 10^6$	$27.5 \cdot 10^6$	$16.8 \cdot 10^9$

between fuzzy sets [16] can be discovered, as proposed in [17]. From these rules, summaries as "$\mathcal{X}$ flows are $\mathcal{Y}$" can be identified, where the qualifier $\mathcal{R}$ is the condition ($\mathcal{X}$) of the rule and the summary $\mathcal{S}$ is the conclusion ($\mathcal{Y}$) of the rule.

Original association rules were defined for transactional data and binary valued attributes. An association rule has the following form: $A_1 \wedge A_2 \wedge \ldots \wedge A_n \rightarrow A_{n+1}$, and states that those items for which attributes $\{A_i\}, i \in \{1 \ldots n\}$ take value 1, will also take value 1 for attribute A_{n+1}. An equivalence between linguistic summaries and association rules can be considered if the summarizer and the qualifier are interpreted as the consequent and the antecedent of an association rule respectively. Then, the confidence of a rule can be interpreted as the combination of the linguistic quantifier and the truth value of the rule.

Two basic measures of the quality of an association rule are usually considered: the support and the confidence. The support is defined as the fraction of the number of items supporting the set of attributes $\{A_i\}, i \in \{1 \ldots n+1\}$ in the data collection. The confidence is defined as the fraction of the rows supporting $\{A_i\}, i \in \{1 \ldots n+1\}$ among all items supporting $\{A_i\}, i \in \{1 \ldots n\}$. While the support is a measure of the statistical significance of a rule, the confidence is a measure of its strength. The most interesting rules are those with a support higher than a minimal threshold and a high confidence.

Generalized association rules are redefined for fuzzy linguistic summaries mining as follows:

$A_1 \text{is} f_1 \wedge \ldots \wedge A_n \text{is} f_n \rightarrow A_{n+1} \text{is} f_{n+1} \wedge \ldots \wedge A_{n+m} \text{is} f_{n+m}$,

where f_i are fuzzy linguistic variables. A number of algorithms for association rules mining have been proposed. For the implementation described in the next section, the Apriori algorithm for fast discovery of association rules [18] is used.

V. EXPERIMENTAL RESULTS

An experimental tool for generating NetFlow linguistic summaries with the two approaches described, `flow-lsummary`, has been implemented. The tool allows for the definition of fuzzy linguistic variables and protoforms of interest for the on-line mode in a configuration file. Both the on-line mode and the data mining mode can be executed on NetFlow records in the format of the widespread flow-tools [19] suite, as well as in Cflowd format. IPv4, IPv6 and NetFlow versions 1, 5 and 9 are supported.

In order to assess the performance of the method implemented, we have generated linguistic summaries for a number of flow records. Some of them are generated from packet level captures and some other are actual NetFlow measurements. Overall properties of these records are shown in table IV. The following NetFlow records were analyzed:

- WIDE-F-1-Aug: network trace taken on August 1, 2007 at samplepoint-F of the WIDE backbone [20], a 155 Mbps trans-pacific link.
- CAIDA-OC48-0-Apr: trace 20030414-0 from an Internet backbone OC-48 link (2.4 Gbps capacity) from CAIDA [21],
- CRAWDAD-Fall03: Dartmouth/campus data set [22] from the Community Resource for Archiving Wireless Data (CRAWDAD), recorded at a wireless campus network covering 18 buildings.

Very few flows are high throughput	[0.997]
Most flows are short lived	[0.717]
Almost all flows are mice	[0.997]
Most flows have an average packet size medium	[1]
About 1/3 flows are packet elephants	[0.941]
Most bulk flows are medium throughput	[1]
Very few bulk TCP flows are long-lived	[0.977]

Fig. 3. Simple on-line linguistic summary of the WIDE-F-1-Aug NetFlow collection (truth values between brackets).

Most flows are low throughput	[1]
Almost all flows are short lived	[0.973]
Almost all flows are mice	[0.960]
About 1/2 flows have an average packet size small	[0.812]
About 2/3 flows are packet mice	[0.868]
Few mice flows are high throughput	[1]
Almost all Bulk TCP flows are short-lived	[0.999]

Fig. 4. Simple on-line linguistic summary of the CAIDA-OC48-0-Apr NetFlow collection (truth values between brackets).

A. Predefined set of Summaries

The method presented in this paper has been found to provide insightful and easy to read summaries of flow collections. Simple on-line summaries for the NetFlow collections analyzed (see table IV) are shown in figures 3, 4 and 5. In simple on-line mode, `flow-lsummary` shows one summary about each of the considered protoforms (see table III for the basic set), i.e., only the most relevant fact concerning each protoform is shown.

The user can ask for an unlimited number of summaries per protoform. For instance, in the case of the CAIDA-OC48-0-Apr record and the protoform for throughput distribution qualified by transfer size, the user would get the following full summary with truth values: almost all bulk flows are medium throughput [0.86], and few bulk flows are high throughput [0.89], and very few mice flows are medium throughput [0.86], and few mice flows are high throughput [1].

B. Mining Association Rules for Extracting Linguistic Summaries

We discuss a sample set of summaries identified using the Apriori algorithm for association rules mining. Though the amount of association rules found can be overwhelming, a few simple filtering rules can significantly reduce the number

Very few flows are high throughput	[0.999]
Most flows are short lived	[1]
Almost all flows are mice	[0.999]
Most flows have an average packet size medium	[1]
Most flows are packet mice	[0.914]
Almost all bulk flows are medium throughput	[1]
Almost all bulk TCP flows are short-lived	[1]

Fig. 5. Simple on-line linguistic summary of the CRAWDAD-Fall03 NetFlow collection (truth values between brackets).

of rules to analyze. In particular, we disregarded those rules with a low support or with a low confidence (truth) value. Many interesting rules were found for the NetFlow records analyzed. We list as examples a selection of them:

- "Most DNS request flows occur both during the day and at night, are mice and short lived", with confidence 0.970, in the WIDE-F-1-Aug collection.
- "Most flows at night are mice", with confidence 0.890, and "Most flows during the day are mice", with confidence 0.998 in the CAIDA-OC48-0-Apr collection.
- "Most SSH traffic occurs during the day, and consists of short lived mice flows", with confidence 0.892 in the CRAWDAD-Fall03 collection.

Linguistic summaries provide a novel method to describe qualitative relations in NetFlow collections using natural language. Thus, by using association rules mining to find relevant summaries we have a suitable method for addressing a problem related to flow analysis: finding invariants in traffic, what is known as one the major goals of Internet Science [15].

VI. Conclusions

We have addressed network traffic analysis at the flow level from the perspective of linguistic summaries. Two approaches for summarizing NetFlow collections have been developed: 1) on-line summarization via a predefined and configurable set of potential interesting protoforms, and 2) discovery of hidden relevant summaries by means of association rules mining.

A tool that implements both approaches has been developed. Experimental results for a set of benchmark NetFlow collections confirm linguistic summaries as an alternative look into network flow statistics useful for both network users and practitioners. The method presented is a novel technique to generate simple and human-interpretable reports, but also provides a promising technique for finding invariants in network traffic and advancing Internet Science. This can be seen as a first step towards natural language based knowledge discovery for Internet Science.

Acknowledgement

We acknowledge the MAWI Working Group from the Wide Integrated Distributed Environment (WIDE) project [20] for kindly providing their flow collections and support. We are also indebted to the Cooperative Association for Internet Data Analysis (CAIDA), for providing their OC48 data collection [21]. Support for CAIDA's OC48 Traces Dataset is provided by the National Science Foundation, the US Department of Homeland Security, DARPA, Digital Envoy, and CAIDA Members. We used the Dartmouth/campus data set [22] from the Community Resource for Archiving Wireless Data (CRAWDAD). Our work has benefited from the use of measurement data collected on the Abilene network as part of the Abilene Observatory Project (http://abilene.internet2.edu/observatory/).

References

[1] Cooperative Association for Internet Data Analysis, "CAIDA Visualization Tools," http://www.caida.org/tools/visualization/.

[2] J. Sommers, P. Barford, and W. Willinger, "SPLAT: A Visualization Tool for Mining Internet Measurements," in *7th Passive and Active Network Measurement Workshop*, Mar. 2006, pp. 31–40.

[3] C. Estan, S. Savage, and G. Varguese, "Automatically Inferring Patterns of Resource Consumption in Network Traffic," in *SIGCOMM 2003*, Karlsruhe, Germany, Aug. 2003, pp. 137–148.

[4] R. R. Yager, "A New Approach to the Summarization of Data," *Information Sciences*, vol. 28, pp. 69–86, 1982.

[5] ——, "Database Discovery Using Fuzzy Sets," *International Journal of Intelligent Systems*, vol. 11, 1996.

[6] J. Kacprzyk and R. R. Yager, "Linguistic Summaries of Data Using Fuzzy Logic," *International Journal of General Systems*, vol. 30, no. 2, pp. 133–1504, Jan. 2001.

[7] J. Kacprzyk and S. Zadrożny, "Linguistic database summaries and their protoforms: Towards natural language based knowledge discovery tools," *Information Sciences*, vol. 173, no. 4, Mar. 2005.

[8] "Cisco IOS NetFlow," http://www.cisco.com/en/US/products/ps6601/products_ios_protocol_group_home.html, Nov. 2007.

[9] B. Claise *et al.*, "Specification of the IPFIX Protocol for the Exchange of IP Traffic Flow Information," Internet Engineering Task Force, IPFIX Working Group, Revision 26, Sep. 2007, Internet Draft.

[10] S. Shalunov and B. Teitelbaum, "TCP Use and Performance on Internet2," in *ACM SIGCOMM Internet Measurement Workshop*, San Francisco, USA, 2001.

[11] L. A. Zadeh, "A Computational Approach to Fuzzy Quantifiers in Natural Languages," *Computers and Mathematics with Applications*, vol. 9, pp. 149–184, 1983.

[12] R. R. Yager, " On Ordered Weighted Averaging Operators in Multicriteria Decision Making," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 18, pp. 183–190, *** 1988.

[13] L. A. Zadeh, "A Prototype-Centered Approach to Adding Deduction Capability to Search Engines-the Concept of Protoform," in *First International IEEE Symposium on Intelligent Systems*, vol. 1, Sep. 2002, pp. 2–3.

[14] ——, "The concept of a linguistic variable and its application to approximate reasoning," *Information Sciences*, vol. 8, no. 3, pp. 199–249, 1975.

[15] A. Broido, Y. Hyun, R. Gao, and kc claffy, "Their Share: Diversity and Disparity in IP Traffic," in *5th Passive and Active Measurement Workshop (PAM)*, Antibes Juan-Les-Pins, France, 2004, pp. 113–125.

[16] M. Delgado, N. Marín, D. Sánchez, and M.-A. Vila, "Fuzzy Association Rules: General Model and Applications," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 2, pp. 214–225, Apr. 2003.

[17] J. Kacprzyk and S. Zadrożny, "Linguistic Summarization of Data Sets Using Association Rules," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, St. Louis, USA, May 2003, pp. 702 – 707.

[18] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo, *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, 1996, Fast Discovery of Association Rules, pp. 307–328.

[19] M. Fullmer *et al.*, "flow-tools," http://www.splintered.net/sw/flow-tools/, Nov. 2007.

[20] Widely Integrated Distributed Environment (WIDE) Project, MAWI Working Group, "Packet traces from wide backbone," http://tracer.csl.sony.co.jp/mawi/, 2006.

[21] CAIDA OC48 Trace Project, "CAIDA OC48 Traces 2003-04-24 (collection)," http://imdc.datcat.org/collection/1-0018-N=CAIDA+OC48+Traces.

[22] D. Kotz, T. Henderson, and I. Abyzov, "CRAWDAD data set dartmouth/campus (v. 2007-02-08)," Downloaded from http://crawdad.cs.dartmouth.edu/dartmouth/campus, Feb. 2007.