Attacking DBSCAN for Fun and Profit

Authors: Jonathan Crussell, W. Philip Kegelmeyer

SAND2015-3895C

Presenter: Jonathan Crussell, jcrusse@sandia.gov

Theme area: Extreme Scale Data, Knowledge, and Analytics

Many security applications depend critically on clustering. However, we do not know of any clustering algorithms that were designed with an adversary in mind. An intelligent adversary may be able to use this to her advantage to subvert the security of the application. Already, adversaries use obfuscation and other techniques to alter the *representation* of their inputs in feature space to avoid detection. As one example, spam email often roughly mimics normal email. We have investigated a more nuanced and informed attack, in which an adversary attempts to subvert clustering analysis by feeding in carefully crafted data points.

This can be effective because clustering is often applied to "found data" -- data whose origin and integrity may be uncertain. Continuing with the above example, an analyst may wish to use clustering to collect, from a large email corpus, all the emails from the same spam campaign. In this case, anyone may have sent emails that would be included in the data set. Since the data being analyzed is "found data", the analyst has little knowledge about what the structure of the clusters should look like in the absence of any adversarially supplied data points. Without secondary analysis, such as manual investigation (or the remediation method we invented and mention below), the adversary may therefore alter the clustering structure to her advantage.

Specifically, we have explored what we call a "confidence attack", where an adversary seeks to poison the clusters to the point that the defender loses confidence in the utility of the system. This may result in the system being abandoned, or worse, waste the defender's time investigating false alarms.

In particular, we have explored how an attacker can subvert DBSCAN, a popular density-based clustering algorithm. While our attacks generalize to all DBSCAN-based tools, we focused our evaluation on AnDarwin, a tool designed to detect plagiarized Android apps, where the apps themselves are processed to form a compact representation suitable for scalable analysis. AnDarwin detects app plagiarism by tightly clustering apps using DBSCAN such that if two apps are in the same cluster, they are very likely to be copies of each other. AnDarwin can then use these clusters and the developer information associated with each app to identify clusters that contain apps from more than a single author. For clusters with more than a single author, one or more of the authors must have plagiarized apps from the others.

Our attacks show how an adversary can merge arbitrary clusters by connecting them with a series of "data mines" that "bridges" the space between the two clusters. Surprisingly, even a small number of merges can greatly degrade the clustering performance that we measure using the accuracy of app plagiarism detection. We also show that the analyst has limited recourse when relying solely on DBSCAN -- she can alter clustering parameters to increase the number of data mines that the adversary has to craft, however, this will make her to miss plagiarizing apps.

Finally, we have created a remediation process that uses machine learning and features based on outlier measures that are orthogonal to the underlying clustering problem to detect and remove injected points. This provides a way for the analyst to sanitize her data set and remove points that are suspected of being adversarially created.

This work connects to the themes of the workshop along a couple of dimensions. The application (identifying the validity of computer executables) is clearly of cybersecurity interest, and we have demonstrated ways to degrade (and perhaps restore) our trust in doing that identification reliably. Further, the effectiveness of the attack stems from design choices (in both the data representation and the clustering algorithm) that were driven by the "extreme scale data" requirements inherent in clustering the millions of constantly arriving executables. The general point, investigated via one example in this white paper, is that it is important to be aware that clever and effective handling of extreme scale data can inherently introduce exploitable vulnerabilities.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.