Application-Layer DDoS Attacks with Multiple Emulation Dictionaries

Michele Cirillo, Mario Di Mauro, Vincenzo Matta, Marco Tambasco

DIEM, University of Salerno, Italy

Abstract

We consider the problem of identifying the members of a botnet under an application-layer (L7) DDoS attack, where a target site is flooded with a large number of requests that emulate legitimate users' patterns. This challenging problem has been recently addressed with reference to two simplified scenarios, where either all bots pick requests from the same emulation dictionary (total overlap), or they are divided in separate clusters corresponding to distinct emulation dictionaries (no overlap at all). However, over real networks these two extreme conditions are difficult to realize, and the intermediate situation is observed where the emulation patterns of distinct bots belong to *partially overlapped* dictionaries. This intermediate situation introduces significant sophistication in the bot identification problem. In order to address this issue, we provide an analytical characterization of the pairwise cluster interaction, which is exploited to devise an identification rule to discriminate legitimate users from bots and to identify the individual bot clusters.

L7-DDoS Attacks with Emulation Dictionaries

- In L7-DDoS attacks (e.g., *Mirai* [1]), a botnet impairs a network target through a huge number of L7 requests (e.g., HTTP messages)
- In a sophisticated variant of L7-DDoS, a botnet emulates legitimate traffic by gleaning admissible messages from an *emulation dictionary* [2]
- Classic entropy-based approaches [3] are ineffective against such attacks since the traffic activities of *individual* bots are not suspicious

Main innovation: Formalization of a challenging scenario with bots organized in *clusters* accessing multiple *overlapped* emulation dictionaries (Fig. 1)

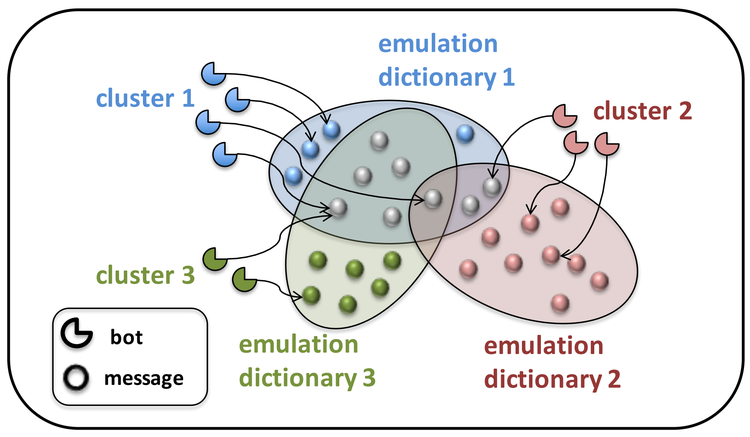

Figure 1: DDoS attack with multiple emulation dictionaries.

Network Indicators

At time t, bots belonging to cluster i pick legitimate messages from the emulation dictionary $\mathscr{E}_i(t)$, whose cardinality grows over time at rate:

$$\alpha \triangleq \lim_{t\to\infty} \frac{|\mathscr{E}_i(t)|}{t}.$$

The *transmission* activity of a subnet $\mathcal{S}$ is summarized by the empirical Transmission Rate:

$$\hat{\lambda}_{\mathcal{S}}(t) \triangleq \frac{N_{\mathcal{S}}(t)}{t}.$$

The *message content* variability is summarized by the empirical Message Innovation Rate (MIR):

$$\hat{\rho}_{\mathcal{S}}(t) \triangleq \frac{|\mathscr{D}_{\mathcal{S}}(t)|}{t}.$$

$N_{\mathcal{S}}(t) \to$ no. of transmissions in $\mathcal{S}$ up to time t.
$\mathscr{D}_{\mathcal{S}}(t) \to$ set of *distinct messages* in $\mathcal{S}$ up to t.

Multiple Emulation Dictionaries

Intersection and Overlap Degree

Intersection of the emulation dictionaries of clusters i and j: $\mathscr{E}_{ij}(t) \triangleq \mathscr{E}_i(t) \cap \mathscr{E}_j(t)$.

Overlap degree of the emulation dictionaries of clusters i and j (see Fig. 2):

$$\lim_{t\to\infty} \frac{|\mathscr{E}_{ij}(t)|}{|\mathscr{E}_i(t)|} = \lim_{t\to\infty} \frac{|\mathscr{E}_{ij}(t)|}{|\mathscr{E}_j(t)|} = \omega_{ij} = \omega_{ji} \in (0,1).$$

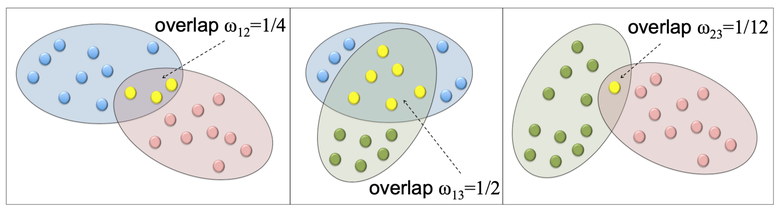

Figure 2: Pairwise overlaps with different values of ω_{ij}.

MIR of the joint subnet $\mathcal{B}_i \cup \mathcal{B}_j$:

$$\hat{\rho}_{\mathcal{B}_i\cup\mathcal{B}_j}(t) \xrightarrow{\text{m.s.}} \rho_{\mathcal{B}_i\cup\mathcal{B}_j} = \omega_{ij}\,\mathscr{R}(\alpha, \lambda_{\mathcal{B}_i} + \lambda_{\mathcal{B}_j}) + (1-\omega_{ij})[\mathscr{R}(\alpha, \lambda_{\mathcal{B}_i}) + \mathscr{R}(\alpha, \lambda_{\mathcal{B}_j})]$$

where $\mathscr{R}(\alpha,\lambda) \triangleq \alpha\lambda/(\alpha+\lambda)$ and the symbol $\xrightarrow{\text{m.s.}}$ denotes mean-square convergence as $t \to \infty$.

The following inequalities hold:

$$\underbrace{\mathscr{R}(\alpha, \lambda_{\mathcal{B}_i} + \lambda_{\mathcal{B}_j})}_{\rho_{\text{tot}}} \le \rho_{\mathcal{B}_i\cup\mathcal{B}_j} \le \underbrace{\mathscr{R}(\alpha, \lambda_{\mathcal{B}_i}) + \mathscr{R}(\alpha, \lambda_{\mathcal{B}_j})}_{\rho_{\text{sum}}}.$$

BotClusterBuster Algorithm

Intuition: Pairwise checks to establish if a *pivot* node p and a *test* node τ form a botnet.

- Let $\gamma(t) = \theta\hat{\rho}_{\text{tot}}(t) + (1-\theta)\hat{\rho}_{\text{sum}}(t),\ \theta \in (0,1)$
- Compare the MIR against the threshold $\gamma(t)$: $\hat{\rho}_{\{p,\tau\}} \le \gamma(t) \Rightarrow$ estimated botnet $\{p,\tau\}$

Three cases (see Fig. 3):
I) p or τ normal: $\hat{\rho}_{\{p,\tau\}} \approx \hat{\rho}_{\text{sum}}(t)$
II) p and τ bots in the same cluster: $\hat{\rho}_{\{p,\tau\}} \approx \hat{\rho}_{\text{tot}}(t)$
III) p and τ bots in distinct clusters i and j: $\hat{\rho}_{\{p,\tau\}} \approx \omega_{ij}\hat{\rho}_{\text{tot}}(t) + (1-\omega_{ij})\hat{\rho}_{\text{sum}}(t)$

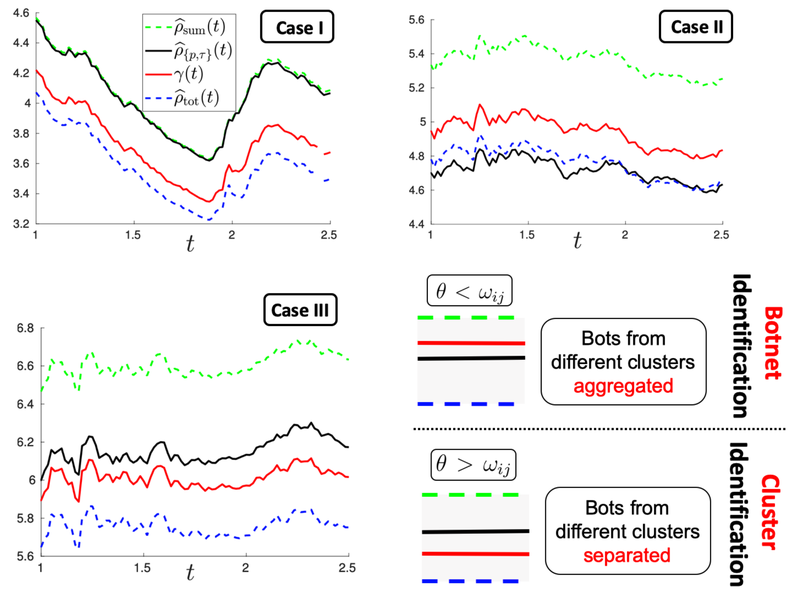

Figure 3: *Top*. MIR evolution in cases I and II. *Bottom*. Case III, where p and τ are in different clusters: on the left we see the MIR evolution, on the right the corresponding Botnet/Cluster identifiability behavior, depending on the condition $\theta \gtrless \omega_{ij}$.

Experimental Setting and Results

- **Normal users**: real-world traffic collected @ Co.Ri.Tel. Laboratory (DIEM, University of Salerno)
- **Botnet**: simulated with emulation dictionaries built by using legitimate patterns from the *normal* dataset

Algorithm: $\widehat{\mathcal{B}}(t)$=BotClusterBuster(traffic patterns until time t, θ, κ, ξ)

$\mathcal{N} = \{1, 2, \ldots, N\}$
for $p \in \mathcal{N}$ **do**
  $\widehat{\mathcal{B}}(p;t) = \{p\}$
  for $\tau \in \mathcal{N} \setminus \{p\}$ **do**
    $\gamma(t) = \theta\hat{\rho}_{\text{tot}}(t) + (1-\theta)\hat{\rho}_{\text{sum}}(t)$
    if $\hat{\rho}_{\widehat{\mathcal{B}}(p;t)\cup\{\tau\}}(t) \le \gamma(t)$ **then** $\widehat{\mathcal{B}}(p;t) = \widehat{\mathcal{B}}(p;t) \bigcup \{\tau\}$
  end
  if $|\widehat{\mathcal{B}}(p;t)| = 1$ **then** $\widehat{\mathcal{B}}(p;t) = \emptyset$
  % begin cluster expurgation
  if $\hat{\lambda}_{\widehat{\mathcal{B}}(p;t)}(t) \le \frac{\xi\kappa}{1+\kappa}\hat{\lambda}_{\mathcal{N}}(t)$ **then** $\widehat{\mathcal{B}}(p;t) = \emptyset$
  % end cluster expurgation
end
$\widehat{\mathcal{B}}(t) = \bigcup_{p=1}^{N} \widehat{\mathcal{B}}(p;t)$

- **Max Rule**: retain only the maximum-size clusters
- **Union Rule**: retain all clusters
- **Expurgation Rule**: discard spurious clusters by exploiting the transmission activity carried by the candidate clusters

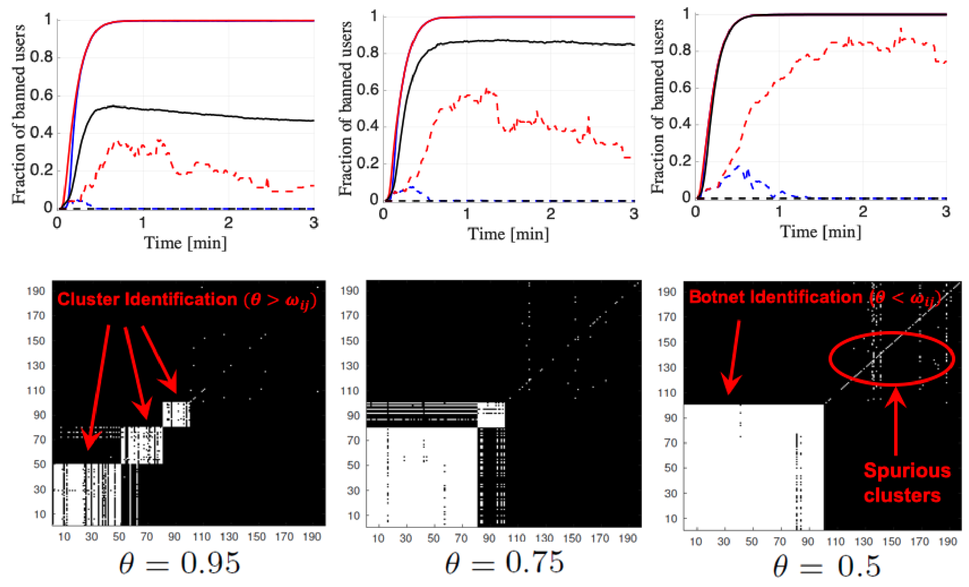

Figure 4: *Left*. BotClusterBuster algorithm. *Three top-right panels*. Performance of BotClusterBuster. Continuous [dashed] curves represent the expected fraction of correctly [uncorrectly] banned users for the **Max**, **Union**, and **Expurgation** rules. *Three bottom-right panels*. Cluster/Botnet identification. As θ decreases, the algorithm looses the ability of identifying clusters, but preserves the ability of identifying the botnet. *Network Setting*: 100 normal users, 100 bots; 3 clusters of 50, 30, 20 bots; $\alpha = 10$; $\omega_{12} = 3/4, \omega_{13} = \omega_{23} = 1/2$.

References

[1] C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas, "DDoS in the IoT: Mirai and Other Botnets," *Computer*, 50 (7), 2017.

[2] V. Matta, M. Di Mauro, and M. Longo, "DDoS Attacks with Randomized Traffic Innovation: Botnet Identification Challenges and Strategies," *IEEE Trans. Inf. Forensics Security*, 12 (8), 2017.

[3] Y. Xiang, K. Li, and W. Zhou, "Low-rate DDoS attacks detection and traceback by using new information metrics," *IEEE Trans. Inf. Forensics Security*, 6 (2), 2011.