CS586 - Grad Project - Submission 3 - Final
Instructor: Charles Winstead
Student: Scott Griffy

Contents

1 Intro

I created a scraper, database, and web interface that allows users to ask questions about the usage of different crypto libraries on github.
I ended up targeting 9 specific C crypto libraries due to the accuracy while querying them and their popularity:
openssl - https://www.openssl.org/
libsodium - https://github.com/jedisct1/libsodium/blob/master/README.markdown
wolfssl - https://www.wolfssl.com/
gnutls - https://www.gnutls.org/
libtomcrypt - https://github.com/libtom/libtomcrypt/blob/develop/README.md
mbedtls - https://tls.mbed.org/
nettle - https://www.lysator.liu.se/ nisse/nettle/
libgcrypt - https://www.gnupg.org/software/libgcrypt/index.html
matrixssl - https://github.com/matrixssl/matrixssl/blob/master/README.md
I used the Github API to populate my database, using the 'requests' package of python.

2 Changes

Because of unavailablility in the Github API and time constraints, some changes have occurred in the schema.
Organizations don't have company information.
A repo's "bugs" actually references it's number of open issues.
The "ContributedTo" table has been removed.
"ProvidesFunction" and "CryptoFunction" have been removed.
"instances" of "UsesLanguage" has been removed.

"type" of CryptoUsingRepo has been removed.
"license" of CryptoProvidingRepo has been removed.
"is_security_related" of Topic has been removed.
But I did add in an extra column into "UsesCryptoLibrary", the "num_indicators", which lets me build confidence scores on whether a library is actually being used.
Determining whether or not a library is being used ended up being more difficult than I expected, and querying each repo for each crypto library took a lot of time, so asking my most important question (which are the most used crypto libraries on github) took a lot more time, but I wanted to get a good answer.

3 Questions

Here are 20 questions I can ask of my collected data
Some of these queries had conditional ommitted due in order to simplify them. These only filtered out results for bad crypto libraries that haven't been removed frmo the db.
Because my database has 881 repos, I can't really show the full results for some of the questions
I've changed around some of the questions because I had to shrink my ER diagram due to time constraints and problem with the github API. The originals will be listed at the end.
What are the most used crypto libraries on github?

```
select C.repo_name,count(*) from "gitsec.UsesCryptoLibrary" U left join "gitsec.
   CryptoProvidingRepo" C on U.crypto_library=C.repo_name where num_indicators>0
   group by C.repo_name order by count(*) desc;
      repo_name      | count
---------------------+-------
 openssl/openssl     |   252
 ARMmbed/mbedtls     |    66
 gnutls/gnutls       |    43
 gpg/libgcrypt       |    32
 breadwallet/nettle  |    21
 jedisct1/libsodium  |    15
 wolfSSL/wolfssl     |    11
 libtom/libtomcrypt  |    11
 matrixssl/matrixssl |     1
(9 rows)
```

Which crypto libraries have the most forks?

```
select R.full_name,R.forks from "gitsec.Repo" R inner join "gitsec.CryptoProvidingRepo
   " C on R.full_name=C.repo_name order by R.forks desc;
      full_name      | forks
---------------------+-------
 openssl/openssl     |  4022
 ARMmbed/mbedtls     |   974
 jedisct1/libsodium  |   951
 libtom/libtomcrypt  |   263
 wolfSSL/wolfssl     |   229
 matrixssl/matrixssl |    23
 gpg/libgcrypt       |    17
 breadwallet/nettle  |    11
 gnutls/gnutls       |     8
(9 rows)
```

Which crypto libraries have the most stars?

```
select R.full_name,R.stars from "gitsec.Repo" R inner join "gitsec.CryptoProvidingRepo
   " C on R.full_name=C.repo_name order by R.stars desc;
      full_name      | stars
---------------------+-------
 openssl/openssl     |  8795
 jedisct1/libsodium  |  6226
 ARMmbed/mbedtls     |  1664
 libtom/libtomcrypt  |   759
 wolfSSL/wolfssl     |   497
 matrixssl/matrixssl |    91
 gpg/libgcrypt       |    38
 gnutls/gnutls       |    18
 breadwallet/nettle  |     3
(9 rows)
```

Which crypto libraries have the most pull requests?

```
select R.full_name,R.pulls from "gitsec.Repo" R inner join "gitsec.CryptoProvidingRepo
   " C on R.full_name=C.repo_name order by R.pulls desc;
      full_name      | pulls
---------------------+-------
 openssl/openssl     |  7726
 ARMmbed/mbedtls     |  2241
 wolfSSL/wolfssl     |  1941
 jedisct1/libsodium  |   688
 libtom/libtomcrypt  |   457
 matrixssl/matrixssl |    14
 gpg/libgcrypt       |     5
 gnutls/gnutls       |     0
 breadwallet/nettle  |     0
(9 rows)
```

Which crypto libraries have the most bugs?

```
select R.full_name,R.bugs from "gitsec.Repo" R inner join "gitsec.CryptoProvidingRepo"
    C on R.full_name=C.repo_name order by R.bugs desc;
      full_name      | bugs
---------------------+------
 ARMmbed/mbedtls     |  610
 openssl/openssl     |  604
 libtom/libtomcrypt  |   35
 wolfSSL/wolfssl     |   26
 matrixssl/matrixssl |   17
 jedisct1/libsodium  |    7
 gpg/libgcrypt       |    3
 breadwallet/nettle  |    1
 gnutls/gnutls       |    0
(9 rows)
```

Which crypto libraries have the most subscribers?

```
select R.full_name,R.subscribers from "gitsec.Repo" R inner join "gitsec.
   CryptoProvidingRepo" C on R.full_name=C.repo_name order by R.subscribers desc;
      full_name       | subscribers
----------------------+-------------
 openssl/openssl      |         834
 jedisct1/libsodium   |         377
 ARMmbed/mbedtls      |         174
 libtom/libtomcrypt   |          92
 wolfSSL/wolfssl      |          64
 matrixssl/matrixssl  |          17
 gnutls/gnutls        |           9
 breadwallet/nettle   |           7
 gpg/libgcrypt        |           6
(9 rows)
```

Which crypto libraries have the most topics (tags)?

```
select R.full_name,count(T.topic_name) from "gitsec.Repo" R inner join "gitsec.
   CryptoProvidingRepo" C on R.full_name=C.repo_name left join "gitsec.RelatedToTopic
   " T on T.repo_name=R.full_name group by R.full_name order by count(T.topic_name)
   desc;
      full_name       | count
----------------------+-------
 wolfSSL/wolfssl      |    20
 openssl/openssl      |     6
 ARMmbed/mbedtls      |     4
 libtom/libtomcrypt   |     4
 jedisct1/libsodium   |     3
 breadwallet/nettle   |     0
 gnutls/gnutls        |     0
 matrixssl/matrixssl  |     0
 gpg/libgcrypt        |     0
(9 rows)
```

Which crypto libraries are related to the most forks? (which cyrpto libraries are used in projects with a high number of forks)

```
select C.repo_name,sum(R.forks) from "gitsec.Repo" R inner join "gitsec.
   UsesCryptoLibrary" U on R.full_name=U.library inner join "gitsec.
   CryptoProvidingRepo" C on U.crypto_library=C.repo_name where U.num_indicators > 0
   group by C.repo_name order by sum(R.forks) desc;
      repo_name       |  sum
----------------------+--------
 openssl/openssl      | 203491
 ARMmbed/mbedtls      |  58778
 breadwallet/nettle   |  40214
 gnutls/gnutls        |  40055
 gpg/libgcrypt        |  29673
 jedisct1/libsodium   |  13869
 libtom/libtomcrypt   |  10809
 wolfSSL/wolfssl      |   7272
 matrixssl/matrixssl  |    765
(9 rows)
```

Which crypto libraries are related to the most stars?

```
select C.repo_name,sum(R.stars) from "gitsec.Repo" R inner join "gitsec.
   UsesCryptoLibrary" U on R.full_name=U.library inner join "gitsec.
   CryptoProvidingRepo" C on U.crypto_library=C.repo_name where U.num_indicators > 0
   group by C.repo_name order by sum(R.stars) desc;
      repo_name      |  sum
---------------------+--------
 openssl/openssl     | 711273
 ARMmbed/mbedtls     | 184022
 gnutls/gnutls       | 144715
 breadwallet/nettle  | 123926
 gpg/libgcrypt       |  95033
 jedisct1/libsodium  |  56503
 libtom/libtomcrypt  |  36254
 wolfSSL/wolfssl     |  25847
 matrixssl/matrixssl |   1892
(9 rows)
```

Which crypto libraries are related to the most pull requests?

```
select C.repo_name,sum(R.pulls) from "gitsec.Repo" R inner join "gitsec.
   UsesCryptoLibrary" U on R.full_name=U.library inner join "gitsec.
   CryptoProvidingRepo" C on U.crypto_library=C.repo_name where U.num_indicators > 0
   group by C.repo_name having sum(R.pulls) > 0 order by sum(R.pulls) desc;
     repo_name      |  sum
--------------------+-------
 openssl/openssl    | 21234
 jedisct1/libsodium |  4377
 gpg/libgcrypt      |  3689
 gnutls/gnutls      |  3689
 ARMmbed/mbedtls    |  2241
 breadwallet/nettle |   626
(6 rows)
```

Which crypto libraries are related to the most bugs?

```
select C.repo_name,sum(R.bugs) from "gitsec.Repo" R inner join "gitsec.
   UsesCryptoLibrary" U on R.full_name=U.library inner join "gitsec.
   CryptoProvidingRepo" C on U.crypto_library=C.repo_name where U.num_indicators > 0
   group by C.repo_name having sum(R.bugs) > 0 order by sum(R.bugs) desc;
     repo_name      | sum
--------------------+------
 openssl/openssl    | 2905
 ARMmbed/mbedtls    |  610
 breadwallet/nettle |  241
 jedisct1/libsodium |  175
 gpg/libgcrypt      |  168
 gnutls/gnutls      |  168
(6 rows)
```

Which crypto libraries are related to the most topics (tags)?

```
select C.repo_name,count(DISTINCT T.topic_name) from "gitsec.Repo" R inner join "
   gitsec.RelatedToTopic" T on R.full_name=T.repo_name inner join "gitsec.
   UsesCryptoLibrary" U on R.full_name=U.library inner join "gitsec.
   CryptoProvidingRepo" C on U.crypto_library=C.repo_name where U.num_indicators > 0
   group by C.repo_name having count(DISTINCT T.topic_name) > 0 order by count(T.
   topic_name) desc;
     repo_name      | count
--------------------+-------
 openssl/openssl    |   533
 ARMmbed/mbedtls    |   185
 gnutls/gnutls      |   127
 gpg/libgcrypt      |    82
 breadwallet/nettle |    44
 jedisct1/libsodium |    32
 libtom/libtomcrypt |    31
 wolfSSL/wolfssl    |    31
(8 rows)
```

Which crypto libraries are related to the most languages?

```
select C.repo_name,count(DISTINCT L.language_name) from "gitsec.Repo" R inner join "
   gitsec.UsesLanguage" L on R.full_name=L.repo_name inner join "gitsec.
   UsesCryptoLibrary" U on R.full_name=U.library inner join "gitsec.
   CryptoProvidingRepo" C on U.crypto_library=C.repo_name where U.num_indicators > 0
   and (C.hide_from_ui <> '1' or C.hide_from_ui is null) group by C.repo_name having
   count(DISTINCT L.language_name) > 0 order by count(DISTINCT L.language_name) desc;
      repo_name      | count
---------------------+-------
 openssl/openssl     |   148
 ARMmbed/mbedtls     |   123
 gnutls/gnutls       |   115
 breadwallet/nettle  |   104
 gpg/libgcrypt       |   103
 jedisct1/libsodium  |    96
 libtom/libtomcrypt  |    78
 wolfSSL/wolfssl     |    58
 matrixssl/matrixssl |    19
(9 rows)
```

What other programming languages is OpenSSL most related to?

```
select L.language_name,count(U.crypto_library) from "gitsec.UsesCryptoLibrary" U inner
    join "gitsec.UsesLanguage" L on L.repo_name=U.library where U.num_indicators>0
   and U.crypto_library='openssl/openssl' group by L.language_name order by count(L.
   language_name) desc;
      language_name        | count
---------------------------+-------
 C                         |   240
 Shell                     |   213
 C++                       |   210
 Makefile                  |   206
 Objective-C               |   149
 Python                    |   145
 Perl                      |   117
 HTML                      |   109
 M4                        |   105
 CMake                     |    85
 Roff                      |    84
 Assembly                  |    80
 Batchfile                 |    79
 JavaScript                |    74
 CSS                      |    60
 Lua                      |    50
 Ruby                     |    48
 Yacc                     |    45
 PHP                      |    43
...
(148 rows)
```

What other programming languages is mbedtls (polarssl) most related to?

```
select L.language_name,count(U.crypto_library) from "gitsec.UsesCryptoLibrary" U inner
    join "gitsec.UsesLanguage" L on L.repo_name=U.library where U.num_indicators>0
   and U.crypto_library=’ARMmbed/mbedtls’ group by L.language_name order by count(L.
   language_name) desc;
      language_name       | count
--------------------------+-------
 C                        |    63
 C++                      |    60
 Shell                    |    58
 Makefile                 |    57
 Objective-C              |    48
 HTML                     |    43
 Python                   |    42
 Assembly                 |    38
 Perl                     |    36
 Batchfile                |    35
 CMake                    |    31
 JavaScript               |    29
 Roff                     |    23
 M4                       |    22
 CSS                      |    20
 Lua                      |    17
 Awk                      |    15
 Ruby                     |    15
 Yacc                     |    14
 PHP                      |    13
...
(123 rows)
```

What other programming languages is gnutls most related to?

```
select L.language_name,count(U.crypto_library) from "gitsec.UsesCryptoLibrary" U inner
    join "gitsec.UsesLanguage" L on L.repo_name=U.library where U.num_indicators>0
   and U.crypto_library=’gnutls/gnutls’ group by L.language_name order by count(L.
   language_name) desc;
      language_name       | count
--------------------------+-------
 C                        |    40
 Shell                    |    39
 C++                      |    38
 Makefile                 |    35
 Objective-C              |    32
 Python                   |    29
 Perl                     |    27
 HTML                     |    25
 M4                       |    22
 Batchfile                |    18
 CMake                    |    17
 Assembly                 |    17
 Roff                     |    16
 JavaScript               |    15
 CSS                      |    13
 Ruby                     |    12
 Yacc                     |    10
 Lua                      |    10
 PHP                      |    10
 Awk                      |    10
...
(115 rows)
```

Which crypto libraries have the most followed authors?

```
select U.name,C.repo_name,U.followers from "gitsec.CryptoProvidingRepo" C inner join "
   gitsec.Repo" R on R.full_name=C.repo_name inner join "gitsec.User" U on U.name=R.
   owner order by U.followers desc;
    name     |      repo_name      | followers
-------------+---------------------+-----------
 jedisct1    | jedisct1/libsodium  |      1446
 gnutls      | gnutls/gnutls       |         0
 libtom      | libtom/libtomcrypt  |         0
 wolfSSL     | wolfSSL/wolfssl     |         0
 openssl     | openssl/openssl     |         0
 breadwallet | breadwallet/nettle  |         0
 gpg         | gpg/libgcrypt       |         0
 ARMmbed     | ARMmbed/mbedtls     |         0
 matrixssl   | matrixssl/matrixssl |         0
(9 rows)
```

Which crypto libraries authors are following the most people?

```
select U.name,C.repo_name,U.following from "gitsec.CryptoProvidingRepo" C inner join "
   gitsec.Repo" R on R.full_name=C.repo_name inner join "gitsec.User" U on U.name=R.
   owner order by U.following desc;
    name     |      repo_name      | following
-------------+---------------------+-----------
 jedisct1    | jedisct1/libsodium  |        91
 gnutls      | gnutls/gnutls       |         0
 libtom      | libtom/libtomcrypt  |         0
 wolfSSL     | wolfSSL/wolfssl     |         0
 openssl     | openssl/openssl     |         0
 breadwallet | breadwallet/nettle  |         0
 gpg         | gpg/libgcrypt       |         0
 ARMmbed     | ARMmbed/mbedtls     |         0
 matrixssl   | matrixssl/matrixssl |         0
(9 rows)
```

Which crypto libraries do the oldest github repositories use?

```
select
  R.full_name,
  to_timestamp(R.created_at),
  (select crypto_library from "gitsec.UsesCryptoLibrary" U inner join "gitsec.
     CryptoProvidingRepo" P on U.crypto_library=P.repo_name where U.library=R.
     full_name and (P.hide_from_ui <> '1' or P.hide_from_ui is null) and
     num_indicators > 0 order by num_indicators desc limit 1),
  (select num_indicators from "gitsec.UsesCryptoLibrary" U inner join "gitsec.
     CryptoProvidingRepo" P on U.crypto_library=P.repo_name where U.library=R.
     full_name and (P.hide_from_ui <> '1' or P.hide_from_ui is null) and
     num_indicators > 0 order by num_indicators desc limit 1),
  R.stars
from "gitsec.Repo" R order by R.created_at desc limit 10;
          full_name           |      to_timestamp      |   crypto_library   |
             num_indicators | stars
------------------------------+------------------------+--------------------+---------------

 rubinius/rubinius            | 2008-01-12 00:00:00-08 | jedisct1/libsodium |
               156 |  2803
 IoLanguage/io                | 2008-02-22 00:00:00-08 |                    |
                   |  1965
 beanstalkd/beanstalkd        | 2008-03-31 00:00:00-07 |                    |
                   |  4838
 codahale/bcrypt-ruby         | 2008-05-07 00:00:00-07 |                    |
                   |  1408
 davidfstr/rdiscount          | 2008-05-30 00:00:00-07 |                    |
                   |   761
 PromyLOPh/pianobar           | 2008-06-10 00:00:00-07 | gpg/libgcrypt      |
                 3 |  1524
 toland/qlmarkdown            | 2008-07-08 00:00:00-07 |                    |
                   |  2576
 taf2/curb                    | 2008-07-15 00:00:00-07 | openssl/openssl    |
                 2 |  1161
 FreeRADIUS/freeradius-server | 2008-07-20 00:00:00-07 | openssl/openssl    |
                55 |   881
 git/git                      | 2008-07-23 00:00:00-07 | openssl/openssl    |
                 2 | 25107
(10 rows)
...
```

Which crypto libraries do the newest github repositories use?

```
select
  R.full_name,
  to_timestamp(R.created_at),
  (select crypto_library from "gitsec.UsesCryptoLibrary" U inner join "gitsec.
     CryptoProvidingRepo" P on U.crypto_library=P.repo_name where U.library=R.
     full_name and (P.hide_from_ui <> '1' or P.hide_from_ui is null) and
     num_indicators > 0 order by num_indicators desc limit 1),
  (select num_indicators from "gitsec.UsesCryptoLibrary" U inner join "gitsec.
     CryptoProvidingRepo" P on U.crypto_library=P.repo_name where U.library=R.
     full_name and (P.hide_from_ui <> '1' or P.hide_from_ui is null) and
     num_indicators > 0 order by num_indicators desc limit 1),
  R.stars
from "gitsec.Repo" R order by R.created_at desc limit 10;
             full_name              |      to_timestamp      | crypto_library  |
                num_indicators | stars
------------------------------------+------------------------+-----------------+------------

 Meituan-Dianping/Logan             | 2018-09-26 00:00:00-07 | ARMmbed/mbedtls |
               134 |  1156
 rianhunter/wasmjit                 | 2018-09-25 00:00:00-07 |                 |
                   |   780
 wangzheng0822/algo                 | 2018-09-24 00:00:00-07 |                 |
                   |  1371
 Tencent/MMKV                       | 2018-09-17 00:00:00-07 | openssl/openssl |
                18 |  5519
 xoreaxeaxeax/rosenbridge           | 2018-08-09 00:00:00-07 |                 |
                   |  1845
 aergoio/litetree                   | 2018-08-06 00:00:00-07 |                 |
                   |  1283
 BlindMindStudios/StarRuler2-Source | 2018-07-17 00:00:00-07 | openssl/openssl |
                 5 |  1039
 rhysd/vim.wasm                     | 2018-07-01 00:00:00-07 |                 |
                   |  3130
 hnes/libaco                        | 2018-06-30 00:00:00-07 |                 |
                   |  1509
 TelegramMessenger/MTProxy          | 2018-05-29 00:00:00-07 | openssl/openssl |
                11 |  1588
(10 rows)
\begin{lstlisting}
Which organizations house the most crypto libraries?\\
\begin{lstlisting}
select O.name,R.full_name from "gitsec.Organization" O inner join "gitsec.Repo" R on R
   .organization_name=O.name inner join "gitsec.CryptoProvidingRepo" C on C.repo_name
   =R.full_name;
      name       |      full_name
-----------------+---------------------
 no_organization | jedisct1/libsodium
 wolfSSL         | wolfSSL/wolfssl
 gnutls          | gnutls/gnutls
 libtom          | libtom/libtomcrypt
 ARMmbed         | ARMmbed/mbedtls
 openssl         | openssl/openssl
 breadwallet     | breadwallet/nettle
 nss-dev         | nss-dev/nss
 gpg             | gpg/libgcrypt
 matrixssl       | matrixssl/matrixssl
(10 rows)
```

Original questions:

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 scraping 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?