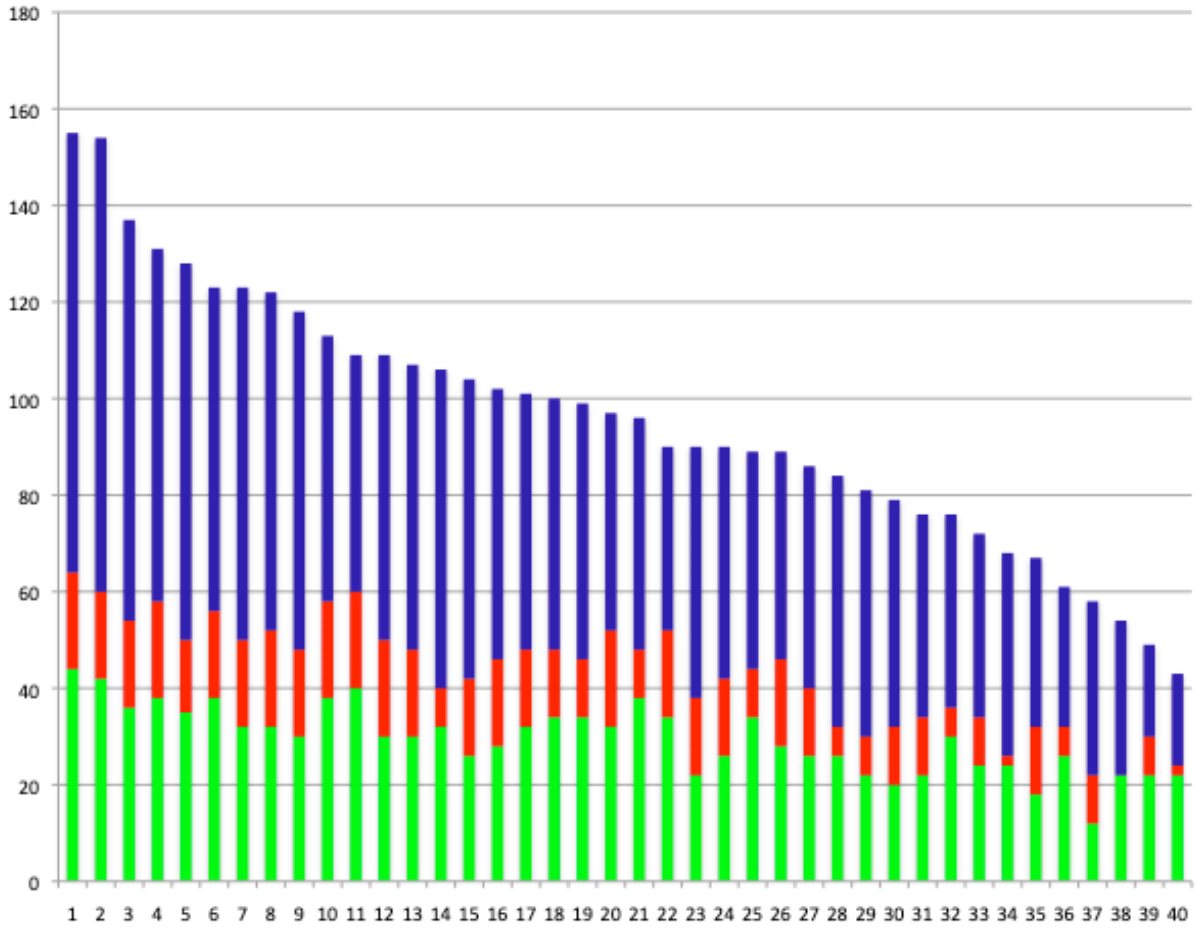
COS 109 Final, Fall 2018

I graded this myself. The median grade was 97, with quartiles at 113 and 76. The closest point of comparison would be the 2014 exam, where the corresponding numbers were 108, 127, and 88. The colors in the graphs below are for parts 1, 2 and 3, reading up from the bottom.



1. (50 points, 2 each) Short Answers. Circle the right answer or write it in the space provided.

(a) Alice wants to digitally sign a document and is wondering whether to use AES or DES. How would you advise her?

AES is clearly better **DES is clearly better** **both work well** **neither is any good**

Neither. AES and DES are symmetric or secret key algorithms. Since all parties must know the key to decrypt, neither can be used for signing: which of the people who know the key did the signing?

(b) ***Very roughly*** how much disk space would it take to store the source code for a version of the AES algorithm written in the C programming language?

a few KB **a few MB** **a few GB** **a few TB** **a few PB**

A few KB. I passed around this very listing one day; it fit comfortably on 4 or 5 pages. Not many got this right, perhaps missing the "source code" part.

(c) A news story says that there are probably about 100 billion planets in the Milky Way. If astronomers want to give each planet a unique number, how many bits would that number have?

37. 100 * 10^9 is about 2^7 * 2^30. Easy question, almost everyone got it.

(d) A major intellectual property event finally happened on January 1, 2019; in explanation, an article in *Smithsonian.com* said "We can blame Mickey Mouse for the long wait." Which of these general legal practice areas was the author referring to?

contract copyright entertainment patent tort trademark trade secret

Copyright. The 20-year delay in the 1998 copyright term extension act that protected Mickey Mouse (and others) expired on Jan 1.

(e) If you interview at a company that wants to tell you proprietary information about their business but prevent you from legally revealing it to anyone else, which one(s) of these might they require you to sign?

EULA FERPA GPL IANAL NDA PITA SOPA TLA

NDA. Most people got this.

(f) In December 2018, Amazon announced that it is designing and will fabricate its own CPU chips. Which of these companies is likely to be most affected by Amazon's action?

Apple Facebook Google Intel Microsoft Netflix

Intel. The other companies either already make their own chips or haven't said they would. Most people got this one too.

(g) Add these two binary numbers:

```
011010100100111
100101011011001
---------------
```

1000000000000000. The carry from the rightmost pair just carries across the whole thing. This is a good example of the observation that if you understand the material, you can just write down the answer. People who laboriously converted each binary number to decimal often got there in the end, but at the price of a lot of work, and sometimes they got the wrong answer.

(h) Suppose that I use SHA-3 to compute a cryptographic hash of a message that I am going to send to a friend. Then I change "bk" to "BK" in one place in the message and recompute the hash. What is the relationship between the first cryptographic hash value and the second?

the same 2 bits are different 2 bytes are different about half the bits differ every bit is different

About half the bits differ. Otherwise it would be far too easy to defeat the purpose of hashing. Not well handled, even though it was demonstrated in class and appears in the book.

(i) Put these names into chronological order of when they made the contribution(s) that caused them to be mentioned in COS 109, by writing the numbers 1 through 5 on them.

Tim Berners-Lee Jeff Bezos Bill Gates Ada Lovelace Lady Trumpington

Ada Lady T Bill Tim Jeff. A surprising number thought that the Web (Berners-Lee) predates the founding of Microsoft (Gates).

(j) Which of these would be the most appropriate name for someone working at a Certificate Authority?

Alice Bob Eve Mallory Trent

Trent, the trusted third party.

(k) Sort the following list of RGB colors, expressed in hexadecimal, into order of *increasing* amount of green. Label them from 1 (least green) to 5 (most green).

ACCEDE BEADED BEDDED DECADE DEFACE

3 1 4 2 5. The colors are numbers expressed in base 16, so A < B < C, etc., and CA < CE. Again, there's no need for arithmetic (and especially not for converting to decimal).

(l) Each of the following files is exactly 100 MB long and each contains typical information of the type indicated by the filename extension. Which one of these files will likely be smallest after Lempel-Ziv compression is applied to it?

`F.gif` `F.jpg` `F.mpg` `F.mp3` `F.png` `F.txt` `F.zip` no way to tell

F.txt. All of the others are already compressed so further compression is unlikely to do much.

(m) Circle the correct answer(s): Bletchley Park is where Alan Turing …

was born **spent his childhood** **invented the Turing machine**

wrote his PhD thesis **did his crypto work** **died** **is buried**

was knighted by Queen Elizabeth II **was pardoned by Queen Elizabeth II**

Did his crypto, developing special-purpose computers to decrypt German Enigma traffic, as discussed at length in class, even with some of the instructor's photographs.

(n) What kind or level of programming language is being described in this excerpt from David Auerbach's 2018 book *Bitwise: A Life in Code*? "Store this number here, retrieve this number from there, add or subtract these two numbers, and branch to different bits of code depending on some condition or other."

Assembly. Most people got this.

(o) Which of these entities would I have to deal with if I want to acquire radio frequency spectrum for a new wireless service in the USA? Circle the correct answer(s).

AT&T **FCC** **FTC** **GCHQ** **IETF** **NIST** **Verizon** **WIPO** **WTF**

FCC. Most people got it.

(p) A *Wall Street Journal* story says that communications with US drone airplanes in places like Afghanistan are not encrypted, and some officers are worried that adversaries "could manipulate the drone video feeds to hide battlefield movements." What ***specific*** kind of attack would this be?

botnet **denial of service** **man in the middle** **tailgating** **virtual machine** **war driving**

man in the middle. Most people got this too.

(q) Modern computers can efficiently process integers of several sizes, usually 1, 2, 4, 8, and sometimes 16 bytes long. Which of these is the least number of bytes that could be used for storing a binary number representing the population of California?

1 **2** **4** **8** **16** **none are big enough**

4. 2 bytes is only 65,000; 4 bytes is up to 4 billion. Not as well handled as it should have been.

(r) Amazon.com and the government of Brazil both want to own the top-level domain `.amazon`. What organization is responsible for deciding who gets the domain name?

ICANN. Most people got this.

(s) Suppose that Microsoft improves the satellite images used by Bing Maps from a resolution of 3 meters to a resolution of 1 foot. By approximately what factor will Microsoft have to increase the amount of disk space it uses to store the new images?

81 or 100 or the like. The linear measure improves by a factor of 9 or 10, so the area is that squared. Not very well done.

(t) Homeopathy involves diluting purportedly beneficial substances with water by powers of 100; the notation "10C" means dilution by a factor of 100^{10}, and "20C" means a factor of 100^{20}. Suppose that a homeopath who took COS 109 wants to dilute something to approximately 15C by diluting by powers of 2. What power of 2 corresponds most closely to 15C?

100. "15C" is 100^15 or 10^30, which is closest to 2^100. Pretty well done.

(u) The speeds of supercomputers are measured in floating-point operations per second, or "flops". Which one of these would be the most representative speed for the fastest of today's supercomputers?

100 Mflops **100 Gflops** **100 Tflops** **100 Pflops** **100 Eflops** **100 Zflops** **100 Yflops**

100 Pflops. Recall our classroom discussion and visit to top100.com?

(v) If I use my laptop in my office to search for "car repair," I get a list of local service stations like Princeton Sunoco on Nassau St. Which of these mechanisms is the most likely way that a search engine can so accurately guess where I am?

API **coookies** **Ethernet address** **GPS** **IP address** **nslookup** **traceroute**

IP address. Most people got it.

(w) "It is convenient to group the binary digits into tetrads, groups of 4 binary digits." (John von Neumann) What synonym or alternative terminology might be used today instead of *tetrads*?

Hexadecimal. I also accepted nibble. Most people got it.

(x) RSA Labs used to sponsor a factoring challenge: they published a list of very large integers and challenged the public to factor them. RSA-768, the largest challenge number that has been factored so far, is 768 bits long and has 232 decimal digits. If there were an RSA-798, how many decimal digits would it most likely have?

241 (or 240 or 242). 30 more bits is a factor of 2^30 or 10^9, so about 9 or 10 more digits.

(y) Suppose that a digital camera takes pictures that are exactly 2 MB in size. The total number of different possible photos is

2000000 **2000000^2** **$2^{2000000}$** **2000000^{256}** **$256^{2000000}$**

256^2000000. Each byte has 256 potential values.

2. (20 points) Understanding Programs

(a) The following Javascript code is supposed to simulate flipping a fair coin *exactly* 1,000 times. At the end, it should print the number of heads and tails. Sadly, it doesn't quite work. Fix the errors. You do not need to rewrite it if you clearly indicate the changes you would make. (This is a question about correct logic; don't worry about syntax. The `Math.random` expression is correct: each call of `Math.random()` produces a new random floating-point value between 0 and 1. The `alert` statement is syntactically correct as well.)

```
var i = 1;
var heads = 0;
alert("heads = " + heads + " tails = " + tails);
while (i < 1000) {
   var r = Math.random();  // random number r >= 0, < 1.0
   if (r >= 0.5) {
      heads = heads + 1;
   } else {
      tails = 1;
   }
}
i = 0 // fixed
heads = 0
tails = 0 // added
```

```
while (i < 1000) {
   r = Math.random()
   if (r >= 0.5) {
      heads = heads + 1
   } else {
      tails = tails + 1 // fixed
   }
   i = i + 1 // added
}
alert(("heads = " + heads + " tails = " + tails) // moved to here
```

One of many ways to fix it up. Reasonably well handled, though the increment of **i** was frequently omitted.

(b) If you want to simulate an unbalanced coin that comes up heads 3/4 of the time and tails only 1/4 of the time, what change(s) would you make to the program above to achieve this, after it has been corrected?

if (r >= 0.25) … is all that's needed.

(c) Suppose that the Toy machine has a **MUL** instruction that multiplies the accumulator contents by a value from the RAM and places the result back in the accumulator. What does this program print when given the sequence of numbers **5 4 3 2 1 0** as its input?

```
BOT   GET             get a number from user, place it in accumulator
      IFZERO  TOP     if accumulator content is zero, go to location TOP
      SUB     3       subtract 3 from accumulator content
      STORE   MID     store accumulator content in location MID
      MUL     MID     multiply accumulator content by content of MID
      MUL     MID     multiply accumulator content by content of MID
      PRINT           print content of accumulator
      GOTO    BOT     go to location labeled BOT
TOP   STOP
MID   0               when execution begins, this location will contain 0
```

8 1 0 -1 -8

(d) There is at least one place in the program of part (c) where an arbitrary sequence of instructions could be inserted between two existing lines without affecting the program's behavior. Identify one such place.

Before or after the line labeled TOP.

3. (110 points, 5 each) Miscellaneous

(a) The basic organizational unit in Excel is the worksheet, which at least through Office 2008 consists of an array of rows numbered 1 through 65,536 and columns labeled A, B, C, ..., Z, AA, AB, …, AZ, BA, BB, …, through IV. Give a plausible *technical* explanation for why the last column has the label IV instead of something that might seem more natural, like ZZ. Please be brief; I will stop reading after about 10-12 words.

Column names are in base 26 starting at 1, and IV is 256, so **the column number fits in one byte**. A number of people thought that there was some association with Roman numerals.

(b) Princeton logs all your Internet connections, including source IP address, the IP address you visit, your Ethernet address, and the Unix standard time (the number of seconds since 1970) at the beginning of the connection and at the end.

(i) If IPv4 addresses are used, how many bytes would be required to store this information for one connection, in the most straightforward and conventional representation?

22. 4 + 4 + 6 + 4 + 4. Partial credit for those who didn't include the end time.

(ii) How many bytes would be required if IPv6 were used instead of IPv4?

46. 16 + 16 + 6 + 4 + 4. Partial credit for those who didn't include the end time.

(c) This partial Unix directory listing shows size, modification date and time, and filename for five files. Exactly which pair(s) of files do I have to compare byte by byte to determine whether or not they have identical contents?

```
1247   Oct 29 16:04   f1.doc
1254   Oct 28 16:05   f1.docx
1254   Apr 22 20:03   f1copy.docx
1255   Sep 20 08:51   f2.txt
1254   Aug 20 08:51   f3.xls
```

All pairs of the three files with the same size. We beat this to death in class, problem sets and Q/A sessions.

(d) The first half of the first byte of an IP packet contains the version number of the protocol.

(i) Write out the bit patterns that one might most reasonably expect for IPv4 and IPv6.

IPv4 _____ IPv6 _____

0100 0110, the binary representations of 4 and 6.

(ii) What is the largest version number that this scheme allows for, in decimal?

15, which is 1111 in binary.

(e) As data travels across the Internet, it is subjected to a fair amount of processing. For each of the following statements, circle the most appropriate answer.

IP packets have serial numbers to ensure that they are processed in the right order	**true**	**false**
IP packets that arrive out of order have to be resent	**true**	**false**
a long IP packet is broken into multiple Ethernet packets	**true**	**false**
Ethernet packets are reassembled into IP packets at each router along the way	**true**	**false**
If an IP packet is lost or damaged in transit, that is detected by the intended recipient	**true**	**false**

false, false, true, true, false.

(f) Morse code uses combinations of one to five dots and/or dashes to represent letters, digits, and punctuation marks. For example, E is a single dot (·), A is dot-dash (·−), and Q is dash-dash-dot-dash (−−·−). Suppose you are designing a new version of a Morse-like code, in which every character will consist of some combination of *exactly* 6 dots and/or dashes. Describe ***briefly*** how you would ***systematically*** assign upper case letters and digits to combinations of 6 dots and dashes. Write down enough of your characters or explain how you create them so clearly that there is *no ambiguity* about your design.

Use binary! A = 0, B = 1, C = 3, etc. That's all you had to say. Some people laboriously wrote out a consistent set of patterns, a lot of work though for full credit if it was correct. Others relied on arm-waving ("and then the same for the other letters and digits"), and surprisingly many explained how to distinguish upper and lower case, even though the question only asks for upper case.

(g) From a Senate bill introduced by Ron Wyden (D-Oregon): "A covered Internet service provider may not, for purposes of measuring data usage or otherwise, provide preferential treatment of data that is based on the source or the content of the data."

(i) What is the most likely topic or issue that this bill deals with?

Net neutrality.

(ii) Name two companies or types of companies that might reasonably be on opposite sides of this issue.

Comcast v Netflix, or Verizon vs Vonage, or similar. I was looking for pairs where the ISP has a

clear conflict of interest with someone else's traffic. I don't think that Facebook (a popular answer) is part of this, but overall graded generously.

(h) A deep-space communications system continuously reports status information about some piece of equipment by sending a stream of status reports. There are three possible status values: OK, High and Low. 98% of the time, the status is OK, while High and Low each occur only 1% of the time. Give an encoding of the three values into three different bit patterns that will minimize the average number of bits sent over a long period of time. Your encoding does not have to use the same number of bits for each status, but there must be no ambiguity about how to decode a sequence of values as they arrive at the receiver.

0 10 11. Way too many people said things like 0 1 10, which is ambiguous, and even more used patterns with 3 or more bits, which does not minimize. Not well handled overall.

(i) A *Mersenne prime* is a prime number of the form $2^n - 1$ where n itself is prime, for example $31 = 2^5 - 1$. In December 2018, a new Mersenne prime was discovered, the largest known so far: $2^{82589933} - 1$. It has 24,862,048 digits in its decimal representation.

(i) If it is written out in binary, how many binary digits does it have?

82589933. As we discussed repeatedly, 2^n-1 has n bits, all of them 1's, as in the example above.

(ii) How many of those binary digits are zero?

None! I had thought this question would be a freebie.

(j) The *NY Times* (12/10/18) says that companies are continuously tracking 200 million US cell phones as many as 14,000 times per day per phone. Suppose that an average phone reports its number and its position to an accuracy of one yard or meter 1,000 times/day. ***Very roughly*** how many gigabytes of tracking information are uploaded by all these phones every day in total? (Hint: the circumference of the Earth is 25,000 miles or 40,000 km.) Be precise about your assumptions about how information is represented.

3000 GB? Location to nearest meter is a **pair** of numbers in the range up to 40M, so call it 4 bytes each. Phone number is 10 digits, or about 5 bytes. So each message is something like 15 bytes. 200M phones * 1000 uploads/day * 15 bytes/upload is about 3 * 10^12, or 3000 GB. Some people were misled by the "14,000"; not penalized.

(k) From a press release in January 2019: "Almost fifteen years ago, Lexar announced a 1GB SD card. Today, we are excited to announce 1TB of storage capacity in the same convenient form factor." Assume (unrealistically) that this improvement is a smooth exponential process.

(i) What was the growth rate of storage capacity per month during this time?

4%. A factor of 1000 in 15 years is doubling every 1.5 years or 18 months. 72/18 = 4.

(ii) If Lexar continues this rate of progress, in what year will they announce a 1PB SD card?

2034, which is 2019 + 15.

(l) The US subset of ASCII uses only 7 bits; the leftmost (8^{th}) bit is always zero. Suppose that instead we use the leftmost bit to give each character ***odd*** parity. Without parity, the hexadecimal value of the ASCII digit `0` is 30, and the other digits follow in numerical order. Write down the *hexadecimal* values of the digits `1` through `5` with that additional parity bit.

31 32 B3 34 B5. Easiest done by writing binary (0 is 00110000, 1 is 00110001, etc.), then inserting the parity bit. Partial credit for leaving it in binary instead of hex.

(m) Suppose that you have just been appointed as Princeton's Dean of Admissions, and you want to experiment with machine learning to evaluate applications, specifically to predict each applicant's final GPA when they graduate, given only information from their application. ***Briefly but precisely*** explain how you would do this: what data you would use and how you would process it.

Choose features like HS GPA, standing, SAT scores, etc. Use data from PU graduates as a training set.

The machine learning algorithm decides how to weight the features, not you; that's the whole point. I really wanted to see something about training on PU grads and it was helpful to mention supervised learning, of which this would likely be an example.

(n) An article in the *NY Times* (12/17/18) describes "dynamic billboards" that change what they display according to who is near them; the advertising is targeted to specific motorists who are driving by. ***Briefly*** (!) describe how this might work: how the billboard knows you are nearby, how it knows what advertisements might appeal to you, what information is stored where and communicated from where, and the like. No long essays or arm-waving, please, just a concise description of how this could work.

My guess: app(s) in your phone send your location to a server that knows about you from your online behavior and lets advertisers bid; the winner's advert is sent to the billboard via wireless or cell or wired connection. The critical pieces are the GPS-enabled phone, your known identity, and your profile. It's not much different from what already happens. I doubt that it would by a vision system recognizing you in your car, and generic advertisements based on car type are not as focused as this is said to be.

(o) Supercomputers are often organized as a "mesh" where each processor is connected to its nearest horizontal and vertical neighbors on a rectangular grid. Suppose that there are **N** processors, each processor is an identical rectangular box, and the boxes fill a large room from floor to ceiling.

(i) How many connections to neighbors does a typical processor have?

6. four sides, top, bottom

(ii) How does the total number of connections grow in proportion to **N**?

linear, or N or 6N.

(iii) If technology improves so that the current length, width and height of each processor can be shrunk by a factor of two, about how many processors would now fit in the room?

8N. Each device is half as big in each of 3 dimensions.

(p) "A common grayness silvers everything" (*Andrea del Sarto*, Robert Browning). The color "gray" describes any color that has equal amounts of red, green and blue.

(i) In the standard 3-byte representation of RGB colors, how many different shades of gray are there? (Hint: it's not 50.)

256. There are 256 red values, and the green and blue have to be the same as the red.

(ii) There are two shades of gray that could be called "medium" gray because they are at the middle of the range of shades. Write out both of these colors in hexadecimal.

7F7F7F, 808080

(q) An old TV commercial shows a truck stopping at an IBM help desk at the side of a 2-lane country road. The person at the desk says to the amazed driver, "The boxes told us you were lost -- there are RFID tags on the cargo to help track shipments." Like most advertising, this might get some technical details wrong. Circle the answers below according to whether or not the system could reasonably include such a mechanism.

the truck location is tracked by satellite cameras	**probable**	**improbable**
a GPS satellite detects signals from the RFID tags	**probable**	**improbable**
a cell phone in the truck broadcasts its location to a GPS satellite	**probable**	**improbable**
a cell phone in the truck sends a signal to nearby cell towers	**probable**	**improbable**
nearby cell towers use Bluetooth to detect the RFID tags	**probable**	**improbable**

Satellites are too far away and are also passive; Bluetooth is too short-range.

(r) An IPv4 network address with **N** bits in the network part is written in dotted decimal notation as `d.d.d.d/N`, where each `d` is an integer between 0 and 255 and **N** is an integer between 0 and 32. For example, one Princeton network is `128.112.0.0/16`. Suppose that the Department of Tendentious Literary Analyses (TLA) has the subnet `128.112.128.0/23`.

(i) How many host computers can there be on the TLA subnet simultaneously?

512. The TLA subnet has 9 bits (32-23). Not understood by most.

(ii) What is the lowest possible host address on the TLA subnet, in dotted decimal?

128.112.128.0.

(iii) What is the highest possible TLA subnet address, in dotted decimal?

128.112.129.255. Maybe easiest seen by writing binary; the last two bytes of the maximum possible address would be 10000001 11111111. Overall, this question proved tough. We went through a very similar one in the final Q/A session; I hope the attendees had an easier time because of that.

(s) Suppose, not unrealistically, that **N** high-tech companies are involved in a bunch of lawsuits.

(i) If each company sues each other company, how does the number of lawsuits grow in proportion to or as a function of **N**?

N^2. Beaten to death: if the problem says "each does something to all the others", it's quadratic.

(ii) Companies may also band together in groups of various sizes to sue companies that are not in the group; for instance if **N** were 4, we might have A suing B, C and D; A and B suing C and D; A, B and C suing D; and so on. If all possible combinations of companies initiate such suits, how does the number of possible lawsuits grow in proportion to **N**?

2^N. This should have reminded you of the key distribution problem from one of the later problem sets. Both parts were reasonably well handled.

(t) The professor in a class with **N** students normally returns problem sets that he has laboriously sorted by student name. For each of the following, give a single expression in **N** (e.g., 2^N) that tells how the work is proportional to or depends on the size of the class in the worst case.

– If the professor uses an efficient algorithm, how much work does he have to do to sort the problem sets?

N log N

– How many problem sets does the first student have to look at to find her problem set in the sorted pile, if she uses an efficient algorithm?

log N

– How many problem sets in total must be looked at by all the members of the class when the pile is sorted, if each in turn uses an efficient algorithm to find his or her own problem set?

N log N

– If the professor fails to sort the problem sets, how many problem sets does the first student now have to look at to find her problem set in the unsorted pile?

N

– How many problem sets in total must be looked at by all members of the class when the pile is unsorted?

N^2

(u) [1 point each] Random quickies.

If you use HTTP to access a web site, your ISP does not know which site it is	**true**	**false**
If you use HTTPS to access a web site, your ISP does not know which site it is	**true**	**false**
If you use Tor to access a web site, your ISP does not know which site it is	**true**	**false**
The SHA-3 algorithm resulted from a worldwide competition run by CERN	**true**	**false**

Citizens of the European Union who live in the USA are protected by the GDPR	**true**	**false**
A lossless compression algorithm can shrink any input data by some amount	**true**	**false**
A two-factor device is used for efficient testing of primes in cryptographic processes	**true**	**false**
A Turing machine is a purely mathematical idea, not something that could be built	**true**	**false**
Zipf's Law is the theoretical basis of the LZ compression algorithm	**true**	**false**
"If a website has a privacy policy, that means the site won't share user information with other sites or companies without permission."	**true**	**false**

Reasonably well done overall, on average better than just guessing