2. Methods to mask, anonymize, or perturb cyber threat indicator data while maximizing data utility

Performers: Dr. Tim Menzies (NC State), Rebecca Stoerts (Duke)

Two researchers with extensive experience in privacy and data anonymization will focus on continuing to improve mechanisms for masking and anonymizing cyber threat data while maintaining its underlying utility with the development of various algorithms. Dr. Tim Menzies will continue to develop a refine the **LACE2** privacy+learning algorithm to exploit the difference between (a) generalizing across over a population vs (b) looking specifics up on one individuals. **LACE2** privatizes data by combining row pruning and column pruning. The resulting data sets hold just a tiny fraction of the original data; e.g. 10% to 1% of the original data. As to ensuring the privacy of the remaining data, **LACE2** system clusters the pruned data then mutated the cluster centroids by a random amount up to, but not more than, half the distance to its nearest neighbor. Such a mutation policy preserves the topology of the example space, while preserving very little of the data for any individual. Given databases of phishing information, the **LACE2** algorithm will be applied to increasingly larger datasets and data collected from different sites. This will allow Dr. Menzies to test if (a) sharing data from multiple sites enables better early warning for phishing attacks (i.e. better than just using local data) and (b) if such shared data can be privatized before sharing, while still enabling the creation of early warnings about security issues.

Deliverables:

- Code for the LACE2 prototype
- Example phishing data sets that other researchers could use to explore other phishing management technologies
- Performance results of LACE2 executing over some target domain (currently, early detection of phishing attacks):
    - Learner performance results when learners are applied to (a) raw data or (b) data privatized by LACE
    - Learner performance results where learners execute over a stream of data containing sequences of phishing reports (this will test how early we can classify new examples as phishing attacks, as well as stress testing our transfer learning methods)
    - Information theoretic measures of the information content within LACE2's privatized data (this will allow us to assess the effectiveness of LACE2's privatization.
    - Plots of runtimes vs data sets size for LACE2 (this will test how well this schema scales)
- Plots of runtimes vs data sets sizes for homomorphic encryption (this is an alternate technology to LACE2, with a reputation of being very slow). These runtime lots will allow for a comparative assessment of LACE2 vs alternate methods

LAS-G Staff: Jen Wiechmann, Austin Allshouse, Deb Crawford, Sandra Harrell-Cook