Latent Botnet Discovery Via Spam Clustering – Proof of Concept

Philip Tom

Overview

Unsolicited email, also known as spam, is sent in huge volumes everyday. Much of this spam is sent from botnets, networks of compromised computers. By examining a large collection of spam, it is possible to discover the latent relationships between the messages and infer the existence of botnets and their involvement with spam.

When examining a collection of spam, the same message can be observed multiple times from multiple sources, which often span multiple countries. This implies that either a massive, internationally distributed marketing firm is sending the messages or that the sources belong to the botnets. Since spam exists as an economy of scale, the latter is more probable.

The goal of this proof of concept is to identify the senders of the spam, and then cluster the messages based on similarity metrics. The resulting clusters should identify machines that are working together as part of a botnet.

Clustering Method

Messages are compared and clustered based on three simple assumptions:
1) All messages sent from the same IP address are part of the same cluster.
2) All messages with identical message bodies are part of the same cluster.
3) All messages with identical subject lines are part of the same cluster.

A cluster is seeded with a single message. Then, these three assumptions are repeatedly applied until the cluster ceases to grow.

Smarter similarity metrics can be used in the future. However, this paper aims to demonstrate that the inherent relationships between spam messages can be used to extract information about their senders. Thus, these simple metrics suffice for the scope of this paper.

Data

This proof of concept uses spam collected from hundreds of thousands of spam traps. The data set consisted of nine days of spam from the end of December 2007.

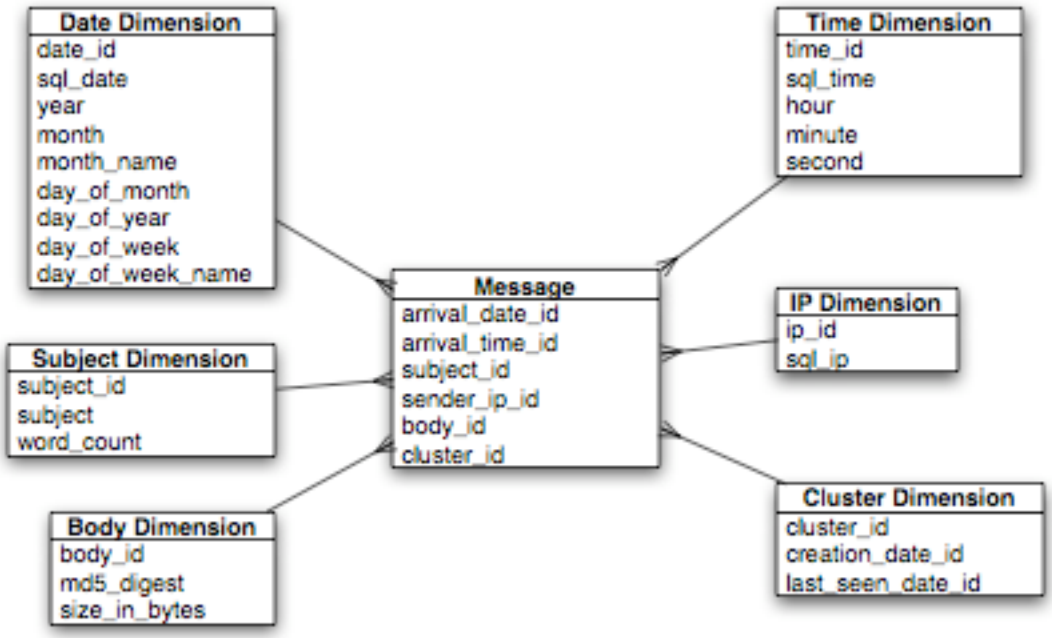
The sender of the spam is identified by its IP address. The originating IP address of the message is determined by examining the Received header lines. To avoid false positives, the data set is filtered to only messages with a single Received line. This line is generated by the MTA of the spam trap, and can thus be trusted. This reduces the data set to 1,731,227 messages, about 20% of its original size.

An MD5 digest of the message body is used as a simple metric to identify duplicate messages. Also, for simplicity, multi-part messages are excluded due to their unique part separators.

The full text of the subject is used for subject comparison.

Model

Dimensional modeling works remarkably well for exploring spam. The star schema creates a many to many relationship between all the message attributes. This allows for clustering based on arbitrary attributes.



Date Dimension
The date dimension allows queries to slice the data over the calendar. This table is pre-populated.

Time Dimension
The time dimension allows queries to slice the data over the 24-hour clock. This table is pre-populated.

Subject Dimension
This table describes message subjects. The Subject email header is extracted to populate this table. Currently, the table consists of the text of the subject and the number of words in the subject.

Body Dimension
The body dimension describes the body of the message. For this paper, the only attributes are the MD5 digest and the size in bytes of the body.

IP Address Dimension
The IP address dimension describes the origin of the message. Currently, only the address is examined. Later, other columns will be added, for example: CIDR, country, connection type, etc. These fields will be useful when analyzing the individual clusters.

Cluster Dimension
The cluster dimension is a placeholder for future analysis. Currently, the cluster dimension is only used to maintain cluster id integrity.

Message Fact
The message fact table contains a row for each message in the data source. The fields are entirely foreign keys to the dimension tables.

Clustering Queries

Clustering pseudocode:

```
for each cluster in clusters
  expand cluster

for each unclustered message in messages
  create cluster
  add message to cluster
  expand cluster
```

Cluster expansion pseudocode:

1) Grow cluster by IP address

```
update sdbf_message
   set cluster_id = ?
 where (cluster_id <> ? or cluster_id is null) and
       date_id <= ? and
       sender_ip_id in (select sender_ip_id
                          from sdbf_message
                         where cluster_id = ?)
```

2) Grow cluster by body

```
update sdbf_message
   set cluster_id = ?
 where (cluster_id <> ? or cluster_id is null) and
       date_id <= ? and
       body_id in (select body_id
                     from sdbf_message
                    where cluster_id = ?)
```

3) Grow cluster by subject

```
update sdbf_message m
   set cluster_id = ?
```

```
  from sdbd_subject s
 where (m.cluster_id <> ? or m.cluster_id is null) and
       m.date_id <= ? and
       m.subject_id in (select subject_id
                          from sdbf_message
                         where cluster_id = ?) and
       m.subject_id = s.subject_id and
       (s.word_count > 1 or length(subject) > 10)
```

Results

Applying the algorithm to the data set yields 30038 clusters. Only 16 clusters contain 100 or more ip addresses each. The following table shows the top clusters sorted by the number of IP addresses. The columns are the database cluster id, the number of messages, the number of unique IP addresses, the number of unique subjects, and the number of unique bodies.

cluster_id	messages	ips	subjects	bodies
1	1437287	325878	99919	331028
62	26623	1313	451	25992
59	11322	962	19	15
68	1065	609	2	1065
69	4476	514	59	85
10477	5521	283	5	9
953	722	275	149	333
175	310	209	2	309
379	240	184	7	9
18219	5581	153	15	5212
3924	2934	150	20	2934
144	377	125	22	377
242	307	124	4	3
134	3399	114	48	169
209	156	105	4	155
198	1117	101	174	1100

Cluster 1 contains the 85% of all of the messages in the test data set. Spam related to "rolex" watches accounts for half of this cluster. The rest is mostly related to gambling, adult content, and sexual enhancement.

Cluster 62 is the second largest cluster by IP count. The majority of the messages in the cluster are about credit and loans.

Clusters 59, 69, 134, 198, 242, 953, 10477 all contain Chinese spam.

Cluster 68, 175, 209 all contain delivery failure notifications from other MTAs. These clusters are a good indication that the clustering algorithm is accurate enough to distinguish between spam sent from botnets and delivery notifications sent from legit MTAs.

Cluster 18219 contains spam advertising marijuana.

Cluster 379 contains Japanese spam.

Cluster 3924 lacks a central strong theme.

Cluster 144 contains spam about stocks and working from home.

Some of the smaller clusters will coalesce with better similarity metrics. For example, cluster 25338 contains 824 messages sent from one ip address. There are 275 subjects and 275 bodies. The subjects look like:

```
Gift? Is not to late! Did you know where to purchase great watch?    bob
Gift? Is not to late! - Do you want Rolex? or any other extra watch?    jim
Gift? Is not to late!  -Gucci or  Louis Vuitton products    nancy
```

The varied text in the subject and body were enough evade the simple similarity metrics used in this algorithm. A query of the data set, shows that 3745 subjects started with "Gift? Is not to late!". 11224 messages match those subjects and comes from 129 IP addresses. These messages span 104 clusters. Hence, a better similarity metric could have identified these messages and coalesced the 104 clusters into a single cluster.

Conclusion

Clustering spam is a viable method of passively identifying botnets. Even with the simplest metrics, the algorithm can reveal the groups of IP addresses sending spam in unison. More sophisticated similarity metrics will yield more accurate clusters. In addition, applying the algorithm to the full data set will produce a more detailed picture of the origins of spam. The resulting clusters can be analyzed and cross-referenced with other data sources to produce new information about the botnets involved.