Abstract. Spamming botnet has been widely used to send a considerable number of spam for evading the blacklist from information security communities and internet service provider (ISP). Spamming botnet makes the anti-spam be more difficult due to dispersed sending mechanism with large scale sending capability and non-obviously network statistics features (e.g., relatively low network traffics, dynamic network addresses and rapid changes of spam signatures). Because characterizing the spamming botnet behavior in straightforward way is difficult, finding the latent behavior of spamming botnet could block the source of spamming bot in dynamic way.

In this poster, we focus on characterizing spamming botnets by leveraging both spam semantic intentions and spam delivering sources. From the observation of open relay simple mail transfer protocol (SMTP) servers, the activities of spamming botnet could be observed from the delivered spam semantic. Different groups of spamming bots would send the similar subjects of advertisement spam. Furthermore, we proposed a semantic graph analysis approach to dynamic differentiate the different intentions from the sources of spamming bots. The semantic graph analysis approach models the topic of each spam by extracting the tokens and modeling the spam topics from different sources of spamming bots. The topic model of each spam is a token feature vector that is estimated by latent dirichlet allocation (LDA) [1]. The combination of each spam topic model given by the same source could be regarded as the topic model of specified spamming bot. All of analyzed data could be captured from the open relay SMTP servers. According to constructed topic models of different sources of spamming bots, we leverage the independent component analysis (ICA) [2] used to find the latent behavior by decomposing the topic model of each spam into the both independent component vector and base vector. Based on the observed independent components, the density-based approach is applied to observe the distinct groups of spamming behavior.
In our analysis, we make use spam dataset collected from Taiwan’s ISP by open relay SMTP server. The quantity of incoming spam achieve to hundred thousands of spam from several open relay SMTP servers per day in average, and the sources of unique IP address are around thousand per week. The overlapped spam senders (i.e., spammers IP) from different open relay server are less then 5%. Most of spam contents language include both traditional Chinese and simple Chinese. The data collected period is from December 2010 to June 2011. Figure 1 shows one of the clustering result from proposed method. We can see different advertisement intensions of spam bots could be clustered together, for example, the left bottom cluster is pornography category. According to our inspection, the different group of spamming bots are recruited by different advertisers.

In this study, we incorporate the semantic graph with independent component analysis and density-based clustering approach to differentiate the spamming behavior from spamming botnet. In this work, we make following contributions: (1) finding the indirect relations between spamming botnet and advertisers intentions, (2) proposed a novel analysis mechanism to analyze the spamming botnet behavior, (3) alleviate the impact of dynamic spamming bot source addresses. This work is still ongoing, the systematic analysis procedure and complete validation will be complete in the future.

References