Sheldon Ruiz
ECE 478
Project 2 Report

Introduction

Problem Statement

The goal of this project was to analyze and categorize internet topology data at the Autonomous System (AS) level. All initial data was sourced through the Center for Applied Internet Data Analysis (CAIDA). Using CAIDA, data about the classification of AS's (content, transit/access, and enterprise) was used to create a distribution of AS classification.

Furthermore, data on AS relationships were analyzed. AS relationships consisted of links, both p2p (peer-to-peer AS) and p2c (provider-to-customer). These links were tracked to each AS in the CAIDA AS Relationships dataset, and the analysis consisted of AS node degrees, AS provider sets, and IP prefix association. AS relationship data was used to infer another AS classification distribution and compared to the original distribution provided by CAIDA. Additionally, AS node degrees were used to infer a list of Tier 1 ASes. The size of the Tier 1 AS set as well as the top 10 ASes, and their respective AS-to-organization mapping, were reported.

Lastly, the customer cone of every AS was analyzed. The top 15 ASes ranked by customer cone size and top 15 ASes ranked by customer cone in percentage of IP addresses reachable by p2c links.

Responsibilities

All coding and reporting was developed and written solely by Sheldon Ruiz

Methodology

2.1 AS Class Distribution

AS Classification data was downloaded from the CAIDA database (Link: *http://www.caida.org/data/as-classification/* (file name: 20150801.as2types.txt.gz)). The classification database was imported into Excel, and Excels scripting functions were used to count the number of instances of classification per ASes. A pie chart was then produced from this counted data.

2.2 Topology Inference Through AS Links

AS Relationship data was downloaded from the CAIDA database (Link: *http://www.caida.org/data/as-relationships/* (file name: 20170901.as-rel2.txt.bz2)).
A python script then read in the data and added all the ASes to a database with their corresponding p2p and p2c links as a dictionary data object. Furthermore, another CAIDA database on AS IP Prefixes was downloaded and parsed by the python script (Link: *http://www.caida. org/data/routing/routeviews-prefix2as.xml*). Using this database, each AS was given a recorded IP Prefix(es) when applicable. Any AS listed in the IP Prefix data sheet

that was not in the original AS Relationship database was added to the AS list with no connections. Next, each node was calculated its degree by summing the total number of p2p and p2c links corresponding to that node. A histogram of AS node degree distribution and IP space was produced from this data. Lastly, each node (AS) was assigned a classification based on its degree:

$$Class = \begin{cases} Enterprise, & for: degree \leq 2,\ p2p = 0,\ p2c = 0 \\ Content, & for: p2p \geq 1,\ p2c = 0 \\ Transit, & for:\ p2c \geq 1 \end{cases}$$

AS Classification through the above formulas was used to produce another pie chart to compare the classification data downloaded from the CAIDA database in section 2.1.

2.3 Inference of T1 ASes

Once each node contained a recorded degree, the data was sorted by AS with highest degree. After each node obtained a rank by degree, a greedy heuristic was executed to infer the T1 AS set. The heuristic is described as follows:

1. Initialize the T1 clique, $S = \{AS_1\}$, where AS_1 was the highest ranked AS
2. If AS_2 was connected to all other nodes in set S, add it to the set. if not, move to next highest ranked node

This set was reported through a table featuring the top 10 ASes in the set *S*. A CAIDA database was then downloaded containing links between organizations and ASes (Link: *http://www.caida.org/data/as-organizations/*). ASes were then mapped to the organization name that owned them.

2.4 Customer Cones and AS Rank

Using a Python dictionary of all ASes mapped to their corresponding p2p and p2c links, each node was submitted to a recursive DFS algorithm that mapped all of the children within an AS's p2c tree. The goal of the algorithm was to find all p2c children and sub children for a given node. Once the fully mapped tree for a given AS node was obtained, the size of the tree (# of nodes) was recorded. Next a Python set was used to record each IP prefix for every node in the tree. The number of unique IPs was calculated by iterating through the set of IP prefixes. An IP space for a prefix was calculated by subtracting the prefix length from 32, then raising 2 to the result. This was repeated for each AS. Finally, the number of ASes in the tree, the advertised IP prefixes, and the corresponding IP addresses were compared to the total number of ASes, IP prefixes, and IP addresses by a percentage, respectively. The top 15 ASes through customer cone in number of ASes and percentage of IP addresses were reported in two tables.

Results

2.1 AS Classification

This pie chart was developed in Excel through the data from CAIDA on AS Classification:

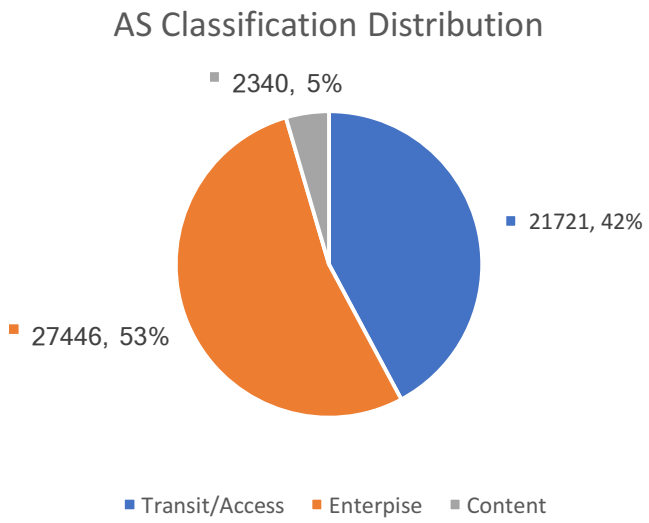

Figure 1: AS Classification Distribution CAIDA

The CAIDA data shows Enterprise ASes taking the majority of the AS space at 53%, with Transit ASes at 42%, and Content representing only a small portion (5%).

2.2 Topology Inference

This histogram plots the distribution of ASes by degree, that is – the sum of their p2p and p2c links:

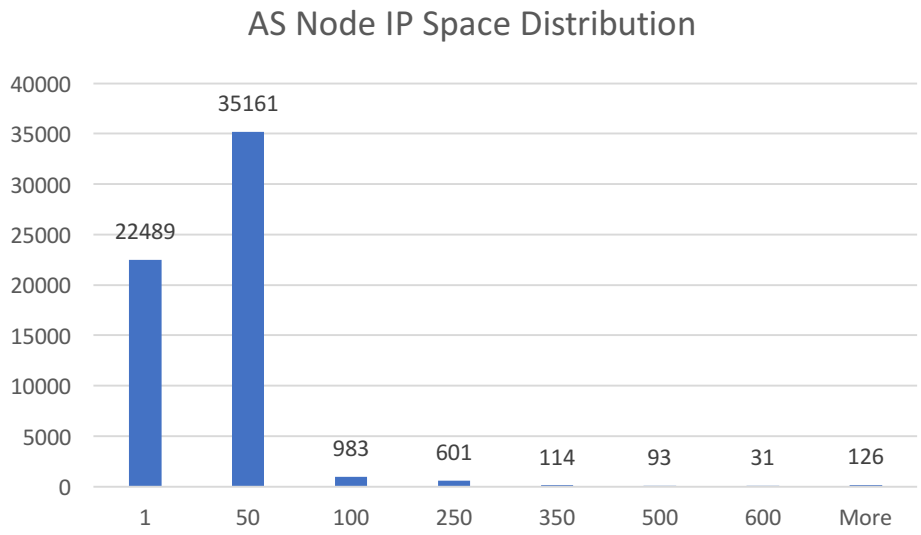

Figure 2: AS Node Degree by Inference

Here, the vast majority of all ASes fall into the '1' bin (51864 individual ASes). This means that most ASes contained at most only one p2p *or* p2c connection with other nodes. Though much smaller compared to the highest ranked bin, the next most featured bin was the one containing 100-200 degree nodes (3578 ASes). Only 40 ASes contained a degree higher than 1000.

Next, this histogram represents the distribution of ASes by IP Space (# of IP unique IP prefixes):

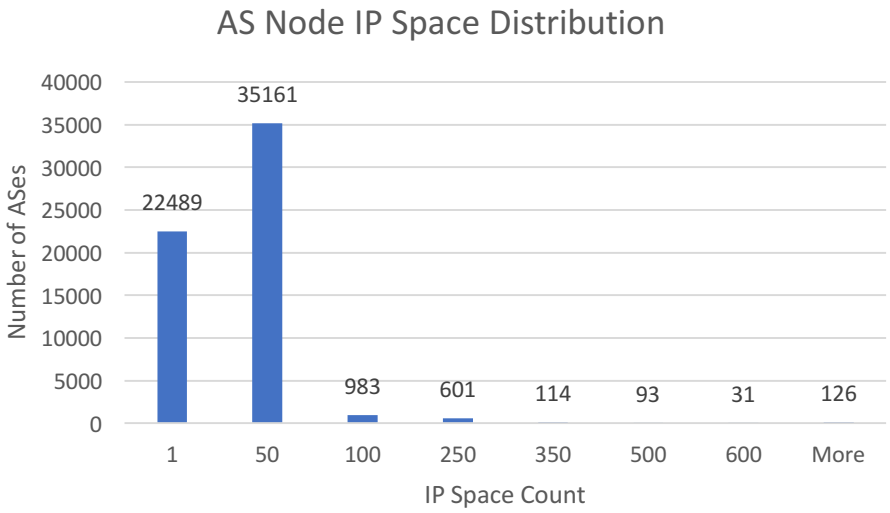

Figure 3: AS Distribution by IP Prefix Count

As seen by the chart, most ASes contained less than or equal to 50 IP prefixes. ASes with between 2 and 50 IP Prefixes were the most prominent (35161 ASes) followed by ASes with only one or less IP prefix (22489 ASes).

Additionally, the AS nodes were once again classified to Transit, Content, or Enterprise nodes. Here, this classification was inferred based on node degree. The definition of which was described in Methodology->2.2. This is the inferred distribution based on node degree:

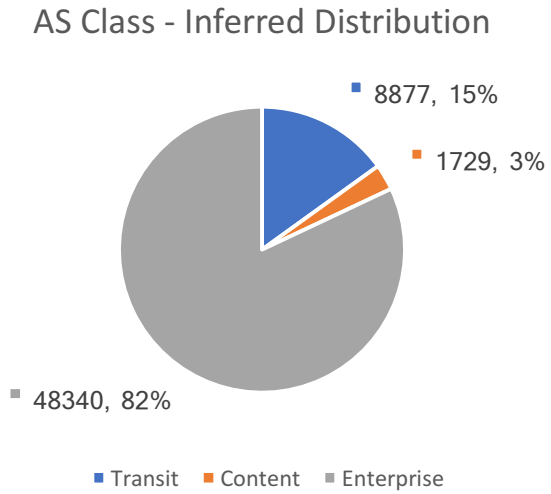

Figure 4: AS Classification by Inference on Node Degree

Compared to the original distribution, Enterprise ASes still take up the majority, but this majority has increased from 53% -> 82%. Inference places 48340 nodes as Enterprise compared to the original 27446 in the CAIDA database. Content nodes were roughly similar (2340 -> 1729), but Transit nodes lost most of its classifications to Enterprise (21721 -> 8877). The reason for this could be an error in the classification system. Or perhaps many ASes chose not to reveal themselves originally, but through p2c and p2p connections more Enterprise nodes were revealed and Transit nodes reclassified. Additionally, ASes that were inferred from the IP CAIDA datasheet could have inflated the Enterprise node count. Lastly, some nodes (654) remained unclassified by the sorting rules:

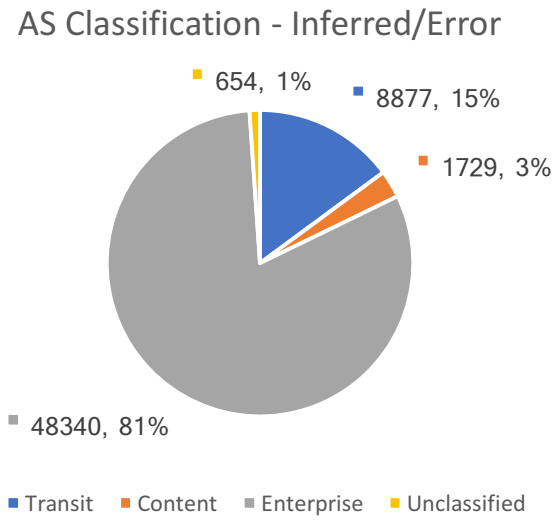

Figure 5: Inferred Classification Error

2.3 T1 Inference

After modifying the T1 search algorithm to continue looking for nodes even after the first node that failed to connect to all other T1 nodes, a total of 19 T1 ASes were found:

Rank	AS	Org Name
1	6939	Hurricane Electric, Inc.
2	174	Cogent Communications
3	3356	Level 3 Communications, Inc.
4	3549	Level 3 Communications, Inc.
5	7018	AT&T Services, Inc.
6	209	Qwest Communications Company, LLC
7	2914	NTT America, Inc.
8	6461	Zayo Bandwidth
9	1299	Telia Company AB
10	3257	GTT Communications Inc.
Total	19	

Table 1: T1 Inference Expansion

Above, the top 10 of the T1 ASes are displayed, along with their mapped organization ownership. 'Level 3 Communications, Inc.' owned ASes that took both rank 3 and 4 in the T1 inference, and 'Hurricane Electric, Inc.' took 1st with AS 6939. The Rank 1 T1 node, 6939, had a degree of 6512.

2.4 Customer Cones

Below are the top 15 ASes ranked by the customer cone in number of ASes:

			customer cone					
AS		**AS**	*number of*			*percentage of*		
Rank	**AS #**	**Degree**	ASes	IP Prefix	IPs	ASes	IP Prefix	IPs
1	174	5366	**50818**	628479	2937784809	85.27%	86.42%	74.44%
2	3356	4892	**50620**	630558	2981136111	84.94%	86.70%	75.54%
3	1299	1608	**48814**	625170	2971649329	81.91%	85.96%	75.30%
4	21320	59	**44556**	574511	2683960148	74.76%	78.99%	68.01%
5	20965	102	**44555**	574510	2683958100	74.76%	78.99%	68.01%
6	2914	1649	**43068**	571169	2597109129	72.26%	78.54%	65.81%
7	6453	719	**42132**	576812	2732957168	70.69%	79.31%	69.25%
8	3257	1566	**42120**	564206	2430293631	70.67%	77.58%	61.58%
9	5511	163	**40868**	551463	2526748796	68.57%	75.83%	64.03%
10	3491	602	**39810**	545847	2477218365	66.80%	75.05%	62.77%
11	6939	6512	**39196**	538341	2426787427	65.77%	74.02%	61.49%
12	209	1892	**37858**	513947	2248988254	63.52%	70.67%	56.99%
13	2603	712	**37846**	499934	2144365484	63.50%	68.74%	54.34%
14	701	1234	**37786**	518503	2451185565	63.40%	71.29%	62.11%
15	6762	412	**36560**	503843	2154181457	61.34%	69.28%	54.58%

Table 2: Customer Cones Ranked by # of ASes

The number of ASes in each customer cone row has been bolded for convenience. Interestingly, the highest ranked ASes in this category do not necessarily contain the highest degree. This is because node degree is also dependent on p2p connections, while customer cone is only dependent on p2c, in addition to considering all the children of each p2c connection in its tree. The highest ranking AS contained 50818 total ASes within its customer cone tree.

This is the table featuring the top 15 ranked ASes in customer cone by IP percentage:

			customer cone					
AS		**AS**	*number of*			*percentage of*		
Rank	**AS #**	**Degree**	ASes	IP Prefix	IPs	ASes	IP Prefix	IPs
1	3356	4892	50620	630558	2981136111	84.94%	86.70%	**75.54%**
2	1299	1608	48814	625170	2971649329	81.91%	85.96%	**75.30%**
3	174	5366	50818	628479	2937784809	85.27%	86.42%	**74.44%**
4	6453	719	42132	576812	2732957168	70.69%	79.31%	**69.25%**
5	21320	59	44556	574511	2683960148	74.76%	78.99%	**68.01%**
6	20965	102	44555	574510	2683958100	74.76%	78.99%	**68.01%**
7	2914	1649	43068	571169	2597109129	72.26%	78.54%	**65.81%**
8	5511	163	40868	551463	2526748796	68.57%	75.83%	**64.03%**
9	3491	602	39810	545847	2477218365	66.80%	75.05%	**62.77%**
10	701	1234	37786	518503	2451185565	63.40%	71.29%	**62.11%**
11	3257	1566	42120	564206	2430293631	70.67%	77.58%	**61.58%**
12	6939	6512	39196	538341	2426787427	65.77%	74.02%	**61.49%**
13	1239	451	35984	517197	2381796764	60.38%	71.11%	**60.35%**
14	209	1892	37858	513947	2248988254	63.52%	70.67%	**56.99%**
15	4766	524	34808	490139	2233018066	58.40%	67.39%	**56.58%**

Table 3: Customer Cone Ranking by IP Percentage

Though similar to Table 2, Table 3 shows some rearranging in the rankings comparatively. Interestingly, some ASes contained a greater percentage of IP Prefixes, yet still contained a larger percentage of the total IP space due to the IP Prefixes holding a smaller prefix length and thus allowing a larger IP space. The highest ranking AS was able to reach 75.54% of IPs within its customer cone using p2c links.