Transitional Relevance and Research Summary Statement

**Name:** Sarah Elder

**Hard Problem Area:** Predictive Security Metrics

**Overall Research Goal(s):**

Many security vulnerabilities could have been avoided had security threat mitigations, e.g. input sanitation, been properly implemented during software development. Security requirements are an important aspect of improving software security through a secure software development lifecycle. The goal of this research is to assist developers and other software development practitioners in understanding and applying security best practices by finding an efficient means of creating a minimal but thorough set of security requirements based upon from the available product documentation and security policies. These policies may include HIPAA, corporate policies, and other regulations. After we conduct our preliminary studies to gain a better understanding of the current research and industry landscape, we will focus on identifying areas of similarity and overlap between related statements in policies and regulations. By identifying and understanding the areas where documents are related, we can better develop a comprehensive, concise model of the information in this documentation that can be used to develop security requirements.

**Broad Impact:**

Policies and Regulations plays an important role in security in practice. These documents are generally influenced, and often written by folks who have more expertise in software security principles than your average practitioner. Previous research in this area has shown that regulations and policies are often challenging to understand, can have significant overlap, and may even be contradictory. Hence although these documents may provide a good starting point for improving the integration of security into the software development process, they must be understood and the information should be condensed into a form to be used by practitioner.

**Specific Research Goals:**

Our preliminary research will include a Systematic Literature Review (SLR) and a survey. For the SLR we will take a saturation approach. We will iterate through the process of identifying and refining our search until we are no longer identifying new, relevant documents. At this point we will begin the formal review. From the survey, we hope to identify the knowledge sources used in industry to identify and understand security best practices. We also hope to better understand the process through which these knowledge sources are applied to a software product, including the processes through which it is determined how well a software product meets a security standard.

The remainder of this research will focus on how to apply the knowledge gained to better identify security requirements from the given knowledge sources. We will begin this work by looking at ways to better identify related security statements that may come from a variety of different documents. This may include finding was to handle different syntactic

---

structures and levels of abstraction. Ultimately, we will need to accurately relate these statements back to a software system.

**Proposed Data Collection (if applicable):**

There are at least three different areas where we will be collecting distinct data sets. The first is the SLR where we will develop a Quasi-Gold-Standard set of documents that are most relevant to our topic. Of the documents identified so far, many come from the RE conference. However, there are outliers and we have not formally identified inclusion criteria for this set. We will then perform frequency analysis on documents to identify a set of search keywords to use. We will use these keywords to perform an automated search to identify the relevant documents. This process may include some manual searching in addition to the identification of the Quasi-Gold-Standard to supplement or verify the set of documents found through the automated process.

The second component is the survey. While we will base the survey on best practices, we do not yet know what questions should be included. Hence while the use of likert scales and other metrics are highly plausible, more work is necessary before this is finalized.

Finally, once the background research is complete and we have identified the best path forward, we will need to test our process on a variety of documents. Fortunately, there are a number of Federal Regulations easily available. If our process requires product documentation, we are likely to begin with documents from open source systems like OpenMRS. Eventually we would like to conduct a case study or similar on a range of production systems.

**Success Criteria:**

At this point our success criteria are still fairly abstract. From the SLR and survey, we should have a better understanding of the problem, as well as a stronger methodology going forward. Our next goal will be met when we can identify related statements at an appropriate level of detail. However, what that level of detail is will depend on the results from the SLR and the survey. Similarly, finding a way to integrate this information into system requirements and design, will depend on previous findings.

**Anticipated Difficulties, Limitations, and Criticisms:**

For both the survey and the overall analysis, a likely limitation will be the available data. Ideally, the survey questions would be based on formal interviews with industry professionals. However, the development of the survey questions may need to rely heavily on other sources such as information derived from the SLR. Similarly, the resulting process should be evaluated against a wide spectrum of products. However, our initial studies may be focused on easily available, open-source systems which are not always designed to meet the same security standards.

Two potential limitations of the SLR will be the scope of the SLR, and the potential for bias in the Quasi-Gold-Standard. Bias in the Quasi-Gold-Standard could occur for many reasons and would lead to the omission of key documents. The scope of the SLR poses more serious issues. To review every potential related document would be impossible. Instead we must impose restrictions on the timeframe in which the documents were published and the content we are looking for. However, this could also lead to serious omissions. For example, if we focus on ways that NLP and machine learning have been applied to security policies, we may miss techniques that have been applied in other contexts but would be applicable here. These risks will be reduced by iteratively refining the search process, but will not disappear.