**Localizador de Fonte - Encontrando Vírus em Código Fonte a partir repositórios no Github públicos e disponiveis**

***Título:*** SourceFinder: Finding Malware Source-Code from Publicly Available Repositories in GitHub. ***Objetivo:***  Criar um grande banco de dados de código fonte de malware. ***Métodos:*** O primeiro passo é coletar repositórios do GitHub que têm uma chance maior de estar relacionados a malware. A extração de repositórios em escala do GitHub esconde vários. Após a fase de coleta utilizar um algoritmo para o processamento dos dados para repositórios com características de malware. Utilizando Pré-processamento de dados; Pré-processamento de nível de caracteres; Pré-processamento de nível de palavra; Filtragem a nível de entidade; Seleção de um motor ML; Detecção de repositórios de código fonte. ***Resultados:*** Foi realizada a coleta 97K repositórios relacionados a malware da GitHub, ou seja, repositórios recuperados usando palavras-chave de malware através do API da GitHub e empregando técnicas para superar várias limitações. O SourceFinder alcança 89% de precisão e a sistematização por caracteristicas. A Identificação de 7504 repositórios de código fonte de malware, que é indiscutivelmente o maior banco de dados de código fonte de malware disponível para a comunidade de pesquisa, repositórios e autores populares. Também foi gerado um banco de dados com 250 repositórios de malware. A descoberta que o número de novos repositórios de malware em nossos dados mais do que triplica a cada quatro anos e a tendência crescente é interessante e alarmante ao mesmo tempo. Para o estudo dos repositórios de malware foi utilizado três métricas de popularidade: o número de observadores, garfos e estrelas. Encontramos 8 repositórios que dominam as listas top-5 para todas as três métricas. Constatamos que 3% dos autores têm mais de 300 seguidores e também descobrimos que 0,2% dos autores têm mais de 7 repositórios de malware, com os autores mais prolíficos tendo criado 336 repositórios. ***Conclusão:*** O trabalho capitaliza uma grande oportunidade perdida: há milhares de repositórios de código fonte malware no GitHub. Ao mesmo tempo, há uma escassez de código fonte de malware, que é necessária para certos estudos de pesquisa. Sendo indiscutivelmente o primeiro a desenvolver um processo sistemático para extrair repositórios de código fonte de malware em escala do GitHub. O estudo em larga escala fornece algumas tendências de interesse tanto para os repositórios de malware quanto para a dinâmica dos autores de malware e como projeto futuro a utlização de outro modelo para identificação de melhores resultados.