CSC 102 • HOW THE INTERNET WORKS
FINAL TAKE-HOME EXAM **KEY**
DUE 2:00 PM ON TUESDAY, 21 DECEMBER 2010
PLEASE SUBMIT YOUR WORK ELECTRONICALLY VIA MOODLE

You may not consult any person other than the professor when completing this exam, but you may use references online and in print. Wherever you use external sources, leave me a URL or citation. Don't forget to use quotation marks when copying verbatim from a source. There will be a certain amount of Web searching necessary to respond to several of the questions. Wikipedia will be accepted as a source of answers. You may also have to use Google or its equivalent intelligently, and some of the other tools we have studied.

1. **Data Transfer** (12 points)

The Sloan Digital Sky Survey is a detailed digital map of the heavens used by astronomers. The images and associated data released to the public totals 40.5 TB (note: 1TB = 1000 GB = 1,000,000 MB). The astronomy department at Stardust College want to study the survey data, and are trying to figure out a way to get a full copy.

a.) They are connected to the Internet with a very fast ethernet cable that has maximum capacity of 100 Mbps. In the unlikely case that they were able to devote the entire connection bandwidth to downloading the data set, how long would this take?

40.5 TB = 40500000 MB = 324000000 Mb.
324000000 MB/100 Mbps = 3240000 s = 900 hours

b.) A graduate student in the department suggests that since Bigbucks U. has a copy of the full data set on BluRay disks, she could drive there in her car and pick them up. How many disks would be required to record the entire data set, if each has a capacity of 50 GB?

40.5 TB = 40500 GB
40500 GB/50GB per disk = 810 disks

c.) Assuming that the trip from Bigbucks U. to Stardust College takes two and a quarter hours, what is the effective data transfer rate of this operation?

2.25 hours = 135 min. = 8100 sec.
40.5 TB/8100s = 0.005 TB/s = 5 GB/s = 40 Gbps

I also accepted answers using 4.5 hours as the time, for those who figured the answer based upon the round trip.

2. **Communications Protocols** (16 points)

A home network is built around a wireless router that uses DHCP with network address translation. The router assigns local IP addresses in the reserved 192.168.x.x range; it takes the 192.168.0.1 address and gives successively higher addresses to each local connection. The global IP address provided by the household's internet service provider is 192.76.85.245.

a.) The desktop computer connected to the router is at 192.168.0.2 on the local area network. Suppose that it sends out a request for a web page to www.smith.edu. As the packet travels (i) from the desktop to the router and (ii) from the router to www.smith.edu, what would it show as the sender IP address and as the destination IP address?

(i) sender: 192.168.0.2; destination: 131.229.64.19
(ii) sender: 192.76.85.245; destination: 131.229.64.19

b.) When www.smith.edu responds to the web page request, it sends packets back to the desktop computer. As the packet travels (i) from www.smith.edu to the router and (ii) from the router to the desktop computer, what would it show as the sender IP address and as the destination IP address?

(i) sender: 131.229.64.19; destination: 192.76.85.245
(ii) sender: 131.229.64.19; destination: 192.168.0.2

c.) Suppose that a laptop connected to the same router at 192.168.0.4 on the local area network also requests a web page from www.smith.edu at the same time. Packets are sent to the router intended for both recipients. How does the router know where to send them?

Packets are distinguished according to the TCP port number and session ID.

d.) Give at least one advantage of using network address translation for small home networks like this one.

Advantages include better protection from external attacks and the ability to share one global IP address between multiple local hosts.

3. **HTML** (18 points)

Identify the outright errors in the web page source below, as well as instances where it departs from the "best practices" taught in class.

```
<HTML>
<TITLE>Exam Question</TITLE>
<BODY bgcolor="blue">
<H2>HTML Exam Question</H2>

This is part of the <B><EM>CSC 102 Fall 2010</B></EM> final exam.
<P>
You should try to find all the errors &ampersand bad habits in this page.
<FONT color="gray">It is very sneaky, with lots of hidden mistakes.</FONT>

<HR>~~~~~~~~</HR>

<UL list-style-type="decimal">
<LI style="list-style-position: inside">One item</LI>
<li>Another item</li>
A third item
</UL list-style-type="decimal">

<PRE>
A heart:  <3
</PRE>

<BLOCKQUOTE><BLOCKQUOTE>
I love indenting!
</BLOCKQUOTE></BLOCKQUOTE>

<CENTER>Middle!</CENTER

<TABLE width=100%><TR><TD width="90%"></TD><TD>Right!</TD></TR></TABLE>

</BODY>
Last modified:  11 December 2010<BR>
</HTML>
```

One point each for finding the items below; 18 points maximum:

1. *Tags should be lower case (best practice)*
2. *Missing <head> element (error)*
3. *style should be used instead of bgcolor (best practice)*
4. *Main header should be <h1> not <h2> (best practice)*
5. *<b> and <em> tags aren't nested properly (error)*
6. *<p> tag requires close (error)*
7. *&ampersand requires closing semicolon (error)*
8. *Avoid using <font> (best practice)*
9. *<hr /> is a self-closing tag, and does not use a close (error)*
10. * should not be used to make a numbered list; use <ol> instead (best practice)*
11. *Third item in list is missing <li> tag (error)*
12. * closing tag should not have an attribute (error)*
13. *< symbol in heart should be replaced by < (error)*
14. *<blockquote> should appear only once, with style to increase indent (best practice)*
15. *<center> tag should be replaced with style (best practice)*

16. *</center> tag is missing closing angle bracket*
17. *Table cells need contents, like   (error)*
18. *Text appears after </body> closing tag*
19. *Use <strong> instead of <b> (best practice)*
20. *<table> tag should have width="100%" with quotes (best practice)*
21. *Width on table and cells should be set using styles (best practice)*
22. *Inline styles would be better replaced by style rules (best practice)*
23. *<doctype> declaration required for standards compliance (best practice)*

4. **Images** (6 points)

What three image file formats are recommended for web use? Suppose that you are designing a web site for a professional photographer, which will include a logo and samples of work. In this context, specify for each file type which parts of the site you would most appropriately use it for. If you would not use one of the three in this context, explain why.

PNG, JPEG and GIF are recommended. The logo should be PNG, and the photographic samples should be JPEG. There is no reason to use a GIF, as animation is not needed.

5. **HTML Forms** (8 points)

Devise a form that would produce the URL result shown below if submitted using its default values. (In other words, the form is submitted without making any changes and it produces this result.) The four results shown below should come from a checkbox, text input, popup menu, and button, respectively.

```
http://example.com/form.html?new=Yes&name=Anonymous&side=left&ok=ok
```

```
<form action=" http://example.com/form.html" method="get">
<input type="checkbox" name="new" value="Yes"checked="checked"  />
<input type="text" name="name" value="Anonymous" />
<select name="side">
<option value="left" selected="selected">Left</option>
<option value="right">Right</option>
</select>
<input type="submit" name="ok" value="ok" />
</form>
```

6. **Email** (14 points)

Consider the email message source below. (Some irrelevant portions of the email have been redacted.) Please answer the questions that follow.

```
Return-path: <nicholas.r.howe@gmail.com>
Received: from mscreen3.smith.edu (mscreen.smith.edu [131.229.64.72])
      by gwsmtp1.smith.edu with ESMTP; Sat, 11 Dec 2010 22:37:08 -0500
Received: from scmapp1.smith.edu (scmapp1.smith.edu [131.229.64.81])
      by mscreen3.smith.edu (8.14.3/8.14.3) with SMTP id oBC3b9IY007482
      for <nhowe@smith.edu>; Sat, 11 Dec 2010 21:37:09 -0600
Received: from (unknown [209.85.161.177]) by scmapp1.smith.edu with smtp
       id 5a4f_5c4e_187d6dd8_05a1_11e0_bd57_0014221cc49d;
      Sat, 11 Dec 2010 22:37:08 -0500
Received: by gxk27 with SMTP id 27so2968162gxk.36
        for <nhowe@smith.edu>; Sat, 11 Dec 2010 19:37:07 -0800 (PST)
MIME-Version: 1.0
Received: by 10.236.95.41 with SMTP id o29mr5588573yhf.40.1292125027865; Sat,
 11 Dec 2010 19:37:07 -0800 (PST)
Received: by 10.236.108.129 with HTTP; Sat, 11 Dec 2010 19:37:07 -0800 (PST)
Date: Sat, 11 Dec 2010 22:37:07 -0500
Message-ID: <AANLkTimcX54STYqBDWubNdceJpiZsaOmGxZ+EVnThapk@mail.gmail.com>
Subject: Test Email
From: Nicholas Howe <nicholas.r.howe@gmail.com>
To: nhowe@smith.edu
Content-Type: multipart/mixed; boundary=00235448f2594a956304972e4d2c

--00235448f2594a956304972e4d2c
Content-Type: multipart/alternative; boundary=00235448f2594a955804972e4d2a

--00235448f2594a955804972e4d2a
Content-Type: text/plain; charset=ISO-8859-1

This is an email for the CSC 102 fall 2010 final exam.

--00235448f2594a955804972e4d2a
Content-Type: text/html; charset=ISO-8859-1

This is an email for the CSC 102 fall 2010 final exam.<br>

--00235448f2594a955804972e4d2a--
--00235448f2594a956304972e4d2c
Content-Type: image/png; name="sun.png"
Content-Disposition: attachment; filename="sun.png"
Content-Transfer-Encoding: base64
X-Attachment-Id: f_ghldqb2h0

iVBORw0KGgoAAAANSUhEUgAAABAAAAAQCAIAAACQkWg2AAAAAXNSR0IArs4c6QAAAARnQU1BAACx
jwv8YQUAAAAgY0hSTQAAeiYAAICEAAD6AAAAgOgAAHUwAADqYAAAOpgAABdwnLpRPAAAAJtJREFU
OE99UkkSwCAM8mk+zZ9bzQoa2+nBScgCoc36a3P0MtMySogV99SOJwwKdtQbK8haQBxLZVTk4JGT
d3ed4Gmbrvv4H0GplIISd9YY2YM0j1qNiIwtEzrFGmMNFag8kNjowAF3MIkKJsDshwmcKEjrYlzP
+gQCoE9ZpVfKijdGlU0l6ojW4EuHL9iFVwH6h8xyH+50c+3wD/S8XKesH5YgAAAAAElFTkSuQmCC

--00235448f2594a956304972e4d2c--
```

a.) How many tags make up the email header in the message source as shown? (Count the lines, ignoring indented lines.)

Fourteen

b.) What is the text of the message? Why does it apparently repeat itself in the message source?

"This is an email for the CSC 102 fall 2010 final exam."
It repeats because the same message appears in plaintext and HTML formats.
c.) There is an attachment to this message. What type of content does it contain? Could opening this attachment harm your computer, according to the guidelines discussed in class?

The attachment is a PNG image file. According to the guidelines discussed in class, opening this file could not hurt your computer. It is a data file, and opening it would not run an untrusted program.

d.) Suppose that you were suspicious of this message, and wanted to do everything you could to confirm the source. Describe what you can deduce or uncover about where this message came from.

The earliest IP addresses are untraceable. But 209.85.161.177, from which Smith's email server received the message, is registered to mail-gx0-f177.google.com. Since the email sender has a gmail address and the mail was handled by a google mail server, the tentative conclusion is that it comes from a gmail account.

7. **Cryptography** (6 points)

Imagine a world where strong cryptography does not exist, and compare it to a world where strong cryptography is widely available and widely used. Discuss at least three positive or negative aspects of each scenario. (You may count a positive for one as a negative for the other; e.g., you need to discuss at least three areas of significance, and describe why each would be advantageous/disadvantageous in one scenario or the other.)

Without strong cryptography, online commerce as we know it would be impossible. Different payment mechanisms would have to be devised that did not rely on electronic communication. Remote login to computer systems might not be possible because passwords could not be kept secret in transmission. It would be impossible to verify identity online, leading to impersonation and identity theft. On the other hand, if strong cryptography were more widely available and widely used, law enforcement would become more difficult because wiretaps would be useless. Censorship would also be more difficult because it would be hard to detect what content was being viewed. More network resources would be consumed in transmitting email and other messages in encrypted format. And users could lose access to their own data if they forgot the corresponding encryption key.

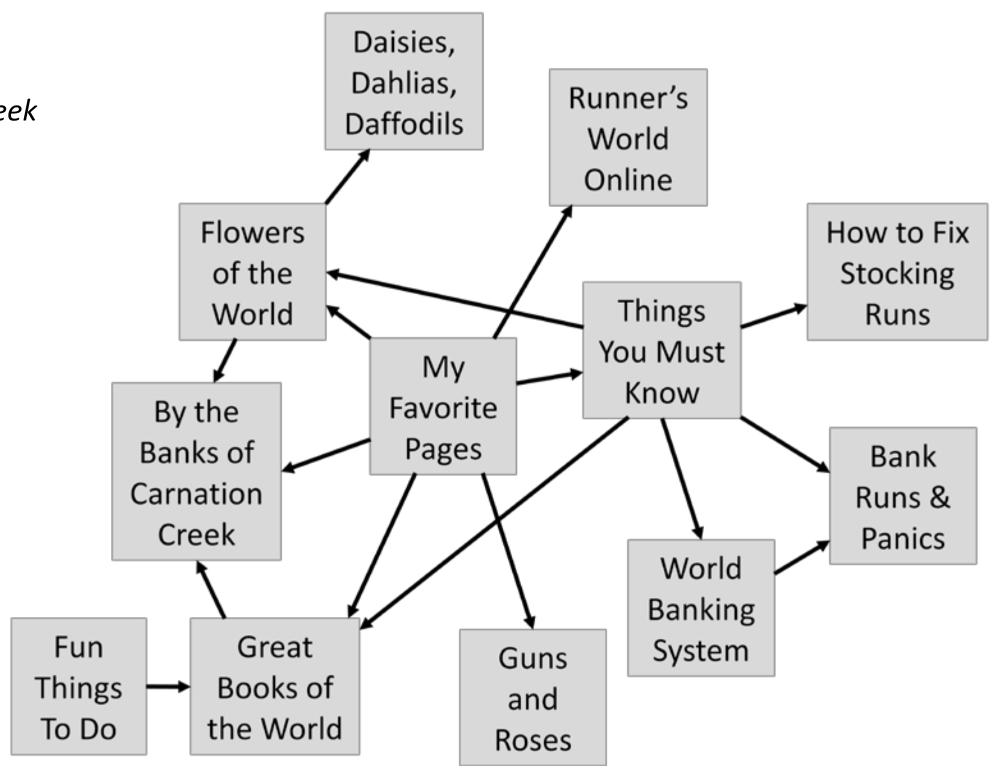
8. **Personal Safety Online** (6 points)

Why do Microsoft and other software companies release security patches for their software, knowing that hackers will examine these closely to learn about vulnerabilities of unpatched systems? Discuss this topic, including its implications for individual computer users like you.

Microsoft presumes that hackers will sooner or later discover and exploit vulnerabilities if they are left unfixed. Thus it is better to fix the loophole and encourage everybody to update their software than to leave it unpatched. The down side is that users who do not take advantage of the freely available patches will be more vulnerable. This is one reason that Microsoft and other software vendors have introduced automatic updating services.

9. **Web Search** (16 points)

For each of the queries below, indicate which pages would be retrieved and their rank order, if a search engine uses a bag-of-words model with query augmentation, and ranks results according to Google's PageRank model. (Note: you are not expected to carry out the full PageRank computation, which is mathematically complex. Just rank according to the principles and example used in class. If you're not sure, include a justification of your choice.)



a.) Query: Flowers

By the Banks of Carnation Creek
Flowers of the World
Daisies, Dahlias, Daffodils
Guns and Roses

b.) Query: Running

Bank Runs and Panics
How to Fix Stocking Runs
Runner's World Online

c.) Query: Bank runs

Bank Runs and Panics

d.) Query: World

Great Books of the World
Flowers of the World
World Banking System
Runner's World Online

Comment: Many people had trouble with this, a sign that I didn't teach it as well as I should have. The best approach is to first figure out what pages will be retrieved, and then determine their relative PageRank to come up with an order. (The rank was only one point for each item.)

For retrieval, a page must contain all the keywords in some form, or a closely related word. ("Query augmentation" is the term indicating that related words must be considered.) Thus "running" will retrieve runs, runner, etc. "Flowers" with query augmentation will retrieve flower, rose, daisy, dahlia, carnation, etc. "Bank runs" will retrieve only documents with both keywords.

PageRank is actually computed on all pages independently of a given query. It is based on the authority in incoming links. Thus pages without any incoming links will have very low PageRank. Those with many will have high PageRank, especially if the links come from other pages with authority. The figure in the course slides on PageRank shows this visually. Thus, in this set of pages the highest PageRank goes to "By the Banks of Carnation Creek" with three incoming links. "Great Books of the World" also has three incoming links, but they are all lower rank so it will be second. "Flowers of the World" is next with two links, and so on.