1. (5 pts) Briefly explain your code architecture and lessons learned. Using Part 3, show a complete trace with 1M input URLs.
   1. First we open the crawler class holding the critical sections needed, then a stats class, and allocated a helper class on the stack, that then the helper class popped from a shared queue of URL links, parsing things like port, path, and the host name, then checking with a unique ip, host, and dns. Then the helper class calls the connection to robots, and then reconnects with a head request. If at any point the connection, host, or any url is not correct the socket is deleted, and reallocated on the next iteration. The largest lesson learned was having a crawler call a heap shared helper is shared memory. This was painful to learn, and took around 10 hours to debug.
   2. TRACE:

Opened URL-input-1M.txt with size 66152005

[ 2] 500 Q 991279 E 8725 H 2050 D 1559 I 1030 R 187 C 130 L 0K

\*\*\* crawling 278.5 pps @ 10.6 Mbps

[ 4] 500 Q 983366 E 16638 H 3310 D 2633 I 1670 R 477 C 338 L 0K

\*\*\* crawling 301.8 pps @ 10.9 Mbps

[ 6] 500 Q 976895 E 23109 H 4291 D 3509 I 2197 R 684 C 475 L 0K

\*\*\* crawling 165.5 pps @ 5.5 Mbps

[ 8] 500 Q 971584 E 28420 H 5124 D 4298 I 2718 R 908 C 622 L 1K

\*\*\* crawling 149.9 pps @ 5.2 Mbps

[ 10] 500 Q 964400 E 35604 H 6160 D 5207 I 3334 R 1242 C 837 L 1K

\*\*\* crawling 192.1 pps @ 6.5 Mbps

[ 12] 500 Q 957615 E 42389 H 7092 D 6067 I 3904 R 1576 C 1050 L 1K

\*\*\* crawling 165.7 pps @ 5.4 Mbps

[ 14] 500 Q 951806 E 48198 H 7904 D 6785 I 4391 R 1914 C 1256 L 2K

\*\*\* crawling 154.1 pps @ 5.0 Mbps

[ 16] 500 Q 945877 E 54127 H 8706 D 7496 I 4886 R 2292 C 1519 L 2K

\*\*\* crawling 173.1 pps @ 5.6 Mbps

[ 18] 500 Q 941324 E 58680 H 9326 D 8062 I 5300 R 2683 C 1744 L 3K

\*\*\* crawling 153.5 pps @ 4.9 Mbps

[ 20] 500 Q 937983 E 62021 H 9786 D 8468 I 5611 R 3034 C 1970 L 4K

\*\*\* crawling 136.9 pps @ 4.9 Mbps

[ 22] 500 Q 935301 E 64703 H 10212 D 8837 I 5883 R 3298 C 2111 L 4K

\*\*\* crawling 91.5 pps @ 3.4 Mbps

[ 24] 500 Q 929759 E 70245 H 11045 D 9553 I 6400 R 3685 C 2388 L 5K

\*\*\* crawling 127.5 pps @ 4.6 Mbps

[ 26] 500 Q 922551 E 77453 H 12067 D 10429 I 7017 R 4078 C 2685 L 5K

\*\*\* crawling 122.5 pps @ 4.2 Mbps

[ 28] 500 Q 914425 E 85579 H 13109 D 11350 I 7672 R 4622 C 3001 L 6K

\*\*\* crawling 144.5 pps @ 4.9 Mbps

[ 30] 500 Q 907169 E 92835 H 14085 D 12194 I 8272 R 5136 C 3304 L 7K

\*\*\* crawling 116.8 pps @ 3.9 Mbps

[ 32] 500 Q 900571 E 99433 H 14979 D 12993 I 8841 R 5700 C 3642 L 8K

\*\*\* crawling 127.0 pps @ 4.7 Mbps

[ 34] 500 Q 893445 E 106559 H 15959 D 13881 I 9483 R 6290 C 3993 L 9K

\*\*\* crawling 129.6 pps @ 4.6 Mbps

[ 36] 500 Q 885943 E 114061 H 17047 D 14835 I 10166 R 7150 C 4521 L 10K

\*\*\* crawling 170.8 pps @ 6.0 Mbps

[ 38] 500 Q 878694 E 121310 H 18102 D 15744 I 10797 R 7825 C 4934 L 12K

\*\*\* crawling 133.4 pps @ 4.6 Mbps

[ 40] 500 Q 872309 E 127695 H 19091 D 16581 I 11405 R 8519 C 5314 L 13K

\*\*\* crawling 130.8 pps @ 4.3 Mbps

[ 42] 500 Q 865556 E 134448 H 20080 D 17498 I 12064 R 9375 C 5847 L 14K

\*\*\* crawling 158.1 pps @ 5.7 Mbps

[ 44] 500 Q 858684 E 141320 H 21131 D 18399 I 12673 R 10319 C 6382 L 16K

\*\*\* crawling 167.5 pps @ 6.1 Mbps

[ 46] 500 Q 852553 E 147451 H 22140 D 19284 I 13302 R 11139 C 6860 L 17K

\*\*\* crawling 138.2 pps @ 4.9 Mbps

[ 48] 500 Q 846352 E 153652 H 23113 D 20145 I 13908 R 11868 C 7309 L 18K

\*\*\* crawling 125.9 pps @ 4.5 Mbps

[ 50] 500 Q 840123 E 159881 H 24045 D 20962 I 14497 R 12644 C 7719 L 19K

\*\*\* crawling 120.5 pps @ 4.2 Mbps

[ 52] 500 Q 833116 E 166888 H 24976 D 21772 I 15053 R 13315 C 8133 L 21K

\*\*\* crawling 99.9 pps @ 3.5 Mbps

[ 54] 500 Q 825622 E 174382 H 25938 D 22620 I 15647 R 14249 C 8630 L 22K

\*\*\* crawling 136.1 pps @ 4.8 Mbps

[ 56] 500 Q 816544 E 183460 H 26916 D 23482 I 16240 R 15021 C 9141 L 23K

\*\*\* crawling 117.0 pps @ 4.0 Mbps

[ 58] 500 Q 808726 E 191278 H 27899 D 24337 I 16851 R 15994 C 9709 L 25K

\*\*\* crawling 135.5 pps @ 4.8 Mbps

[ 60] 500 Q 800686 E 199318 H 29013 D 25321 I 17546 R 17020 C 10376 L 26K

\*\*\* crawling 141.3 pps @ 4.9 Mbps

[ 62] 500 Q 793160 E 206844 H 30031 D 26211 I 18162 R 17739 C 10792 L 27K

\*\*\* crawling 93.2 pps @ 2.9 Mbps

[ 64] 500 Q 786961 E 213043 H 30985 D 27054 I 18738 R 18714 C 11325 L 29K

\*\*\* crawling 120.2 pps @ 4.0 Mbps

[ 66] 500 Q 779594 E 220410 H 31959 D 27887 I 19302 R 19506 C 11932 L 31K

\*\*\* crawling 116.3 pps @ 4.4 Mbps

[ 68] 500 Q 771364 E 228640 H 32948 D 28742 I 19899 R 20408 C 12401 L 32K

\*\*\* crawling 101.9 pps @ 3.6 Mbps

[ 70] 500 Q 763886 E 236118 H 34013 D 29663 I 20536 R 21331 C 13032 L 34K

\*\*\* crawling 116.7 pps @ 4.2 Mbps

[ 72] 500 Q 757973 E 242031 H 34944 D 30491 I 21096 R 22349 C 13604 L 36K

\*\*\* crawling 118.0 pps @ 4.2 Mbps

[ 74] 500 Q 752229 E 247775 H 35900 D 31338 I 21672 R 23613 C 14436 L 38K

\*\*\* crawling 147.5 pps @ 5.2 Mbps

[ 76] 500 Q 746587 E 253417 H 36879 D 32206 I 22282 R 24686 C 14998 L 39K

\*\*\* crawling 111.3 pps @ 4.0 Mbps

[ 78] 500 Q 740216 E 259788 H 37905 D 33111 I 22906 R 26037 C 15658 L 41K

\*\*\* crawling 130.8 pps @ 4.8 Mbps

[ 80] 500 Q 733769 E 266235 H 38967 D 34063 I 23568 R 27250 C 16363 L 43K

\*\*\* crawling 126.8 pps @ 4.6 Mbps

[ 82] 500 Q 727393 E 272611 H 40068 D 35032 I 24222 R 28586 C 17053 L 46K

\*\*\* crawling 136.4 pps @ 4.8 Mbps

[ 84] 500 Q 720692 E 279312 H 41223 D 36060 I 24900 R 29910 C 17787 L 48K

\*\*\* crawling 135.3 pps @ 4.7 Mbps

[ 86] 500 Q 714326 E 285678 H 42355 D 37083 I 25583 R 31229 C 18511 L 50K

\*\*\* crawling 118.7 pps @ 4.2 Mbps

[ 88] 500 Q 708914 E 291090 H 43311 D 37936 I 26171 R 32413 C 19229 L 52K

\*\*\* crawling 115.8 pps @ 4.0 Mbps

[ 90] 500 Q 703176 E 296828 H 44311 D 38808 I 26758 R 33708 C 19980 L 54K

\*\*\* crawling 122.0 pps @ 4.2 Mbps

[ 92] 500 Q 697596 E 302408 H 45247 D 39644 I 27311 R 35082 C 20778 L 56K

\*\*\* crawling 124.8 pps @ 4.3 Mbps

[ 94] 500 Q 691903 E 308101 H 46212 D 40510 I 27906 R 36453 C 21654 L 59K

\*\*\* crawling 132.4 pps @ 4.7 Mbps

[ 96] 500 Q 687003 E 313001 H 47022 D 41239 I 28407 R 37756 C 22422 L 61K

\*\*\* crawling 113.4 pps @ 3.9 Mbps

[ 98] 500 Q 681736 E 318268 H 47818 D 41943 I 28893 R 38915 C 23037 L 63K

\*\*\* crawling 95.9 pps @ 3.3 Mbps

[100] 500 Q 676149 E 323855 H 48744 D 42739 I 29413 R 40210 C 23777 L 66K

\*\*\* crawling 111.9 pps @ 3.8 Mbps

[102] 500 Q 670910 E 329094 H 49635 D 43536 I 29950 R 41320 C 24462 L 68K

\*\*\* crawling 90.9 pps @ 3.2 Mbps

[104] 500 Q 665657 E 334347 H 50516 D 44320 I 30489 R 42576 C 25096 L 70K

\*\*\* crawling 97.6 pps @ 3.5 Mbps

[106] 500 Q 659803 E 340201 H 51559 D 45236 I 31109 R 44306 C 25979 L 72K

\*\*\* crawling 127.2 pps @ 4.3 Mbps

[108] 500 Q 653395 E 346609 H 52605 D 46145 I 31713 R 45923 C 26942 L 75K

\*\*\* crawling 124.2 pps @ 4.4 Mbps

[110] 500 Q 645788 E 354216 H 53719 D 47148 I 32340 R 47310 C 27802 L 78K

\*\*\* crawling 111.5 pps @ 4.0 Mbps

[112] 500 Q 638491 E 361513 H 54787 D 48108 I 32991 R 48990 C 28669 L 80K

\*\*\* crawling 122.2 pps @ 4.3 Mbps

[114] 500 Q 630564 E 369440 H 55755 D 48930 I 33554 R 50628 C 29667 L 84K

\*\*\* crawling 125.0 pps @ 4.5 Mbps

[116] 500 Q 622440 E 377564 H 56737 D 49795 I 34137 R 52515 C 30795 L 87K

\*\*\* crawling 140.4 pps @ 5.2 Mbps

[118] 500 Q 613463 E 386541 H 57863 D 50805 I 34784 R 54493 C 31839 L 90K

\*\*\* crawling 136.9 pps @ 5.0 Mbps

[120] 500 Q 604397 E 395607 H 59034 D 51854 I 35465 R 56497 C 33020 L 95K

\*\*\* crawling 146.3 pps @ 5.3 Mbps

[122] 500 Q 593840 E 406164 H 60184 D 52882 I 36141 R 58348 C 34107 L 98K

\*\*\* crawling 133.6 pps @ 4.8 Mbps

[124] 500 Q 583839 E 416165 H 61374 D 53926 I 36813 R 60355 C 35296 L 102K

\*\*\* crawling 142.5 pps @ 5.0 Mbps

[126] 500 Q 575187 E 424817 H 62474 D 54898 I 37423 R 61950 C 36108 L 104K

\*\*\* crawling 105.4 pps @ 3.7 Mbps

[128] 500 Q 567883 E 432121 H 63499 D 55841 I 38033 R 63557 C 37014 L 107K

\*\*\* crawling 110.8 pps @ 3.9 Mbps

[130] 500 Q 560398 E 439606 H 64533 D 56752 I 38637 R 65517 C 38078 L 111K

\*\*\* crawling 125.8 pps @ 4.7 Mbps

[132] 500 Q 553342 E 446662 H 65508 D 57565 I 39136 R 67124 C 38997 L 114K

\*\*\* crawling 101.2 pps @ 3.7 Mbps

[134] 500 Q 547604 E 452400 H 66363 D 58310 I 39619 R 68198 C 39692 L 116K

\*\*\* crawling 74.9 pps @ 2.6 Mbps

[136] 500 Q 541249 E 458755 H 67215 D 59082 I 40118 R 69537 C 40407 L 118K

\*\*\* crawling 80.8 pps @ 2.8 Mbps

[138] 500 Q 533074 E 466930 H 68248 D 60007 I 40706 R 71216 C 41317 L 121K

\*\*\* crawling 102.7 pps @ 3.5 Mbps

[140] 500 Q 525369 E 474635 H 69276 D 60922 I 41264 R 73078 C 42295 L 123K

\*\*\* crawling 113.5 pps @ 4.0 Mbps

[142] 500 Q 515775 E 484229 H 70342 D 61855 I 41851 R 74905 C 43301 L 126K

\*\*\* crawling 109.9 pps @ 3.8 Mbps

[144] 500 Q 503340 E 496664 H 71465 D 62870 I 42507 R 77109 C 44526 L 130K

\*\*\* crawling 133.2 pps @ 4.6 Mbps

[146] 500 Q 489390 E 510614 H 72669 D 63920 I 43167 R 79313 C 45670 L 133K

\*\*\* crawling 126.6 pps @ 4.3 Mbps

[148] 500 Q 479454 E 520550 H 73722 D 64860 I 43754 R 81276 C 46847 L 137K

\*\*\* crawling 117.5 pps @ 4.1 Mbps

[150] 500 Q 470202 E 529802 H 74922 D 65939 I 44409 R 83187 C 47839 L 141K

\*\*\* crawling 109.2 pps @ 3.7 Mbps

[152] 500 Q 461219 E 538785 H 75992 D 66880 I 45000 R 85139 C 48939 L 144K

\*\*\* crawling 112.9 pps @ 4.0 Mbps

[154] 500 Q 452115 E 547889 H 77116 D 67888 I 45641 R 87405 C 50098 L 147K

\*\*\* crawling 124.1 pps @ 4.3 Mbps

[156] 500 Q 444919 E 555085 H 78165 D 68807 I 46212 R 89507 C 51460 L 151K

\*\*\* crawling 120.0 pps @ 4.2 Mbps

[158] 500 Q 438369 E 561635 H 79174 D 69697 I 46737 R 91255 C 52494 L 154K

\*\*\* crawling 100.6 pps @ 3.5 Mbps

[160] 500 Q 432722 E 567282 H 80110 D 70491 I 47184 R 92905 C 53339 L 156K

\*\*\* crawling 84.3 pps @ 2.9 Mbps

[162] 500 Q 427643 E 572361 H 80975 D 71226 I 47582 R 94407 C 54157 L 159K

\*\*\* crawling 77.0 pps @ 2.7 Mbps

[164] 500 Q 421480 E 578524 H 81908 D 72007 I 48003 R 96000 C 55031 L 161K

\*\*\* crawling 89.7 pps @ 3.1 Mbps

[166] 500 Q 414946 E 585058 H 82795 D 72768 I 48396 R 97499 C 55793 L 164K

\*\*\* crawling 76.5 pps @ 2.7 Mbps

[168] 500 Q 408262 E 591742 H 83661 D 73521 I 48796 R 99046 C 56667 L 167K

\*\*\* crawling 78.7 pps @ 2.7 Mbps

[170] 500 Q 401170 E 598834 H 84631 D 74353 I 49181 R 100905 C 57800 L 170K

\*\*\* crawling 96.6 pps @ 3.5 Mbps

[172] 500 Q 394547 E 605457 H 85571 D 75196 I 49583 R 102586 C 58916 L 174K

\*\*\* crawling 89.9 pps @ 3.2 Mbps

[174] 500 Q 387044 E 612960 H 86634 D 76141 I 49959 R 103884 C 59487 L 176K

\*\*\* crawling 59.5 pps @ 2.0 Mbps

[176] 500 Q 378995 E 621009 H 87777 D 77147 I 50375 R 105366 C 60417 L 179K

\*\*\* crawling 79.9 pps @ 2.8 Mbps

[178] 500 Q 370493 E 629511 H 88909 D 78115 I 50777 R 107028 C 61365 L 182K

\*\*\* crawling 82.1 pps @ 3.0 Mbps

[180] 500 Q 360708 E 639296 H 90139 D 79161 I 51204 R 108732 C 62471 L 186K

\*\*\* crawling 88.0 pps @ 3.1 Mbps

[182] 500 Q 352245 E 647759 H 91294 D 80140 I 51598 R 110105 C 63209 L 188K

\*\*\* crawling 62.6 pps @ 2.3 Mbps

[184] 500 Q 343986 E 656018 H 92517 D 81157 I 51997 R 111906 C 64261 L 191K

\*\*\* crawling 83.4 pps @ 2.8 Mbps

[186] 500 Q 334460 E 665544 H 93920 D 82366 I 52488 R 113579 C 65227 L 194K

\*\*\* crawling 82.8 pps @ 2.9 Mbps

[188] 500 Q 325023 E 674981 H 95310 D 83557 I 52971 R 115488 C 66202 L 197K

\*\*\* crawling 88.0 pps @ 3.0 Mbps

[190] 500 Q 317411 E 682593 H 96439 D 84537 I 53360 R 117043 C 66897 L 199K

\*\*\* crawling 66.3 pps @ 2.3 Mbps

[192] 500 Q 310158 E 689846 H 97492 D 85446 I 53742 R 118549 C 67806 L 202K

\*\*\* crawling 72.9 pps @ 2.5 Mbps

[194] 500 Q 303876 E 696128 H 98413 D 86217 I 54070 R 120005 C 68768 L 205K

\*\*\* crawling 67.6 pps @ 2.3 Mbps

[196] 500 Q 297009 E 702995 H 99449 D 87108 I 54422 R 121552 C 69660 L 207K

\*\*\* crawling 69.2 pps @ 2.5 Mbps

[198] 500 Q 289672 E 710332 H 100500 D 88010 I 54775 R 122939 C 70455 L 209K

\*\*\* crawling 62.3 pps @ 2.1 Mbps

[200] 500 Q 281625 E 718379 H 101670 D 89007 I 55154 R 124927 C 71649 L 213K

\*\*\* crawling 91.7 pps @ 3.1 Mbps

[202] 500 Q 272884 E 727120 H 102864 D 90037 I 55531 R 126422 C 72550 L 215K

\*\*\* crawling 68.8 pps @ 2.4 Mbps

[204] 500 Q 263824 E 736180 H 104138 D 91134 I 55913 R 127816 C 73323 L 218K

\*\*\* crawling 62.1 pps @ 2.1 Mbps

[206] 500 Q 255594 E 744410 H 105255 D 92120 I 56292 R 129204 C 74217 L 221K

\*\*\* crawling 63.8 pps @ 2.2 Mbps

[208] 500 Q 248174 E 751830 H 106304 D 93006 I 56636 R 130868 C 75189 L 223K

\*\*\* crawling 69.3 pps @ 2.4 Mbps

[210] 500 Q 240907 E 759097 H 107313 D 93884 I 56942 R 132123 C 75890 L 226K

\*\*\* crawling 53.8 pps @ 1.9 Mbps

[212] 500 Q 232651 E 767353 H 108547 D 94946 I 57338 R 134285 C 77182 L 230K

\*\*\* crawling 90.0 pps @ 3.3 Mbps

[214] 500 Q 224053 E 775951 H 109978 D 96178 I 57758 R 136057 C 78111 L 233K

\*\*\* crawling 71.1 pps @ 2.4 Mbps

[216] 500 Q 215369 E 784635 H 111301 D 97337 I 58174 R 137735 C 79062 L 236K

\*\*\* crawling 69.1 pps @ 2.5 Mbps

[218] 500 Q 206480 E 793524 H 112524 D 98388 I 58555 R 139984 C 80447 L 241K

\*\*\* crawling 92.1 pps @ 3.3 Mbps

[220] 500 Q 198089 E 801915 H 113628 D 99372 I 58903 R 141411 C 81233 L 243K

\*\*\* crawling 56.6 pps @ 2.0 Mbps

[222] 500 Q 190835 E 809169 H 114746 D 100337 I 59253 R 142635 C 81932 L 244K

\*\*\* crawling 49.6 pps @ 1.7 Mbps

[224] 500 Q 183005 E 816999 H 115939 D 101357 I 59610 R 143982 C 82800 L 247K

\*\*\* crawling 59.7 pps @ 2.1 Mbps

[226] 500 Q 174403 E 825601 H 117190 D 102411 I 59990 R 145937 C 83824 L 250K

\*\*\* crawling 74.5 pps @ 2.5 Mbps

[228] 500 Q 165280 E 834724 H 118535 D 103568 I 60426 R 147720 C 84807 L 253K

\*\*\* crawling 69.0 pps @ 2.4 Mbps

[230] 500 Q 155259 E 844745 H 119858 D 104686 I 60825 R 149465 C 85950 L 256K

\*\*\* crawling 72.6 pps @ 2.4 Mbps

[232] 500 Q 145821 E 854183 H 121126 D 105716 I 61217 R 151591 C 87065 L 259K

\*\*\* crawling 78.9 pps @ 2.7 Mbps

[234] 500 Q 137559 E 862445 H 122196 D 106575 I 61541 R 153334 C 87924 L 262K

\*\*\* crawling 62.5 pps @ 2.3 Mbps

[236] 500 Q 130027 E 869977 H 123156 D 107374 I 61833 R 154787 C 88725 L 264K

\*\*\* crawling 51.7 pps @ 2.0 Mbps

[238] 500 Q 121372 E 878632 H 124181 D 108207 I 62144 R 156552 C 89865 L 268K

\*\*\* crawling 68.9 pps @ 2.4 Mbps

[240] 500 Q 113870 E 886134 H 125319 D 109162 I 62532 R 158196 C 90901 L 270K

\*\*\* crawling 62.6 pps @ 2.0 Mbps

[243] 500 Q 104180 E 895824 H 126678 D 110280 I 62969 R 160168 C 92133 L 273K

\*\*\* crawling 72.5 pps @ 2.5 Mbps

[245] 500 Q 96027 E 903977 H 127948 D 111368 I 63414 R 162231 C 93419 L 277K

\*\*\* crawling 75.9 pps @ 2.7 Mbps

[247] 500 Q 87043 E 912961 H 129272 D 112468 I 63803 R 164115 C 94587 L 281K

\*\*\* crawling 69.2 pps @ 2.4 Mbps

[249] 500 Q 78497 E 921507 H 130549 D 113536 I 64175 R 165915 C 95560 L 283K

\*\*\* crawling 60.3 pps @ 2.1 Mbps

[251] 500 Q 71416 E 928588 H 131696 D 114505 I 64540 R 167732 C 96579 L 287K

\*\*\* crawling 62.8 pps @ 2.2 Mbps

[253] 500 Q 63097 E 936907 H 132921 D 115553 I 64920 R 169611 C 97652 L 290K

\*\*\* crawling 63.3 pps @ 2.2 Mbps

[255] 500 Q 53386 E 946618 H 134144 D 116580 I 65288 R 171548 C 98872 L 294K

\*\*\* crawling 70.2 pps @ 2.3 Mbps

[257] 500 Q 44554 E 955450 H 135291 D 117532 I 65635 R 173394 C 99939 L 298K

\*\*\* crawling 59.1 pps @ 2.2 Mbps

[259] 500 Q 38431 E 961573 H 136113 D 118222 I 65901 R 174787 C 100831 L 300K

\*\*\* crawling 47.4 pps @ 1.6 Mbps

[261] 500 Q 31206 E 968798 H 136996 D 118953 I 66169 R 176139 C 101522 L 302K

\*\*\* crawling 41.1 pps @ 1.5 Mbps

[263] 500 Q 23693 E 976311 H 137919 D 119732 I 66470 R 177782 C 102597 L 306K

\*\*\* crawling 56.4 pps @ 2.0 Mbps

[265] 500 Q 14735 E 985269 H 138954 D 120598 I 66791 R 179448 C 103606 L 308K

\*\*\* crawling 54.5 pps @ 1.8 Mbps

[267] 415 Q 0 E 1000004 H 139400 D 120993 I 66951 R 180576 C 104315 L 310K

\*\*\* crawling 38.7 pps @ 1.3 Mbps

[269] 365 Q 0 E 1000004 H 139400 D 120994 I 66951 R 180576 C 104330 L 310K

\*\*\* crawling 0.4 pps @ 0.0 Mbps

[271] 319 Q 0 E 1000004 H 139400 D 120994 I 66951 R 180600 C 104349 L 310K

\*\*\* crawling 0.8 pps @ 0.0 Mbps

[273] 281 Q 0 E 1000004 H 139400 D 120994 I 66951 R 180600 C 104349 L 310K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[275] 232 Q 0 E 1000004 H 139400 D 120994 I 66951 R 180600 C 104349 L 310K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[277] 181 Q 0 E 1000004 H 139400 D 120994 I 66951 R 180600 C 104349 L 310K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[279] 135 Q 0 E 1000004 H 139400 D 120994 I 66951 R 180600 C 104349 L 310K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[281] 110 Q 0 E 1000004 H 139400 D 120994 I 66951 R 180600 C 104349 L 310K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[283] 74 Q 0 E 1000004 H 139400 D 120994 I 66951 R 180600 C 104349 L 310K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[285] 32 Q 0 E 1000004 H 139400 D 120994 I 66951 R 180600 C 104349 L 310K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

Extracted 1000004 URLs @ 3484/s

Looked up 139400 DNS names @ 421/s

Attempted 180600 robots @ 629/s

Crawled 104349 pages @ 363/s (1689.48 MB)

Parsed 310590 links @ 1082/s

HTTP codes: 2xx = 37210, 3xx = 19920, 4xx = 46115, 5xx = 1104, other = 0

1. (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.
2. (5 pts) Determine the average page size in bytes (across all HTTP codes). Estimate the bandwidth (in Gbps) needed for Google to crawl 10B pages a day.
   1. =
3. (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?
   1. Unique Host: 13.93%
   2. Unique Host has a valid Dns: 100%
   3. Valid 4xx robots: 10.68%
4. (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. Examples of suitable links: irl.cs.tamu.edu/ afcerc.tamu.edu/index.html tamu.edu/ www.cse.tamu.edu/people 2 Examples of false-positives: tamu.edu.org/ www.x.com/tamu.edu/
   1. Crawled 2xx pages: 40
   2. Number of Crawled 2xx pages outside of TAMU: 38380
   3. I obtained this information from parsing each url after receiving the link buffer. After the fragment, query, port, and path was parsed, I was able to count using find function, and checked for the position after tamu.edu to have a null terminator or a backslash.