Jared Jones
CSCE 612
Hw1p3 Report

1. (5 pts) Briefly explain your code architecture and lessons learned. Using Part 3, show a complete trace with 1M input URLs.

My software was structured to allow easy multithreading through the use of simplified function calls that automate the url parsing and information retrieval steps. The main function in 612Hw1.cpp does input checking and deals with urls or an input file. From there all the urls(s) depending on if a file or single url was given, are loaded into a one-way thread safe queue. This queue is a wrapper class around std::vector that is only thread safe one way because it is loaded up by one thread at the beginning, and then all of the consumers take items out of it at the same time, which only requires the pop functionality to be thread safe. From there, the required number of threads are created all sharing the same winsock and are released and begin dequeuing items from the queue and connecting / parsing them. The winsock class has been implemented to be shared between all the threads and is used to collect the relevant data on the progress the threads are making throughout the runtime. This is done with a variety of variables and mutexes in order to protect each item respectively in order to allow for minimal slowdowns. As the consumer threads are connecting and checking links, they are waiting on mutexes and updating the counts of urls that have successful host checks, DNS resolutions, IP checks, connections, etc etc. This is all outputted by the stats thread that is also started if a file input is given and will report the progress of the program. I rewrote a lot of the given winsock example to use string spans (a char* and length struct) because I was having trouble with some web pages not displaying properly because of std output with char*. Winsock also uses a headerParser class I created that takes in http headers, separates them, and pulls out the status code using regex (which was fun to learn and figure out how to use). This assignment also was a great tool on mutexes and edge cases in parallel programming in addition to teaching about what all goes into browsing the web and accessing webpages.

Trace with 1M input URLs

```
Opened URL-input-1M.txt with size 66152005
[  0] 10000 Q1000004 E      0 H      0 D      0 I    0 R    0 C    0 L   0K
      *** crawling 0.0 pps @ 0.0 Mbps
[  2] 10000 Q992067 E   7937 H   7935 D   7000 I 5913 R 4949 C  554 L   9K
      *** crawling 3968.5 pps @ 26.3 Mbps
[  4] 10000 Q940203 E  59801 H  17885 D  15921 I10794 R 9152 C 1342 L  21K
      *** crawling 25932.0 pps @ 30.6 Mbps
[  6] 10000 Q861936 E 138068 H  28163 D  24770 I15015 R12669 C 2149 L  32K
      *** crawling 39133.5 pps @ 34.7 Mbps
[  8] 10000 Q786200 E 213804 H  38185 D  33560 I18903 R15945 C 2939 L  44K
      *** crawling 37868.0 pps @ 34.5 Mbps
[ 10] 10000 Q711180 E 288824 H  48318 D  42500 I22700 R19152 C 3747 L  61K
      *** crawling 37510.0 pps @ 40.8 Mbps
[ 12] 10000 Q636248 E 363756 H  58112 D  51151 I26492 R22381 C 4564 L  69K
      *** crawling 37466.0 pps @ 22.5 Mbps
```

[14] 10000 Q570876 E 429128 H 66131 D 58216 I30081 R25407 C 5360 L 90K
    *** crawling 32686.0 pps @ 31.7 Mbps
[16] 10000 Q511783 E 488221 H 73130 D 64612 I34353 R29185 C 6329 L 102K
    *** crawling 29546.5 pps @ 26.0 Mbps
[18] 10000 Q446046 E 553958 H 80091 D 70882 I38393 R32784 C 7210 L 114K
    *** crawling 32868.5 pps @ 34.0 Mbps
[20] 10000 Q386411 E 613593 H 86993 D 77199 I42499 R36393 C 8152 L 124K
    *** crawling 29817.5 pps @ 30.1 Mbps
[22] 10000 Q336242 E 663762 H 93548 D 83204 I46296 R39818 C 9039 L 133K
    *** crawling 25084.5 pps @ 23.5 Mbps
[24] 10000 Q294117 E 705887 H 100065 D 89097 I50023 R43182 C 9962 L 143K
    *** crawling 21062.5 pps @ 28.0 Mbps
[26] 10000 Q248520 E 751484 H 107189 D 95543 I54074 R46756 C10884 L 155K
    *** crawling 22798.5 pps @ 35.2 Mbps
[28] 10000 Q208385 E 791619 H 112450 D 100232 I56909 R49260 C11528 L 165K
    *** crawling 20067.5 pps @ 21.9 Mbps
[30] 10000 Q200303 E 799701 H 113456 D 101138 I57468 R49760 C11660 L 174K
    *** crawling 4041.0 pps @ 14.7 Mbps
[32] 10000 Q190046 E 809958 H 114741 D 102363 I58193 R50382 C11798 L 187K
    *** crawling 5128.5 pps @ 37.9 Mbps
[34] 10000 Q181250 E 818754 H 115734 D 103270 I58759 R50914 C11955 L 194K
    *** crawling 4398.0 pps @ 16.7 Mbps
[36] 10000 Q171576 E 828428 H 116760 D 104203 I59325 R51427 C12084 L 200K
    *** crawling 4837.0 pps @ 15.5 Mbps
[38] 10000 Q164201 E 835803 H 117685 D 105036 I59843 R51890 C12215 L 208K
    *** crawling 3687.5 pps @ 17.2 Mbps
[40] 10000 Q156352 E 843652 H 118748 D 106003 I60420 R52391 C12335 L 213K
    *** crawling 3924.5 pps @ 16.4 Mbps
[42] 10000 Q149830 E 850174 H 119678 D 106854 I60957 R52869 C12458 L 220K
    *** crawling 3261.0 pps @ 16.5 Mbps
[44] 10000 Q143035 E 856969 H 120664 D 107734 I61468 R53314 C12557 L 226K
    *** crawling 3397.5 pps @ 16.7 Mbps
[46] 10000 Q136357 E 863647 H 121630 D 108587 I61974 R53757 C12682 L 231K
    *** crawling 3339.0 pps @ 13.9 Mbps
[48] 10000 Q130274 E 869730 H 122462 D 109330 I62431 R54174 C12800 L 238K
    *** crawling 3041.5 pps @ 18.0 Mbps
[50] 10000 Q121793 E 878211 H 123672 D 110395 I63062 R54724 C12947 L 243K
    *** crawling 4240.5 pps @ 17.6 Mbps
[52] 10000 Q113387 E 886617 H 124770 D 111363 I63598 R55192 C13078 L 249K
    *** crawling 4203.0 pps @ 15.7 Mbps
[54] 10000 Q105647 E 894357 H 125773 D 112282 I64129 R55677 C13203 L 255K
    *** crawling 3870.0 pps @ 20.7 Mbps
[56] 10000 Q 99526 E 900478 H 126531 D 112955 I64507 R56023 C13320 L 261K
    *** crawling 3060.5 pps @ 21.3 Mbps
[58] 10000 Q 92609 E 907395 H 127375 D 113733 I64933 R56379 C13418 L 271K
    *** crawling 3458.5 pps @ 16.8 Mbps
[60] 10000 Q 84049 E 915955 H 128395 D 114653 I65459 R56876 C13533 L 277K
    *** crawling 4280.0 pps @ 17.5 Mbps
[62] 10000 Q 74624 E 925380 H 129527 D 115671 I65998 R57340 C13642 L 283K

*** crawling 4712.5 pps @ 17.8 Mbps
[64] 10000 Q 67320 E 932684 H 130496 D 116527 I66488 R57782 C13759 L 290K
*** crawling 3652.0 pps @ 12.2 Mbps
[66] 10000 Q 61271 E 938733 H 131327 D 117264 I66911 R58149 C13857 L 295K
*** crawling 3024.5 pps @ 14.3 Mbps
[68] 10000 Q 56200 E 943804 H 131974 D 117865 I67244 R58454 C13945 L 298K
*** crawling 2535.5 pps @ 6.7 Mbps
[70] 10000 Q 50163 E 949841 H 132687 D 118509 I67596 R58750 C14018 L 301K
*** crawling 3018.5 pps @ 13.2 Mbps
[72] 10000 Q 43190 E 956814 H 133573 D 119270 I67988 R59084 C14101 L 312K
*** crawling 3486.5 pps @ 21.0 Mbps
[74] 10000 Q 35371 E 964633 H 134545 D 120139 I68405 R59458 C14195 L 316K
*** crawling 3909.5 pps @ 13.9 Mbps
[76] 10000 Q 28049 E 971955 H 135515 D 120986 I68817 R59792 C14294 L 321K
*** crawling 3661.0 pps @ 19.3 Mbps
[78] 10000 Q 20466 E 979538 H 136520 D 121840 I69183 R60105 C14394 L 327K
*** crawling 3791.5 pps @ 16.3 Mbps
[80] 10000 Q 12784 E 987220 H 137501 D 122703 I69547 R60397 C14465 L 334K
*** crawling 3841.0 pps @ 17.5 Mbps
[82] 10000 Q 6202 E 993802 H 138419 D 123499 I69858 R60656 C14547 L 339K
*** crawling 3291.0 pps @ 12.8 Mbps
[84] 9867 Q   0 E1000004 H 139302 D 124322 I70169 R60941 C14655 L 350K
*** crawling 3101.0 pps @ 24.2 Mbps
[86] 9610 Q   0 E1000004 H 139302 D 124348 I70182 R60960 C14676 L 362K
*** crawling 0.0 pps @ 25.0 Mbps
[88] 9404 Q   0 E1000004 H 139302 D 124348 I70182 R60966 C14682 L 370K
*** crawling 0.0 pps @ 28.4 Mbps
[90] 9196 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14683 L 379K
*** crawling 0.0 pps @ 20.5 Mbps
[92] 9006 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14683 L 388K
*** crawling 0.0 pps @ 23.9 Mbps
[94] 8811 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 398K
*** crawling 0.0 pps @ 22.5 Mbps
[96] 8618 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 412K
*** crawling 0.0 pps @ 21.6 Mbps
[98] 8437 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 418K
*** crawling 0.0 pps @ 23.0 Mbps
[100] 8305 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 428K
*** crawling 0.0 pps @ 20.3 Mbps
[102] 8171 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 434K
*** crawling 0.0 pps @ 17.7 Mbps
[104] 8051 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 441K
*** crawling 0.0 pps @ 17.3 Mbps
[106] 7932 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 450K
*** crawling 0.0 pps @ 18.4 Mbps
[108] 7800 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 456K
*** crawling 0.0 pps @ 20.8 Mbps
[110] 7668 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 464K
*** crawling 0.0 pps @ 21.6 Mbps

[112] 7542 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 473K
*** crawling 0.0 pps @ 29.2 Mbps
[114] 7417 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 485K
*** crawling 0.0 pps @ 27.9 Mbps
[116] 7290 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 496K
*** crawling 0.0 pps @ 24.9 Mbps
[118] 7173 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 504K
*** crawling 0.0 pps @ 18.6 Mbps
[120] 7061 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 511K
*** crawling 0.0 pps @ 16.9 Mbps
[122] 6943 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 519K
*** crawling 0.0 pps @ 19.9 Mbps
[124] 6827 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 528K
*** crawling 0.0 pps @ 23.0 Mbps
[126] 6711 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 535K
*** crawling 0.0 pps @ 15.1 Mbps
[128] 6595 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 542K
*** crawling 0.0 pps @ 20.1 Mbps
[130] 6479 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 549K
*** crawling 0.0 pps @ 13.8 Mbps
[132] 6364 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 558K
*** crawling 0.0 pps @ 17.6 Mbps
[134] 6251 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 565K
*** crawling 0.0 pps @ 14.8 Mbps
[136] 6135 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 573K
*** crawling 0.0 pps @ 18.8 Mbps
[138] 6021 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 579K
*** crawling 0.0 pps @ 17.9 Mbps
[140] 5909 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 586K
*** crawling 0.0 pps @ 14.8 Mbps
[142] 5797 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 591K
*** crawling 0.0 pps @ 17.2 Mbps
[144] 5685 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 596K
*** crawling 0.0 pps @ 12.3 Mbps
[146] 5572 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 604K
*** crawling 0.0 pps @ 16.3 Mbps
[148] 5460 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 609K
*** crawling 0.0 pps @ 13.9 Mbps
[150] 5350 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 617K
*** crawling 0.0 pps @ 14.2 Mbps
[152] 5241 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 625K
*** crawling 0.0 pps @ 16.8 Mbps
[154] 5132 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 631K
*** crawling 0.0 pps @ 15.6 Mbps
[156] 5022 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 637K
*** crawling 0.0 pps @ 16.2 Mbps
[158] 4912 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 644K
*** crawling 0.0 pps @ 13.9 Mbps
[160] 4801 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 650K

*** crawling 0.0 pps @ 12.8 Mbps
[162] 4691 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 656K
*** crawling 0.0 pps @ 13.1 Mbps
[164] 4582 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 664K
*** crawling 0.0 pps @ 23.9 Mbps
[166] 4474 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 672K
*** crawling 0.0 pps @ 15.8 Mbps
[168] 4364 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 678K
*** crawling 0.0 pps @ 22.3 Mbps
[170] 4255 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 684K
*** crawling 0.0 pps @ 18.2 Mbps
[172] 4145 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 690K
*** crawling 0.0 pps @ 15.0 Mbps
[174] 4037 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 698K
*** crawling 0.0 pps @ 17.8 Mbps
[176] 3930 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 704K
*** crawling 0.0 pps @ 19.0 Mbps
[178] 3821 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 711K
*** crawling 0.0 pps @ 18.2 Mbps
[180] 3714 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 720K
*** crawling 0.0 pps @ 24.5 Mbps
[182] 3607 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 725K
*** crawling 0.0 pps @ 11.5 Mbps
[184] 3497 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 731K
*** crawling 0.0 pps @ 17.4 Mbps
[186] 3389 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 738K
*** crawling 0.0 pps @ 24.3 Mbps
[188] 3282 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 743K
*** crawling 0.0 pps @ 15.0 Mbps
[190] 3177 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 748K
*** crawling 0.0 pps @ 13.0 Mbps
[192] 3069 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 755K
*** crawling 0.0 pps @ 17.3 Mbps
[194] 2962 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 760K
*** crawling 0.0 pps @ 14.3 Mbps
[196] 2855 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 764K
*** crawling 0.0 pps @ 12.4 Mbps
[198] 2748 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 769K
*** crawling 0.0 pps @ 12.1 Mbps
[200] 2643 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 777K
*** crawling 0.0 pps @ 16.6 Mbps
[202] 2536 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 783K
*** crawling 0.0 pps @ 12.3 Mbps
[204] 2431 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 790K
*** crawling 0.0 pps @ 17.3 Mbps
[206] 2326 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 797K
*** crawling 0.0 pps @ 17.4 Mbps
[208] 2221 Q 0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 804K
*** crawling 0.0 pps @ 20.1 Mbps

[210] 2116 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 808K
       *** crawling 0.0 pps @ 14.4 Mbps
[212] 2012 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 815K
       *** crawling 0.0 pps @ 14.8 Mbps
[214] 1908 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 822K
       *** crawling 0.0 pps @ 13.1 Mbps
[216] 1803 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 829K
       *** crawling 0.0 pps @ 12.5 Mbps
[218] 1698 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 834K
       *** crawling 0.0 pps @ 16.0 Mbps
[220] 1594 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 840K
       *** crawling 0.0 pps @ 13.3 Mbps
[222] 1491 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 845K
       *** crawling 0.0 pps @ 18.5 Mbps
[224] 1390 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 853K
       *** crawling 0.0 pps @ 22.7 Mbps
[226] 1287 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 863K
       *** crawling 0.0 pps @ 24.7 Mbps
[228] 1184 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 869K
       *** crawling 0.0 pps @ 16.9 Mbps
[230] 1083 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 877K
       *** crawling 0.0 pps @ 15.7 Mbps
[232]   985 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 882K
       *** crawling 0.0 pps @ 14.8 Mbps
[234]   884 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 887K
       *** crawling 0.0 pps @ 16.1 Mbps
[236]   781 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 894K
       *** crawling 0.0 pps @ 15.1 Mbps
[238]   679 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 901K
       *** crawling 0.0 pps @ 17.1 Mbps
[240]   577 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 910K
       *** crawling 0.0 pps @ 18.2 Mbps
[242]   477 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 920K
       *** crawling 0.0 pps @ 21.1 Mbps
[244]   376 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 925K
       *** crawling 0.0 pps @ 14.0 Mbps
[246]   274 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 935K
       *** crawling 0.0 pps @ 21.7 Mbps
[248]   176 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 943K
       *** crawling 0.0 pps @ 19.0 Mbps
[250]   74 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 949K
       *** crawling 0.0 pps @ 15.5 Mbps
[252]   0 Q   0 E1000004 H 139302 D 124350 I70184 R60969 C14684 L 958K
       *** crawling 0.0 pps @ 16.3 Mbps

Extracted 1000004 URLs @ 1000004/s
Looked up 139302 DNS names @ 548/s
Attempted 70184 robots @ 276/s
Crawled 14684 pages @ 57/s (602.71 MB)

Parsed 958913 links @ 3775/s
HTTP codes: 2xx = 14684, 3xx = 39056, 4xx = 5465, 5xx = 727, other = 1

C:\Users\jared.jones280\Downloads\webcrawler-master (1)\webcrawler-master\x64\Release\612Hw1.exe (process 28828) exited with code 0.
To automatically close the console when debugging stops, enable Tools->Options->Debugging->Automatically close the console when debugging stops.
Press any key to close this window . . .

2. (5 pts) Across all pages that came back with a 2xx code, calculate the average number of
HTML links (i.e., out-neighbors) found by the parser. Estimate the size of Google's webgraph (in terms of edges and bytes it occupies on disk) assuming they crawl 1T (trillion) pages. A webgraph here would store each crawled node x and its out-neighbors (y1,y2, …) using adjacency lists, where URLs are represented by 64-bit hashes.

Webpages with 2xx = 14684, containing a total of 958913 links.
958913 links / 14684 pages = 65 links per page.
Webgraph would contain 1 trillion links + 1 trillion * 65 (links per page link) = 66 trillion links total in 64 bit hashes
66 trillion * 64bits = 66x10^12 * 64 = 4.224x10^15 bits
4.224x10^15 bits / 8x10^12 (bits/TB) = 528 TB.

3. (5 pts) Determine the average page size in bytes (across all HTTP codes). Estimate the bandwidth (in Gbps) needed for Bing to crawl 10B pages a day.

602.71 MB in 14684 pages
602.71 / 14684 = .041045 MB/page = 41.045KB/page = 41045 Bytes/page
10B pages * 41045 Bytes/page = 410,450,000,000,000 Bytes
410,450,000,000,000 Bytes / (1x10^9 Bytes/GB) = 410,450 GBps * 8 = 3,283,600 Gbps

4. (5 pts) What is the probability that a link in the input file contains a unique host? What is
the probability that a unique host has a valid DNS record? What percentage of contacted sites had a 4xx robots file?

From the stats above:
Extracted 1000004 URLs @ 1000004/s
Looked up 139302 DNS names @ 548/s
Attempted 70184 robots @ 276/s
Crawled 14684 pages @ 57/s (602.71 MB)
Parsed 958913 links @ 3775/s
HTTP codes: 2xx = 14684, 3xx = 39056, 4xx = 5465, 5xx = 727, other = 1

Unique hosts = 139302 / 1000004 = 13.93%
Valid DNS if already unique host = 124350 / 139302 = 89.27%
Contacted sites with invalid robots file = 70184 - 60969 = 9215 / 70184 = 13.13%

5. (5 pts) How many of the crawled 2xx pages contain a hyperlink to our domain tamu.edu? How many of them originate from outside of TAMU? Explain how you obtained this information.

I made my program output the char* output from the HTMLParserBase into a output.txt file. This gave me a file with a list of all of the found links that were found in the crawled pages.

```
//put found links dump in output file
w->fileWrite.lock();
std::fstream file;
file.open("output.txt", std::ios_base::app | std::ios_base::in);
if (file.is_open()) {
        file << pageLinks<<std::endl;
}
file.close();
w->fileWrite.unlock();
```

After that I took my output file and ran some bash commands to get the following output.

jared@jared-Precision-5520:~/Documents$ grep "tamu.edu" output.txt
http://texasforestinfo.tamu.edu/

http://cotton.tamu.edu/index.html

http://eps.tamu.edu/

http://vpr.tamu.edu/

http://aglifesciences.tamu.edu/

http://regsci.tamu.edu/

http://soiltesting.tamu.edu/webpages/calendar.html

jared@jared-Precision-5520:~/Documents$ grep "tamu.edu" output.txt | wc -l
7

Which show 7 valid links that I went over with visual inspection. Then counting the lines there were 7 hyperlinks to the tamu.edu domain.

I also edited the file output line to include the original link (x) so then I could use grep -C to get the context around the lines and see from where the origin links came from.

```
file <<"[[" << x << "]]" << std::endl << pageLinks << std::endl;
```

Seeing the output below with context you can see that the origin links to all the links with tamu.edu also contain tamu.edu so therefore all of them originate from within the tamu.edu system.

```
jared@jared-Precision-5520:~/Documents$ grep -C 2 "tamu.edu" output.txt
[[http://clovisconcreteco.com/]]
[[http://cotton.tamu.edu/]]
http://cotton.tamu.edu/index.html
[[http://www.gtjaqh.com/]]
http://www.gtjafxgl.com/
--
[[http://pwoodford.net/blog/]]
http://pwoodford.net/blog
[[http://otc.tamu.edu/]]
http://vpr.tamu.edu/
[[http://nextnavy.com/does-warfighting-first-put-ship-handling-second/]]
http://nextnavy.com/
--
[[http://olympionikvarazdat.ru/]]
http://olympionikvarazdat.ru/ov_modul_reklama_Vardan_1_0.jpg
[[http://scsdistance.tamu.edu/]]
http://aglifesciences.tamu.edu/
[[http://perrinelson.com/]]
http://perrinelson.com/
--
[[http://rachelkushner.com/flamethrowers.html]]
http://rachelkushner.com/books.html
[[http://regsci.tamu.edu/]]
http://regsci.tamu.edu/
[[http://streaming.osu.edu/WOSU-NPRnews-wm.htm]]
http://streaming.osu.edu/WOSU/leadin/wosu-am.wax
--
[[http://thelouvertureproject.org/index.php?title=Napol%C3%A9on_Bonaparte_Proclamation_on
_Saint-Domingue_(1799)]]
[[http://soiltesting.tamu.edu/]]
http://soiltesting.tamu.edu/webpages/calendar.html
[[http://seemedlikeagoodidea.ca/]]
```

--
[[http://www.google.dk/search?num=10&query=dansk]]
http://www.google.dk/?sa=X&ved=0ahUKEwiA-ISH7_71AhUQmHIEHScpAIoQOwgC
[[http://texasforestinfo.tamu.edu/]]
http://texasforestinfo.tamu.edu/
[[http://texaspoliceassociation.com/]]
http://texaspoliceassociation.com/index.php
--
[[http://www.josephbcastro.com/]]
http://www.josephbcastro.com/index.html
[[http://eps.tamu.edu/]]
http://eps.tamu.edu/
[[http://www.journeyofwater.co.za/]]
http://www.journeyofwater.co.za/search
Binary file output.txt matches