**Department of Computer Science & Engineering**

**Masters Thesis Defense**

**“Private Record Linkage: A Comparison of Selected Techniques for Name Matching”**

***By Pawel Grzebala***

**Monday, April 11, 2016 – 3:00 pm**

**304 Russ**

**Committee:** **Drs. Michelle Cheatham, Advisor, Keke Chen, and TK Prasad**

**ABSTRACT:**

The rise of Big Data Analytics has shown the utility of analyzing all aspects of a problem by bringing together disparate data sets. Efficient and accurate private record linkage algorithms are necessary to achieve this. However, records are often linked based on personally identifiable information, and protecting the privacy of individuals is critical. This work contributes to this field by studying an important component of the private record linkage problem: linking based on names while keeping those names encrypted, both on disk and in memory. We explore the applicability, accuracy, speed and security of three different primary approaches to this problem (along with several variations) and compare the results to common name-matching metrics on unprotected data. While these approaches are not new, this work provides a thorough analysis on a range of datasets containing systematically introduced flaws common to name-based data entry, such as typographical errors, optical character recognition errors, and phonetic errors. Additionally, we evaluate the privacy level of the q-grams based metrics by simulating the frequency analysis attack that can occur in case of potential data breaches. We show that, for the use case we are considering, the best choice of string metric are padded q-gram based metrics which can provide high record linkage accuracy and are resilient to frequency analysis attack under certain conditions.