Classifying the Types of Autonomous Systems in the Internet

Xenofontas Dimitropoulos
Georgia Tech/CAIDA
fontas@ece.gatech.edu

Dmitri Krioukov
CAIDA
dima@caida.org

George Riley
Georgia Tech
riley@ece.gatech.edu

KC Claffy
CAIDA
kc@caida.org

1. INTRODUCTION

The AS-level topology of the Internet has attracted extensive research attention the last few years. Measuring the topology of the Internet, analyzing the properties of Internet topology graphs, and generating Internet-like synthetic graphs are prominent research topics in the field. Although the Internet AS topology has been studied extensively, less is known about the individual ASs, the entities that comprise aggregation units in the BGP routing system. In general, AS numbers are used by service providers, companies, universities and other organizations that connect to the Internet using BGP. However, the nature of the organizations that use AS numbers has not been systematically investigated yet. Statistical knowledge of the ASs in the Internet is essential not only to identifying the types of ASs that drive AS number exhaustion, but also to modeling the structure and evolution of Internet topology. In this work we initiate development of an Internet AS taxonomy by proposing an initial classification scheme based on empirically observed differences among AS characteristics.

2. MOTIVATION

Annotating the AS topology of the Internet with AS type information is a prerequisite for modeling the evolution of the Internet. Different types of ASs are associated with different growth patterns. For example, Internet Service Providers (ISP) try to grow by attracting new customers and by engaging in business agreements with other ISPs. On the other hand, small companies that connect to the Internet through one or few ISPs are not expected to grow significantly over time. Thus, categorizing different types of ASs in the Internet will help us to identify network evolution patterns and develop accurate evolution models.

Knowledge of AS types is also important for augmenting the AS topology with realistic intra-AS and inter-AS router-level topologies. For example, we expect that the network of a dual-homed university is vastly different from that of a dual-homed small company. The university will likely contain dozens of internal routers, thousands of hosts, and many other network elements (switches, servers, firewalls). On the other hand, the small company will most probably have a single router and a simple network topology. Even if the two networks are completely different, without knowing the AS type information we will fail to associate appropriate router level topologies with a given AS.

Finally, knowing the AS type information is also crucial for understanding how the logically abstracted AS graph reflects actual network entities and organizations in the Internet.

3. GOAL

In this work, we seek to classify the types of organizations that own AS numbers. To do so, we employ information retrieval techniques to analyze data from the Internet Routing Registries (IRR). We effectively classify 20,598 ASs in 8 categories.

4. METHODOLOGY

The IRRs constitute a distributed database in which ASs store information about their routing policies, IP prefixes, contact points, etc. Unfortunately for Internet topology analysis, IRRs frequently contain incomplete, and sometimes inaccurate, entries. We use information retrieval techniques to overcome this challenge. Among the wide range of information in the IRRs, we are interested in the *aut-num* attribute of the RPSL class *aut-num* (the class and the attribute have the same name). This attribute contains a short description or a name of the organization responsible for the AS number. For example, the following are entries for the *aut-num* attribute found in the IRRs: “Oak Ridge National Laboratory”, “Unisys Corporation”. The *aut-num* attribute does not have a standard representation. It usually consists of a short description as in the examples above, but in some cases it just contains an acronym, e.g., “AMOLF”, “SITA”.

To classify the organizations registered in the IRRs we built an expert system that uses standard Text Classification (TC) techniques. Expert systems are a class of TC approaches that require to knowledge-engineer classifiers on how to classify documents. Knowledge-engineering is performed by manually encoding expert knowledge into a set of rules used to categorize text documents. Rules are typically expressed in boolean expressions in Disjunctive Normal Form (DNF). A boolean expression in DNF is a disjunction (OR) of clauses, where every clause is a conjunction (AND) of one or more literals l_i,[1] e.g. $(l_1 \wedge l_2) \vee (l_3 \wedge l_2)$. The text under consideration is assigned to the appropriate category if the corresponding DNF expression is satisfied: **if** (*DNF expression*) **then** (*category*).

We develop an expert system that classifies organization description records T_i, which we extract from the IRRs, in categories C_j. We first perform *extraction indexing* on the set T_i. Extraction indexing is the simplest method for indexing articles, in which the index is composed of the words that appear most frequently in the article. We preprocess all records T_i, converting words into lower case and removing punctuation points. Then we construct the index by computing the highest frequency words and phrases. For the computed index we remove stop words, i.e., words with little semantic content, e.g., "the", "of", "and", etc. We find the following top frequency words (phrases): "inc", "network", "system", "center", "internet", "information", "autonomous", "autonomous system", and "services". We also observe that word (phrase) frequencies follow a Zipf distribution, similar to word frequencies in the English language [1].

Next, we examine the top entries of the obtained index and group together keywords (or key-phrases) of similar semantic content that characterize distinct AS types. For example, we group together the keywords "ixp", "exchange" and "nap" that describe Internet eXchange Points (IXP). We create 8 groups of keywords (or key-phrases) that correspond to the following AS categories:

C_1: ISPs.
C_2: IXPs.
C_3: Network Information Centers (NIC); organizations responsible for managing and allocating Internet resources, like IP addresses and AS numbers.
C_4: Universities, colleges, schools or research centers.
C_5: Military networks or military-related organizations.
C_6: Government or local administration networks.
C_7: Hospitals and health centers.
C_8: Companies that own AS number(s), but as opposed to members of C_1 do not provide Internet connectivity services.

We say that an AS is of type C_j if its record T_i contains one or more of the keywords associated with C_j. Note that a record T_i may belong to multiple categories C_j. For example, we classify the record "DoD network information center" in both of the categories C_3 and C_5.[2] Similarly, we classify the record "NETIS TELECOM Inc. Yaroslavl region ISP provider Russia" in both of the categories C_1 and C_8.[3] We then investigate if one the multiple categories is more representative of the AS type. To do so, we identify *strongly coupled keywords*, meaning keyword pairs of different categories that often appear in the same record. We examine records with strongly coupled keywords, determine if one of the multiple categories is more relevant to the AS and encode our preference for this category into our expert system using DNF expressions. In certain cases it is not possible to uniquely classify a record, since it inherently belongs to multiple categories. For instance, the record "UK Defence Research Agency" naturally belongs in both of the categories C_4 and C_5. However, for the cases that a certain category is clearly more appropriate, we introduce a DNF expression. For example, the record "DK Ministry of Education" is initially classified to both C_4, due to keyword "education", and C_6, due to keyword "ministry". Since the record refers to a government organization, we modify the initial classification rule: **if** (T_i contains "education") **then** (C_4) to: **if** ((T_i contains "education") **and not** (T_i contains "ministry")) **then** (C_4). This modification, enables us to uniquely classify all the records with both of the keywords "education" and "ministry". By examining strongly coupled keywords, we build an effective system of DNF expressions for use in classification.

Table 1: Number of ASs in each category.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
ASs	7,072	227	1,413	2,161	421	275	312	8,569
%	34.33	1.1	6.86	10.49	2.04	1.34	1.51	41.6

5. RESULTS

We implement the above algorithm and apply it to a set of organization description records we extract from the IRRs. After removing description records for private AS numbers, we are left with a set of 32,689 records. Among these, we effectively classify 20,598 ASs, which corresponds to 63.01% of the total number of records. Table 1 shows the per category classification statistics. Among the classified ASs, 41.6% are companies; 34.33% are ISPs; 10.49% are universities, colleges, school or research centers; 6.86% are NICs; 2.04% are military-related networks; 1.51% are hospitals and health centers; 1.34% are government and local administration related networks; and 1.1% are IXPs.

6. WORK IN PROGRESS

In this work, we introduce a simple and fundamental question: What kind of entities do Autonomous Systems represent? We provide a first answer using information retrieval techniques to analyze data in the IRRs. We develop a novel classification methodology and classify 20,598 of the Internet ASs into 8 representative categories.

Different types of ASs have different network topologies and infrastructures. Developing accurate topology generators requires the independent analysis of the network properties of these diverse networks. For each type of network, we must understand the requirements that frame its architectural design and evolution. Some requirements will be consistent across all ASs, others will vary as a function of size, geographic span, and local properties including regulatory and market circumstances. The challenge is daunting, but since ASs are the fundamental units of aggregation in the global Internet routing system, formalizing their structural properties is on the critical path between where we stand today and a deeper understanding of the structure of the Internet.

7. REFERENCES

[1] G. Zipf. *Selective studies and the principle of relative frequency in language.* 1932.

[1] A literal is a boolean variable x_i or its negation $\bar{x}_i$.

[2] The keyword "dod" stands for Department of Defense and is associated with military organizations. The key-phrase "network information center" is associated with NICs.

[3] The keywords "isp" and "provider" are associated with ISPs and the keyword "inc" with companies.