CS586 - Grad Project - Submission 1
Instructor: Charles Winstead
Student: Scott Griffy

1 Domain

The database will scrape popular C github repositories (using the github API or traditional web scraping)
I'm hoping to ask questions that are security-oriented, such as querying the most popular crypto libraries and functions within those libraries.
I'm using the term "popular" so that I can scale the size of the database arbitrarily in order to avoid running out of space or scrape time.
I'm planning to focus on the language C, because of the maturity of crypto libraries written in that language.
This means that I'll be mostly collecting data about repositories, such as the number of forks, commits, pull requests. In the repositories table, I'll also include the presence of code snippets like "$\#include < openssl.h >$" which would indicate that the openssl library was used in that project. A number of relation tables will be used to signify use of a crypto library. The repositories table will be the largest table.
Other tables will include Users, Projects, Organizations, and various relations. Users will contribute to a number of repositories and their activity rate will be measured. Popular/active users' commits and forks will be tracked as well to answer questions about which libraries certain users prefer. Projects and Organizations will reference repos and users that work under them, relating themselves to certain crypto libraries. Separate tables will be used for crypto repositories and repositories that use crypto and crypto library repositories will be hand-picked due to the difficulty of automatically detecting crypto library repos.
The distinction between crypto-using libraries and crypto libraries should be made early and not include too many crypto libraries. The selection should focus on generic crypto primitives, and not classify libraries that try to provide functionality for a more specific application as crypto libraries
Querying all of the repositories on github will take a lot of time. This is why only the most interesting repositories will be queried, such as repos with the most stars or forks.

2 Implementation

One potential way to search these libraries would be to use this API method:
https://developer.github.com/v3/search/#search-repositories
Using a HTTP request like this:

```
curl https://api.github.com/search/repositories?q=crypto+language:c&sort=stars&order=
    desc
```

The sort parameter would be used to find popular/relevant repositories quickly.
I then intend to search through the code for interesting portions, such as "#include" functions. Hopefully cloning will not be necessary and instead, another portion of the search API can be used:
https://developer.github.com/v3/search/#search-code

```
https://api.github.com/search/code?q=define+in:file+language:c+repo:torvalds/linux
```

The solution will be coded up in Python (version 3.X), using the "requests" library and builtin json parsing. OAuth hopefully won't be needed, and ZIPs can be downloaded instead of cloning.
I'll setup a Postgres server on my home computer to store the data. Actual scraping may be done in the cloud (or other remote services). DB storage may be done remotely as well if transfering data to my house becomes difficult.

3 Questions

Here are 20+ questions I could ask of my collected data:
What are the most popular crypto libraries?
What are the most popular crypto functions within those libraries?
What are the most common crypto functions that crypto libraries provide?
Which crypto libraries have the most forks?
Which crypto libraries have the most stars?
Which crypto libraries have the most commits?
Which crypto libraries have the most pull requests?
Which crypto libraries have the most bugs?
Which crypto libraries have the most feature requests?
Which crypto libraries have the most topics (tags)?
Which crypto libraries are related to the most forks? (which cyrpto libraries are used in projects with a high number of forks)
Which crypto libraries are related to the most stars?
Which crypto libraries are related to the most commits?
Which crypto libraries are related to the most pull requests?
Which crypto libraries are related to the most bugs?
Which crypto libraries are related to the most feature requests?
Which crypto libraries are related to the most topics (tags)?
What other programming languages is OpenSSL most related to? (would require scrap-

ing more data besides C repositories)
Which crypto libraries do the most prevalent github users use?
Which crypto libraries do the oldest github repositories use?
Which crypto libraries do the newest github repositories use?
Which crypto libraries are most often used in combination?
Which crypto libraries are used in projects with the most varied contributors?
Which projects house the most crypto libraries?
Which organizations house the most crypto libraries?