A post-mortem empirical investigation of the popularity and distribution of malware files in the contemporary web-facing Internet

This short empirical paper investigates a snapshot of about two million files from a continuously updated big data collection maintained by F-Secure for security intelligence purposes. By further augmenting the snapshot with open data covering about a half of a million files, the paper examines two questions: (a) what is the shape of a probability distribution characterizing the relative share of malware files to all files distributed from web-facing Internet domains, and (b) what is the distribution shaping the popularity of malware files? A bimodal distribution is proposed as an answer to the former question, while a graph theoretical definition for the popularity concept indicates a long-tailed, extreme value distribution. With these two questions - and the answers thereto, the paper contributes to the attempts to understand large-scale characteristics of malware at the grand population level - at the level of the whole Internet.