CLAIMS

1. A method for clustering executable files implemented on a computer device, the method comprising:

   creating new clusters of a plurality of executable files by:

      obtaining, by the computer device, the plurality of executable files;

      determining, by the computer device, a file format of each executable file of the plurality of executable files;

      for each file format:

         (i) detecting, by the computer device, repeat sequences of commands of a predetermined length in a given one of the plurality of executable files;

         (ii) determining, by the computer device, at least one frequently occurring sequence of the repeat sequences of commands in the given one of the plurality of executable files; and

      based on the at least one frequently occurring sequence of commands, attributing, by the computer device, the given one of the plurality of executable files to a new cluster;

      iteratively executing, by the computer device, the detecting, the determining, and the attributing until one of: all of the plurality of executable files are attributed to at least one cluster, and until un-attributed files of the plurality of executable files do not contain any repeat sequences of commands; and

      in response to presence of un-attributed files of the plurality of executable files, attributing each of the un-attributed files of the plurality of executable files to a separate new cluster.

2. The method of claim 1, further comprising determining the predetermined length of a given repeat sequence based on a number of executable files where the given sequence was detected.

3. The method of claim 1, wherein the predetermined length of a given repeat sequence is fixed.

4. The method of claim 1, wherein the detecting the repeat sequences further comprises determining a respective entropy value of each one of the repeat sequences and selecting only those thereof for further processing whose respective entropy values exceed a predetermined entropy value threshold.

5. The method of claim 1, wherein the at least one frequently occurring sequence is determined by searching therefor in a hash table.

6. The method of claim 1, wherein the attributing the given one of the plurality of executable files to the cluster is based on a result of calculating, for the given executable file, a respective value of a weighting function comprising a sum of coefficients associated with each one of at least two frequently occurring sequences of commands detected in the given executable file.

7. The method of claim 6, wherein each coefficient of the weighing function is calculated as a respective ratio of a number of executable files where a given frequently occurring sequence is present at least once, to a total number of executable files in the plurality of executable files.

8. The method of claim 1, wherein executable files not containing any repeat sequences are not attributed to any cluster.

9. A method for clustering executable files implemented on a computer device, the method comprising:

creating new clusters of a plurality of executable files by:

obtaining, by the computer device, the plurality of executable files;

determining, by the computer device, a file format of each executable file of the plurality of executable files;

launching, by the computer device, based on a respective file format, each executable file of the plurality of executable files in a respective isolated environment to determine a respective machine code of each one of the plurality of executable files;

generating, by the computer device, based on the respective machine code, a respective sample code by deleting therefrom portions non-indicative of a given executable file being affiliated with any cluster;

detecting, by the computer device, in the respective sample code associated with the given executable file, repeat sequences of commands of a predetermined length;

determining, by the computer device, at least one frequently occurring sequence of the repeat sequences of commands in the respective sample code;

based on the at least one frequently occurring sequence, attributing, by the computer device, the given executable file to a new cluster;

iteratively executing, by the computer device, the detecting, the determining, and the attributing until one of: all of the plurality of executable files are attributed to at least one cluster, and until un-attributed files of the plurality of executable files do not contain any repeat sequences of commands; and

in response to presence of un-attributed files of the plurality of executable files, attributing each of the un-attributed files of the plurality of executable files to a separate new cluster.

10. The method of claim 9, further comprising determining the predetermined length of a given repeat sequence based on a number of sample codes where the given sequence was detected.

11. The method of claim 9, wherein the predetermined length of a given repeat sequence is fixed.

12. The method of claim 9, wherein the detecting the repeat sequences further comprises determining a respective entropy value of each one of the repeat sequences and selecting only those thereof for further processing whose respective entropy values exceed a predetermined entropy value threshold.

13. The method of claim 9, wherein the at least one frequently occurring sequence is determined by searching therefor in a hash table.

14. The method of claim 9, wherein the attributing the given one of the plurality of executable files to the cluster is based on a result of calculating, for the respective sample code, a respective value of a weighting function comprising a sum of coefficients associated with each one of at least two frequently occurring sequences of commands detected in the respective sample code.

15. The method of claim 14, wherein each coefficient of the weighing function is calculated as a respective ratio of a number of executable files where a given frequently occurring sequence is present at least once, to a total number of executable files in the plurality of executable files.

16. The method of claim 9, wherein executable files not containing any repeat sequences are not attributed to any cluster.

17. A system for clustering executable files comprising:

a processor;

a non-transitory computer readable medium storing instructions; and

the processor, upon executing the instructions being configured to:

creating new clusters of a plurality of executable files by:

obtain the plurality of executable files;

determine a file format of each executable file of the plurality of executable files;

for each file format:

(i) detect repeat sequences of commands of a predetermined length in a given one of the plurality of executable files;

(ii) determine at least one frequently occurring sequence of the repeat sequences of commands in the given one of the plurality of executable files; and

based on the at least one frequently occurring sequence of commands, attribute the given one of the plurality of executable files to a new cluster;

iteratively execute the detecting, the determining, and attributing until one of: all of the plurality of executable files are attributed to at least one cluster, and until un-attributed files of the plurality of executable files do not contain any repeat sequences of commands; and

in response to presence of un-attributed files of the plurality of executable files, attribute each of the un-attributed files of the plurality of executable files to a separate new cluster.

18. A system for clustering executable files comprising:

a processor;

a non-transitory computer readable medium storing instructions; and

the processor, upon executing the instructions being configured to:

creating new clusters of a plurality of executable files by:

obtain the plurality of executable files;

determine a file format of each executable file of the plurality of executable files;

for each file format:

(i) launch each executable file of the plurality of executable files in a respective isolated environment to determine a respective machine code of each one of the plurality of executable files;

(ii) generate a respective sample code by deleting from the respective machine code portions non-indicative of a given executable file being affiliated with any cluster;

detect in the respective sample code associated with the given executable file, repeat sequences of commands of a predetermined length;

determine at least one frequently occurring sequence of the repeat sequences of commands in the respective sample code;

based on the at least one frequently occurring sequence, attribute the given one of the plurality of executable files to a new cluster;

iteratively execute the detecting, the determining, and the attributing until one of: all of the plurality of executable files are attributed to at least one cluster, and until un-attributed files of the plurality of executable files do not contain any repeat sequences of commands; and

in response to presence of un-attributed files of the plurality of executable files, attribute each of the un-attributed files of the plurality of executable files to a separate new cluster.