# Technical Report: Predicting future attacks Data analysis of Dshield data set

## I. Introduction

In this report we analyzed 6 months of Dshield logs from May 2008 to October 2008 [1]. The goal of this analysis is to gain insight into spatial-temporal characteristics of malicious IP sources as observed by multiple "sensor" points (networks) over the Internet. We observed, analyzed and correlated malicious activities originated from hundreds of millions IPs to improve early detection and characterization of coordinated attacks (e.g. botnets). Ideally, we would like to able to filter out or counteract attacks at a very early stage or even before they begin.

## II. Data set description

### A. Data Format

In this report we analyzed 6 months of Dshield logs [1]. Dshield is repository of network security logs collected from over 600 different networks located all over the Internet. Every Dshield contributor (subscriber) submit the following informations every time an alert is raised by its network intrusion and detection system (NIDS) [2]:

| |
|---|
| time |
| contributorID |
| src IP |
| src port |
| dst IP |
| dst port |
| protocolID |
| flags (optional) |

TABLE I
Dshield format [2]

where, `time` is the time stamp when the alert was raised, `contributorID` is a unique identifier for the contributing network, `srcIP` and `dstIP` denote respectively the source and the target (destination) IP address[1], `src port` and `dst port` represent the source and the target port respectively, `protocolID` indicates the protocol used (when this information is available) and, finally, `flags` specifies the TCP flags (when available).

Network security logs submitted to Dshield are independently compiled at each contributing network. The Dshield data set offers a broad view of malicious traffic on the Internet as detected by hundreds of different networks. In this sense, it is a richer data set than having only the malicious course seen from a single IP prefix. However, since there is no information on the specific reasons alerts were raised, a data set of this nature is also inherently subject to errors and noise, for instance, due to NIDS misconfigurations and false alerts. In the extreme case, a malicious user could potentially subscribe to Dshield and submit fake reports with the only purpose of poisoning the entire database (poisoning attack). As observed in [1], the amount of noisy data can be considerably reduced by appropriately pre-processing the logs. However, even this pre-processing step can not completely eliminate all erroneous entries. As a consequence, we observe that, any algorithm that uses this type of data set must be designed to be robust to noisy data.

### B. Data set overview

Fig.1 and Fig.1 show the number of logs analyzed, source IPs, destination IPs, and contributor IDs per every day. In the time period considered there was a stream of 10 to 20 millions logs per day, contributed by about 600 different networks. We observe that, every day, there are about 200,000 different destination IPs and about 800,000 different source IPs.

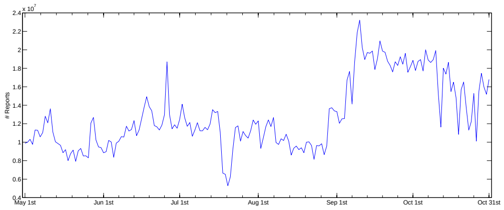

Fig. 1. Number of Dshield logs per day

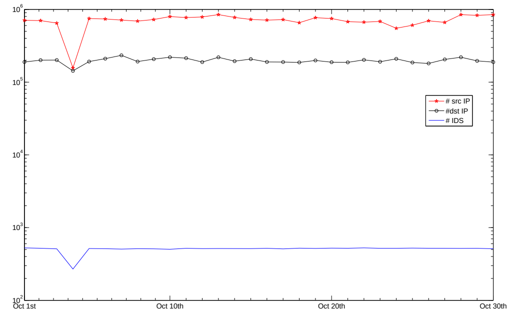

Fig. 2. Number of unique: source IPs, destination (target) IPs and contributors per day

[1] we note that whenever a contributing network does not want to reveal the destination IP, it can partially or fully obfuscate it as described in [2]

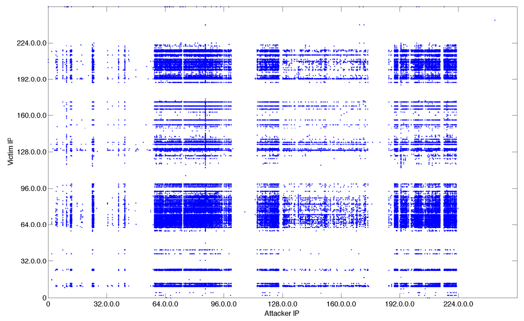

Fig. 3. Pairs (source IP, destination IP) for a single day day

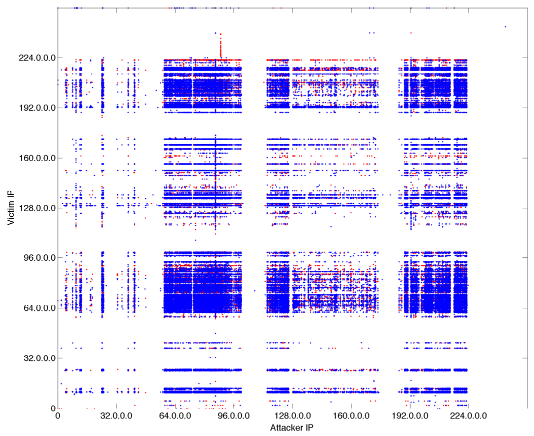

Fig. 4. Pairs (source IP, destination IP) for two different days

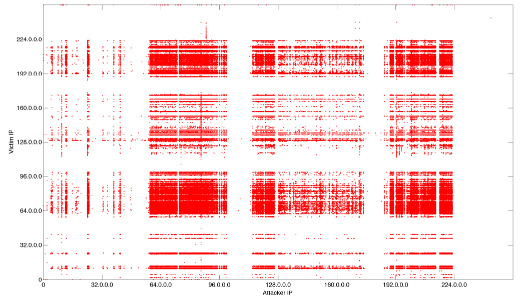

Fig. 5. Pairs (source IP, destination IP) for seven aggregated days

Fig.3, Fig.4 and Fig.5 show the distribution of pairs, (source IP, destination IP) over different time periods: one day, two different days and 7 aggregated days, respectively. All figures show some common patterns: there are both horizontal and vertical gaps, as well as horizontal and vertical areas with very high density. Gaps are mainly due to the following reasons: i) the set of Dshield contributors represents just a sampling of the Internet malicious activities: we do not have complete information; ii) some IP ranges are reserved/assigned but have very little usage, while some other are still unassigned. For instance, we can clearly see gaps corresponding to prefixes 224.0.0.0/3, 176.0.0.0/5, 100.0.0.0/6, 104.0.0.0/6, and so on, which are known to be unassigned IP prefixes [5]; we also observe that the first 64 class A networks, which are mainly assigned to US military, governmental organizations, large IT companies, appear to have fewer IP sources of malicious traffic than the rest of the IP space.

In Fig.4, we compare the location of pairs, (source IP, destination IP), for two different days. We observe that, at a large granularity, we have two very similar configurations. However, at finer granularities the two set of points differ substantially. In some cases, new IPs are within the same source subnet are a previously-seen malicious IP; this is likely due to the use of DHCP, which would allow the same machine to appear at different times with different IPs. In other cases, new IPs belong to previously-unseen network subsets. This can be explained considering the large number of compromised machines controlled by criminal groups. This gives them the luxury of swapping the use of different subsets of the botnet to elude traditional defense mechanism based on blacklists of single IPs.

We also derive the country location of both source and destination IPs: in our dataset the majority of destination IPs are located in Thailand (43.6%), US (40.5%), EU (9.9%). Source IPs are mainly are located in: Thailand (41.3%), US (17.2%), and China (13.1%).

### C. *Temporal analysis*

We studied the temporal dynamics of malicious source IPs. The main findings can be summarized as:

- IPs that attack multiple times do it within few minutes. In fact, we observe that either an IP attack only once or, if it attacks multiple times, consecutive attacks are very likely to happen within 3 to 10 minutes from one other. Fig.6 shows the CDF of consecutive attack time from the same source IP. About 90% of consecutive attacks happen within 3 minutes from each other. While a little percentage, 5%, are separated by several hours.

  This observation is straighten when looking at subnets behavior. Fig.7 shows that 95% of consecutive attacks happen from the same source subnet happen within 4 to 5 minutes.
- Single IPs are usually not active for more than 1-3 consecutive days .

  Fig.8 represents the histogram of consecutive blacklisted period for single IPs for the month of October. We make two main observation: i) as aforementioned, the large majority of IPs are active only 1 to 3 days; ii) a small fraction of IPs are continuously active for several days, up to the entire period of observation. This is particularly interesting also because, we will see, there is a direct correlation between the active time, and the number of attacks sent.

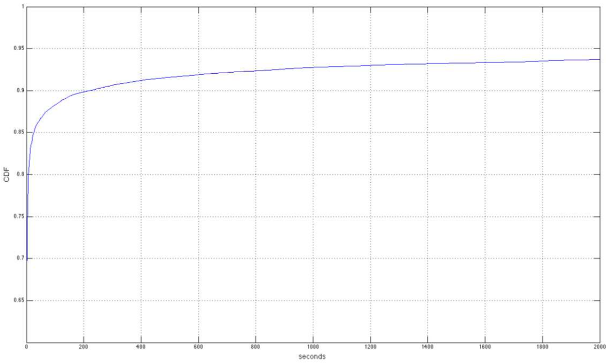

Fig. 6. Interarrival time of attacks from the same source IP

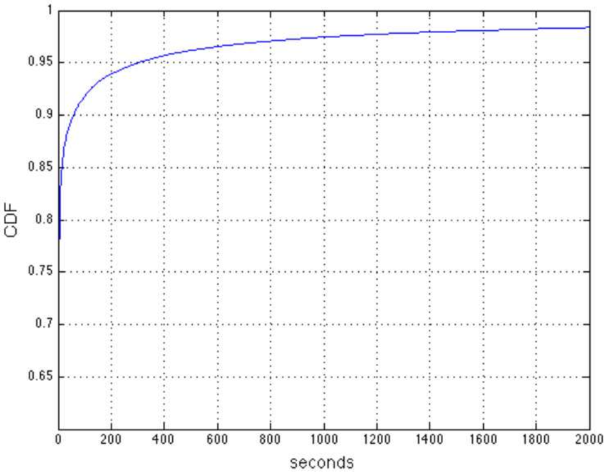

Fig. 7. Interarrival time of attacks from the same source subnet

Fig.9 analyzes the number of IPs that are reported as malicious both day $x$ and on day $x + \Delta t$, where $\Delta t$ is measured in days. The number of IPs that are reported in two consecutive days is about $13\%$. This is a quite small number and it decreases rapidly to $8-6\%$ in 3 days. However, as observed in the previous figure, there is a small fraction of IPs, $4\%$, that is continuously reported as malicious.

### D. Graph

In this section we interpret our data set as a graph, in which nodes are IPs and there is an edge from node $n_1$ to node $n_2$, if and only if, there is a log in the data set that has $n_1$ as source IP and $n_2$ as destination IP. In this section we study some basic properties of this graph.

Per every single day, at the IP granularity, the graph is well approximated by a bi-partite graph. There are, on average, only 221 out of 800,000 different IPs that, in the same day, appear both as source and as destination IP in the Dshield logs. Aggregating all days in October, the number of IPs that are both source and destination is still quite limited: 3957 IPs are both source and destination in October, out of 14,000,000 different source IPs.

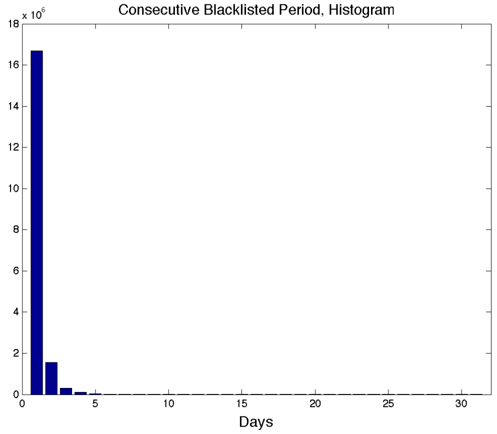

Fig. 8. Consecutive blacklisted time

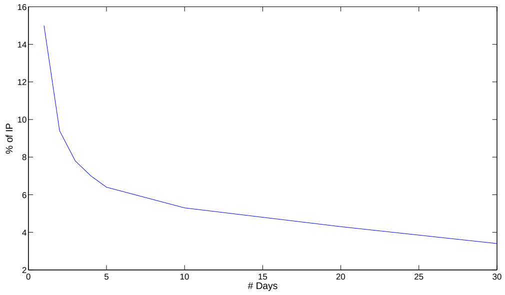

Fig. 9. Percentage of IP blacklisted both in day X and at day X+ΔT

Surprisingly, at the subnet level the situation is much different: 3,805 subnets appear in October both as source and as destination, out of 126,949 source subnets, and 4,070 destination subnets. This implies that, in about one month, almost all destination subnets eventually generate traffic that is classified as malicious. This might be due, for instance, to worms that successfully infected a contributing network and from there, try to keep on spreading to a different network on the Internet.

*1) Fan out:* In this section we study the distribution of out-degrees of nodes in the graph, i.e. the number of destinations attacked by single IP/subnet.

We observed that, the large majority of IPs attack only one destination (Fig.10). This mean that, in order to have an accurate prediction on future attacks, not only is fundamental to explore the correlation between attacks on different destinations but we must also account for the past history of attacks of every contributor individually.

Contrary to individual IPs, subnets are more likely to attack multiple contributors (Fig.11. This is probably due to one of the following reasons: hosts within the same subnet might be infected with different malicious code (e.g. they belong to different botnets), or they might be controlled by the same entity which split its resource to target multiple destinations, finally it is possible that the same machine target different network in different days using a different IP source at each time due to DHCP re-assignement. It is possible that analyzing the time of the attacks will help us to discriminate between these cases.

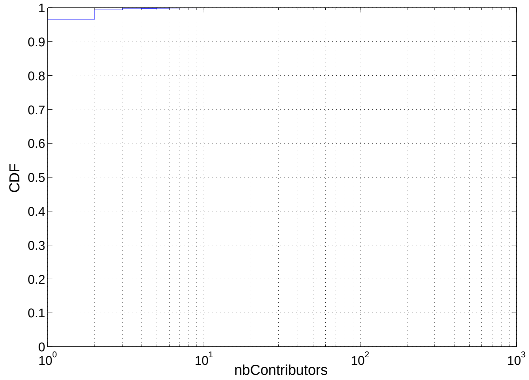

Fig. 10. CDF of the number of contributor attacked by a single IP

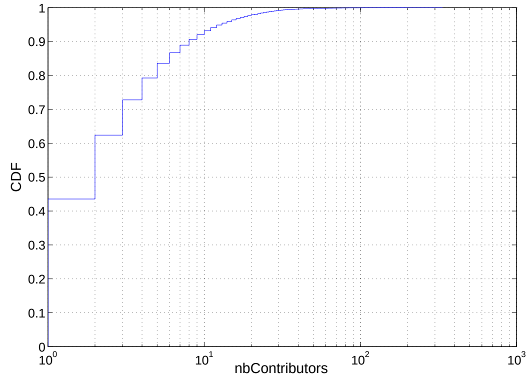

Fig. 11. CDF of the number of contributor attacked by a single subnet

*2) Fan in:* Fig.12 shows that about 50% of contributors provides 100-1000 attack reports. 10% of contributors have less then 10 attacks, while the largest contributor reports 500.000 attacks per day. On the one hand, there are two very large contributors which account for 60% of all data set; on the other hand, there few contributors that report very few attacks. This can be caused both by the attackers' behavior (e.g. depending on the type of services/software ran a target network can be more or less attractive from an attacker's perspective), and on the contributor's behavior: e.g. a misconfigured firewall can raise a significantly larger number of alarms than the actual attacks.

### E. Common Attackers

In this section we go in further details analyzing the correlation between attacks. In particular, we study the amount of *common attackers*, i.e. IP prefixes that attack a group of contributing networks.

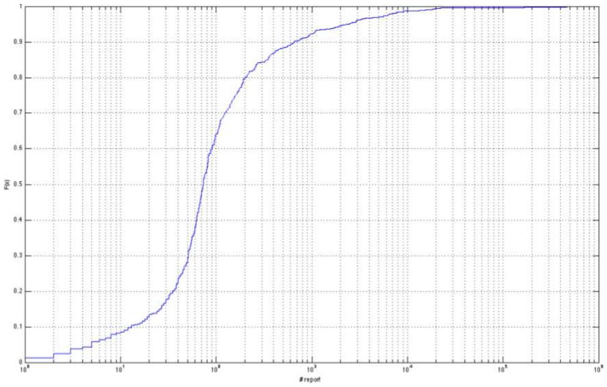

Fig. 12. CDF of the number of attackers per single contributor

Fig.13 shows how the average percentage of common attackers between two contributors (i.e. the number of IPs that attack both contributors) varies when considering only the contributors that provide a number of logs greater or equal to a certain threshold. When all contributors are considered, the average percentage of common attackers is about 62%. However, this is due to small contributors that share a large percentage of their attackers with other contributing networks. When we consider only contributors that provide at least 100 attacks logs, the average percentage of common attackers drops to 44%. This is still a very high number which shows that coordinated attacks by the same sources to the different destination is not a negligible phenomenon.

Fig.14 highlights the fact the when there are shared attacks, those also happens at about the same time. In order words, when a source IP attacks multiple destinations it is likely to attack them at about the same time. A possible scenario that explains these findings is the one in which a botmaster decides to attacks several networks. In this case, all bots coordinates to attack, probably in an order decided by an hit list, the various targets. This also confirm previous findings [6].

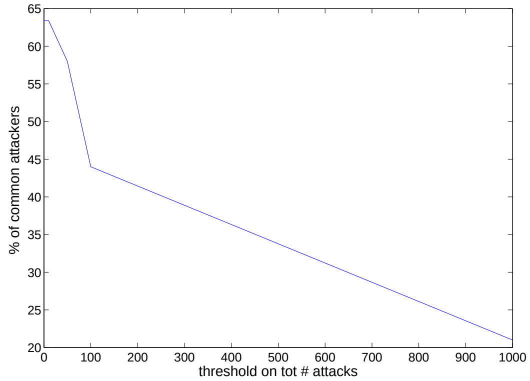

Fig. 13. Common attackers vs Threshold on total number of attacks

In Fig.15 and Fig.16 show the average number of *neighbors*, i.e. networks that shares common attackers with another

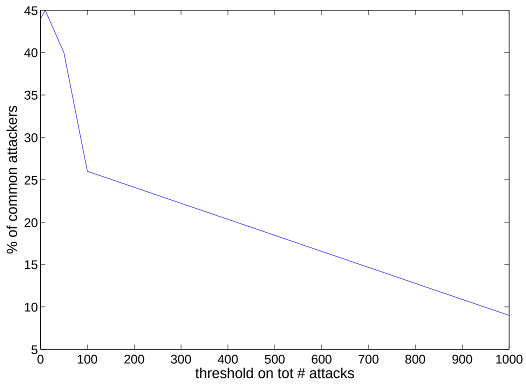

Fig. 14. Common attackers (within 10 minutes) vs Threshold on total number of attacks

network, for one and five aggregated days, respectively. Fig.15 for instance, show that, for a single day, we can cluster contributors in clusters of size 6 considering as neighbors pairs of contributors that share at least 40 common source IPs per day.

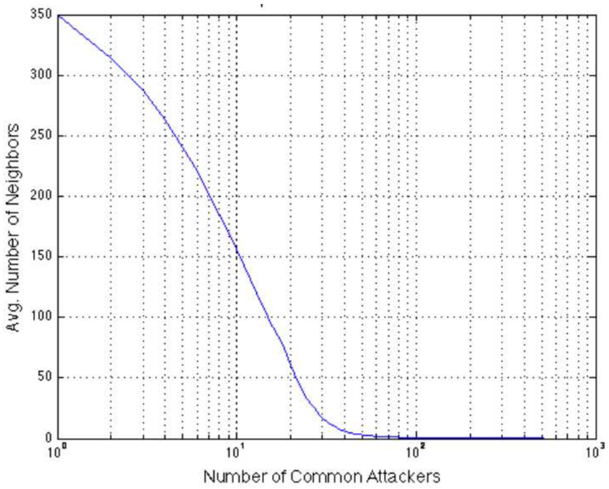

Fig. 15. Average number of neighbors vs Number of common attackers (1 day)

In Fig.17 and Fig.18 we analyzed the common attackers between each contributing network and its *closest* neighbor, i.e. the contributor with whom it shares the largest number of common attackers. Most networks shares 20-60 common attackers (single IPs); however, we also observe both extremes cases: networks with less then 5 shared source IPs and networks with thousands of shared attackers.

### *F. Correlation between attacks volume, duration and fan-out.*

In these section we analyze the correlation between the attack volume, duration, and fan-out of the attackers.

Fig.19 shows an interesting correlation between the fan out of attackers and the duration of the attacks: IPs that attacks one or few contributors tend to attack for a short time, while IPs attacking several contributors tend in general to produce attacks of longer duration.

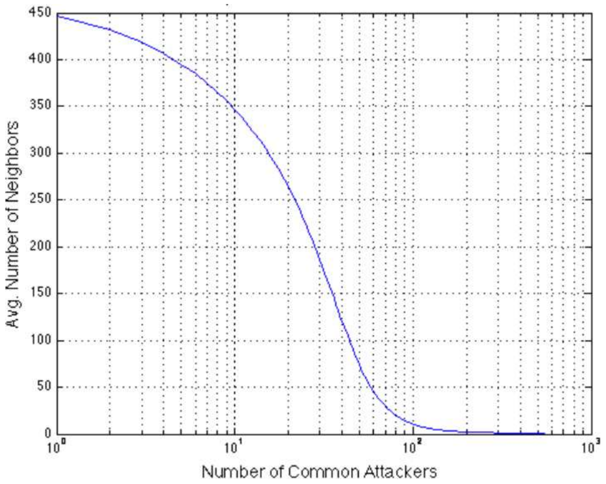

Fig. 16. Average number of neighbors vs number of common attackers (5 aggregate days)

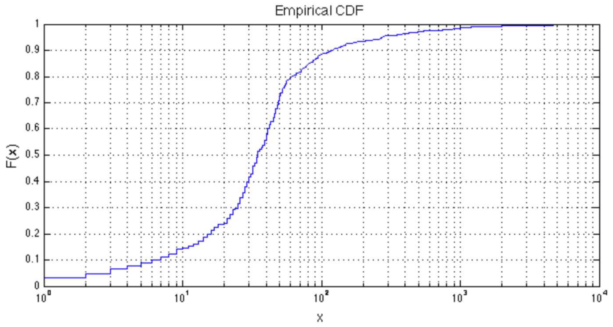

Fig. 17. CDF of common attackers with the closest neighbor (1 day)

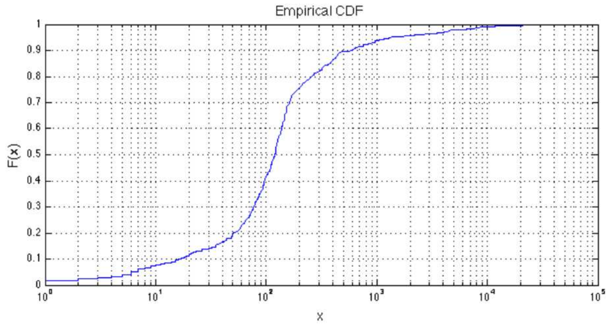

Fig. 18. CDF of common attackers with the closest neighbor (5 aggregate days)

A similar trend can be observed when correlating the duration of the attack with its volume, Fig.20.

A less intuitive figure is the one relating the fan-out of attackers with its volume. While intuitively one could expect these two metric to be simply directly proportional to each other, when analyzing the data we actually observe all kind of different behaviors, Fig.21 and Fig.22

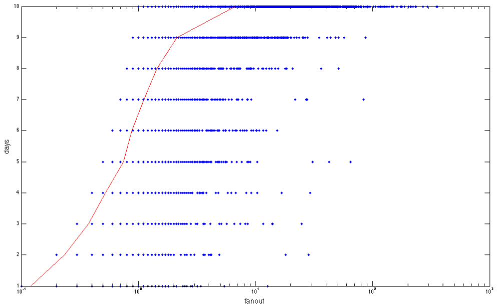

Fig. 19. Fan-out vs duration (days)

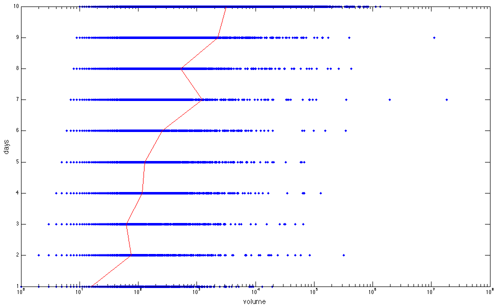

Fig. 20. Volume vs duration (days)

Fig.21 vaguely suggest a linear trend between volume and fan-out. However, a more careful inspection of the data highlight that several different behaviors, Fig.22: for many networks there seems to be a quadratic dependence of the volume on the fan-out; but we also observe different kind of extreme behavior: attackers with high volume focuses on few, or one, contributors as well as attackers with low volume but large fan out (low frequency attacks). It is possible that separating these data, for instance according to the the type of attack, will help us to see these different behaviors clustered in few categories.

In conclusion, is it possible to spot attackers with large fan out, large volume, and large duration. However, there are set of attacker that have only two, or one, of thee properties.

## References

[1] "Dshield: Cooperative Network Security Repository," *http://www.Dshield.org*

[2] Dshield data format *http://www.Dshield.org/specs.html#Dshield_format*.

[3] J. Zhang, P. Porras, J. Ullrich,"Highly Predictive Blacklisting", *in Proc. of Usenix Security 2008*, (best paper award), *http://www.cyberta.org/releases/HPB/*.

[4] A.Toscher, M. Jahrer, R.Legenstain, "Improved Neighborhood-based algorithms for large scale recommender systems", *in Proc. KDD'08*, Las Vegas, Nevada, USA, August 24-27, 2008

[5] `http://www.iana.org/assignments/ipv4-address-space`

[6] Sachin Katti, Balachander Krishnamurthy, and Dina Katabi, "Collaborating Against Common Enemies", *in Proc. ACM IMC '05*

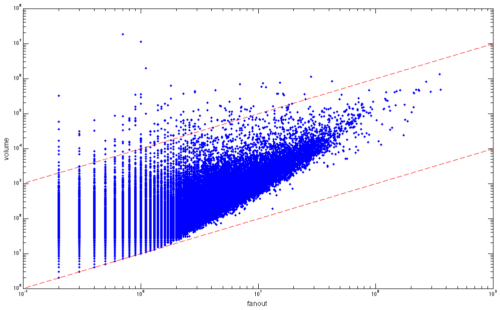

Fig. 21. Volume vs fan-out (log-log scale)

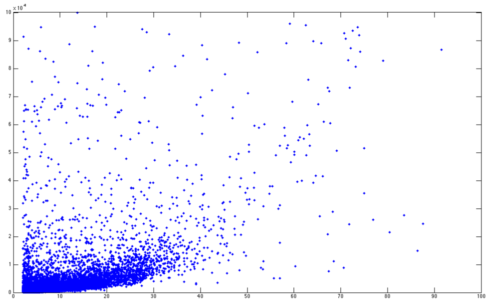

Fig. 22. Volume vs fan-out (lin-lin scale)