# 4. Metrics from dataset

In addition to the data in the given dataset additional metrics can be defined for insight in the spamming behaviour. The metrics defined in this chapter can be derived solely from the data in the dataset. When a metric can only be derived when additional external data is available this will be stated explicitly. The derived metrics are as follows:

1. Unique IP addresses controlled by botnet

Each botnet consists of a Command and Control node that controls all the bots in a botnet. The number of unique IP addresses under control of the Command and Control node signifies the efficiency at which a botnet can achieve the goal of e.g. spamming. It is to be noted that multiple devices can be connected behind a single IP address via Network Address Translation (NAT). The limitation of accuracy due to NAT is noted by for practical reasons ignored in this paper.

Metric interesting for: Internet Service Providers (ISP) to know what botnets are most prevalent to know where support is necessary most for cleaning customers. Also Internet security companies can use the metric to know what software developments add most value to products.

1. Top 10 country per botnet

Botnet signatures can be counted per country. Bots are not bound by geographical of national boundaries. Within each country operate a certain amount of bots that belong to a certain botnet which can be counted. For the top 5 biggest botnets a top 10 of countries is established to show what botnets are best represented in which country. The top 5 biggest botnets metric is established from the first metric “Unique IP addresses controlled by botnet”.

Metric interesting for: ISP’s can learn whether they overall have many infected customers and should have more controls in effect to counter customers getting infected. Internet security companies write software for the world, focussing on one country would be a less profitable endeavour. Governments can learn from his metric if advising the public on safe Internet use is effective relative to other countries.

1. Top 10 Internet Service Providers (ISP) that host botnets

Devices are connected to the Internet via an Internet Service provider (ISP), including bots and botnet Command & Control nodes. Certain Internet Service providers (ISP) could be more likely to host bots than others. This metric allows for the insight which ISPs host the most bots in a top 10 list.

Metric interesting for: ISP’s can learn whether they have many infected customers overall and should have more controls in effect to counter customers getting infected. Internet security companies can profit from this metric by offering their services to ISP’s. The government could use this metric to give incentive to ISP’s to clean up and support customers is keeping their devices clean.

1. Top 10 SPAM sending countries

Not all countries send the same amount of SPAM, nor considering SPAM sent per capita. There is no homogenous set of laws that is international applicable. Neither are digital criminality laws enforced with the same magnitude, which results in disparities in SPAM sent per country. A top 10 of sent SPAM per country gives a metric that can be used to decide which countries pose an additional security risk.

Metric interesting for: ISP’s could enforce stricter email policies for countries via a high SPAM rate. Security companies can use the metric to know what countries need expertise the most and offer expertise. Governments know how their country performs in protecting against SPAM relative to other countries.

1. Top 10 most active botnets and competition

Botnets send a certain amount of SPAM per timeframe denoting the activity of a botnet. Besides the size of a botnet sending SPAM, the activity (i.e. sent SPAM per timeframe) denotes the total amount of SPAM generated. This is another metric to show risk of a botnet.

Metric interesting for: ISP’s can use this metric (in combination with metric 3 for example) to learn whether they should act on cleaning customers. Internet security companies can use the metric to know where to focus software development. Governments know on what botnets to focus public awareness.

1. Botnet activity per country

Each country experiences a different amount of active bots that send SPAM. For each country separately the top 10 of most active botnets can be calculated. This metric is valuable to consider for a company in a certain country whether to invest in countermeasures against the top botnets.

Metric interesting for: ISP’s can merely learn an aggregated view of botnet activity. Internet Security companies can learn to what countries they can offer their expertise besides the top 10 in metric 5. Governments not showing in the top 10 of metric 5 can learn from this metric how well they relatively perform.

1. Number of countries active for the top 10 botnets

For the top 10 biggest botnets in IP count the number of different countries can be calculated. This metric allows for insight in how dispersed botnets are over different countries. Especially when SPAM emails contain links to infect more devices this metric could give insight in what countries a certain botnet is not effective in gaining more bots.

Metric interesting for: ISP’s can learn what botnet poses the biggest threat to customers. Internet security companies can learn on what botnet to focus software development of focus gaining knowledge. Governments know on what botnets to create public awareness and pose the biggest threat.

1. Choropleth map or heat map to show clustering of SPAM sending bots

A choropleth of heat map can show clusters of SPAM activity on a world map. For this visual metric to materialise the need for an IP-to-Coordinates database is necessary. The converted IP addresses into GPS coordinates allows for pinpointing an IP on a world map visually showing clustering of SPAM sending bots.

Metric interesting for: ISP’s could use this metric as a visual addition on the other metric to get a quick visual overview of SPAM origin distribution around the world. Moreover, Internet security companies and governments could use the maps for the same end as the ISP’s.

1. SPAM activity by timeframe

SPAM activity is not the same for a 24 hour period. The world population is awake at different times around the world. This metric gives insight in the amount of SPAM sent per hour in a 24 hour cycle (i.e. one earth day). If there is a difference in SPAM sent per hour this could mean that SPAM sending processes are not fully automated but require human intervention.

Metric interesting for: ISP’s can calculate for which timeframe maybe additional personnel is necessary for support, as the SPAM activity and thus infection rise. The same point can be made for Internet security companies and governments.

1. SPAM sent via Tor node

SPAM is in most cases sent directly from a bot’s IP address. However, it might be possible that certain bot’s use the Tor network to send SPAM. This metric gives insight in how much percent uses the Tor network to send SPAM to remain truly anonymous.

Metric interesting for: This metric is less interesting for ISP’s at their own as they can’t do anything against using Tor. However, Internet security companies and governments could use this metric to get insight in how SPAM is delivered to victims. The government could enact policies against Tor in collaboration with Internet Security companies and ISP’s.

There is one more advanced metrics to be considered, but require additional datasets to be generated. The eleventh metric could be geographically locating the Command & Control nodes of the botnets. By using IP address to GPS coordinate conversion the geographical locations and time stamps of sent SPAM could be used in combination with considering a spike in sent SPAM. During a spike different bots must have gotten a command from the botnet Command & Control (C&C) node to send SPAM. The latency between bots and botnet C&C could be used to roughly estimate where the botnet C&C is located in the world. Because this metric requires coupling multiple datasets together and could have a large inaccuracy due to rough estimation of latency and ignoring that packets could be routed in sub-optimal paths this metric is not pursued in this paper.

This chapter showed that at least ten metrics could be retrieved from the Spamhaus SPAM dataset. It is explained what value can be retrieved from the metrics and what and how actors could retrieve value from the metrics. Both quantitative numeric metrics can be produced as well as metrics, such as a heatmap, that appeal to getting a quick grasp of SPAM production.