Comments

Here are my comments on the two papers submitted for ECON 18. I'll first discuss common elements for the two papers and then add some thoughts that are specific to each paper. I also include an appendix of my comments to the penultimate drafts.

Comments on Both Papers

Overall, I think that both papers are very strong. You have clearly spent many hours on this research, both in the coding/data analysis and in the actual write ups. Both papers are professionally presented along with their associated source code. I suspect that very few other students produced work as substantive as this over Winter Study. Indeed, I think that the quality of these papers surpasses most of the papers submitted in upper level regular semester Economics classes. You should all be proud of this work.

1) Note how easy it is for me to use the built in tools in R to pull your code out of the .Rnw file.

```
> Stangle("Fraulo_Nguyen_Econ18_2009")
Writing to file Fraulo_Nguyen_Econ18_2009.R
>
```

Now I can examine every line of code myself. I can also replicate every aspect of your calculations. That's the way that science should work but, alas, not how it works often enough, at least in economics. Consider the code chunk that starts the analysis:

```
####################################################
## CREATE Z
## Compile the necessary data into z
require(ws.data)
data(secref)
data(yearly)
data(daily.1998)
data(daily.2007)
x <- grab.data(symbols = secref$symbol, years = 1998:2007)
x <- subset(x, tret < 2)
y.1 <- calc.returns(x, d.before = 30, d.after = 0, actual = TRUE)
y.2 <- calc.returns(x, d.before = 0,   d.after = 30, actual = TRUE)
z <- merge(y.1, y.2)
```

This looks fine to me. But note that it would be helpful to have more comments. For example, it is obvious that you need to issue the data() command in order to bring secref into the workspace because you use the symbol column in secref in the call to grab.data(). That makes sense. But why use data on two of the random years? I suppose that you do this because you want to examine this data to find the last and first dates of the data. There are problems with this.

First, you should make a comment to explain that motivation, both for the benefit of anyone looking at this work and for your own benefit, should you revisit the project a few months or more from now. Second, in a later section when you actually calculate the min and max date, you reissue those same data() commands. Better to just do so once. Third, you probably don't want to use the raw data to calculate the min and max date in any event. The reader is not really interested in the date that was the earliest available *in theory*. She wants to know the earliest actual data that you, *in fact*, used. Now, in this case, you would get the same answer if you used, say, the x data frame calculated with grab.data(), but the point still holds.

2) The overall structure of both papers is excellent. The abstracts provide good summaries of the entire papers. The introductions are longer versions of the abstracts. The conclusions are shorter versions of the introductions. A reader could learn a great deal from just the abstract or from just the abstract, introduction and conclusion. For every 100 readers who examine a paper's abstract, only 1 (at most) reads it all the way through. I realize that it is somewhat depressing to think that so many people might read your abstract without reading the rest, but that is the way of the world. So, you are wise to spend so much time ensuring that your abstracts are as good as they can be.

3) Although not perfect, the tables and figures are well done. In almost all cases, I could read just the tables/figures (along with their captions) and have a good sense of what the papers are about. I could also read just the papers, while ignoring the tables/figures, and not miss anything important. A couple of figures don't have captions and a couple of figures are not mentioned in the text of the paper itself.

4) Both papers fail to report any measures of uncertainty. As we discussed in class, that is not your fault since we did not have the time to learn about such measures. But, for any future work you might do, you should be aware of the importance of such measures.

5) I think both papers represent real contributions to the academic literature. After cleaning up some of the remaining issues, you might consider posting them at SSRN. Both papers would make fine jumping off points for independent work or for senior theses.

Comments on Jannen and Pham

As I mentioned in the comments to your previous draft, this is excellent work. You have accomplished much of what we set out to do at the start of Winter Study. It was very generous of you to take up a leadership role in terms of generating code that could be used by everyone in the class, organizing the Google code doc and so on. Much appreciated. Below are just a collection of minor points for you to consider.

Your introduction is good but probably too detailed. You probably don't need to go into so much detail as to the actual results just yet. Save something for the body of the paper.

You use "modernity of the data" when you probably mean "recency of the data."

On the bottom of page 2, you report some key results. There are two problems. First, you need to get the significant digits correct. MG may have returns 1.357%, but no one wants to read that last digit. Better would be 1.36% or, even better, 1.4%. Financial research is rarely accurate to more than one decimal place. Second, the numbers you present are not consistent with the words around them. You claim that GH is the weakest result, but its 1.043% spread is larger than the 0.888% spread seen with JT. I suspect that the result that you saw in an initia, draft might have changed once you got rid of some of the data mistakes.

(This is one of the dangers of the replication/Sweave approach. You don't know what the number will be until you actually calculate it, but, by that time, the result may contradict the words you wrap around it. The only solution, obviously, is to ensure that your results --- even as you update them with better error checking, a more refined universe and so on --- are consistent with your description.)

Be precise in your language. On page 3, you reference "200% over consecutive trading days." Is that really what you mean? Probably not. "Consecutive" trading days are not different, as far as I can see, from non-consecutive trading days. We do not treat Friday-to-Mondays differently from Tuesday-to-Wednesdays. You mean 200% daily return, regardless of what day of the week it in.

On the bottom of page 3, you have a date range (1999-01-31 to 2007-05-31) that does not make sense to me. Since you only need 12 months of prior data to form a portfolio, shouldn't the first portfolio be on 1998-12-31? Since you only need 6 months of future returns after portfolio formation, shouldn't the last portfolio be formed on 2007-06-30? The nice thing about reproducible reseach is that I can examine your exact calculations in the .Rnw file:

```
-----------
For this reason, our first portfolios are not formed until
\Sexpr{as.character(sort(big.table$v.date, dec=FALSE))},
and our last portfolios are formed on \Sexpr{as.character(sort(big.table$v.date,
dec=TRUE))}.
-----------
```

Hmmm. Obviously, I now to need check out your R code and see how you constract big.table. Easy! I just run Stangle() on your .Rnw file and search for "big.table". Here is the key code:

```
-----------
big.table <- subset(merged, (merged$v.date > (min(merged$v.date) + 365)) &
               (merged$v.date < (max(merged$v.date) - 182))
               )
-----------
```

And, for that, it is fairly easy to see what went wrong. Although there is nothing clearly wrong with using 365 and 182 *days* as a *hack* to figure out a date that is a year in the past or 6 months in the future, in this particular case, you are losing one month at each end for now good reason. Recall that the first day in our data is in early January 1998 (not December 31, 1997). So, when you insist on a date that is at least 365 days after that first date, you end up with January 31, 1999, when what you really want is December 31, 1998. A similar thing happens on the other end.

Now, obviously this mistake is unlikely to make a difference in your final results, but this exercise highlights how nice it is for one scientists (read: me) to figure out exactly how other scientists (read: you) have conducted their research.

Getting back to more minor points:

On page 6, the table reference does not resolve properly. Also note that both figures need much more extensive captions. See Faulo and Nguyen (2009) for good examples. I am also somewhat suspicious of the data for the horrible month in early 1999. Is that really possible? As you correctly note in the paper, each months portfolio is getting 6 months worth of returns, so there should be significant correlation between one month in the next, as, indeed, we see in most of the other months. For a single month to be such a huge outlier suggests that just one or a few stocks got in the spread portfolios that month, but not the month before or the month after. That's possible, of course, but probably bears further study.

On page 8, the labels in Figure 2 should be more informative. Use "52-week high" instead of "ratio" and "industry momentum" instead of "ret.6m.0.ind". I realize that R is just automatically placing in the names of the variables in those slots, but, given that, you should change those names in the step before plotting. You want to make your figures as clear as possible to the casual reader.

Your extension is very interesting. As I mentioned in class, I do not think that anyone has looked closely at the topic of recency in the published academic literature. This is a real contribution.

Summary: A job well done!

Comments on Fraulo and Nguyen

Overall, this is a solid paper but you would have benefited from a few more days of work. Consider your opening paragraph.

Jegadeesh and Titman (1993) demonstrate that stocks which have performed well (poorly) over the last few months continue to do well (poorly) over the following months, and label this occurrence the "momentum effect". This momentum effect, originally shown by Jegadeesh and Titman (1993) [7] has spurred numerous subsequent attempts to further our understanding of what creates this arbitrage opportunity in the market, such as George and Hwang (2004) [2] who showed the relationship between the 52-week high and momentum as well as Han and Grinblatt (2002) [4] who demonstrated the disposition effect as it relates to momentum. Moskowitz and Grinblatt (1999) (hereafter MG) [3] attempted to show that the momentum effect, originally demonstrated by Jegadeesh and Titman (hereafter JT), appears to be stronger when viewed by industry rather than by individual stocks.

The first setence is excellent. Starting an academic paper with a precise description of what previous researchers have discovered is always a good idea. But then the second sentence is a mess. Why is there a reference "[7]" to the paper here when you did not need one (?) in the first sentence? Why two uses of "originally?" In truth, you don't even need to use this adverb once. By crediting the momentum effect to JT, you are already indicating that they were the ones that came up with the idea.

Recall the comment that I made on your penultimate draft.

Although the Introduction is not as important as the Abstract, it is the second most important part. Make sure that your sentences are precise. "this phenomenon, such as George and Hwang (2004)" is *gibberish.* The academic paper George and Hwang (2004) has nothing to do with phenomenon or characteristics or even, directly, the momentum effect. The reference here it to one of the "attempts", but that occurs so far earlier in the sentence that the connection can not be made by the reader.

You still have exactly this problem in your Introduction! At least I can claim consistency in finding this sentence to be problematic. In this version, George and Hwang (2004) and Han and Grinblatt (2002) are connected to an "arbitrage opportunity in the market," which, of course, is not what you mean.

The last paragraph of the Introduction is adequate but not optimal. The last point you make --- that you find the industry momentum in a time period not considered by MG --- is important enough that it should not just be tacked on as the last sentence in a paragraph devoted to your extension. Make that point earlier and finish up with your extension.

Minor issues: You do not cite all the references that appear in your Bibliography, at least as far as I can tell. For example, footnote 1 should mention Campbell et al (2007) but does not. Maybe I am missing something, but I don't see a reference to Figures 1 and 2 in the text of the paper itself.

Your extension is interesting. As I mentioned in class, there is nothing wrong with using an extension that returns "boring" results. Perhaps it would have been interesting to discover that industry momentum works better (or worse) in smaller industries. But, from a scientific point of view, discovering (and demonstrating) that there is no such effect is just as important.

Your conclusion, like your abstract, is just about perfect.

Summary: A solid paper.

Appendix

Below are the collection of comments I made on the penultimate drafts of the two papers. Many of these comments (but not, alas, all) were incorporated in the final versions. These comments were made in a series of e-mails, all of which were distributed to the entire class; so some of the comments apply to both papers. It is tough to fully understand what is going on without having read those drafts. My goal in reprinting the comments here is to give prospective students some insight into the knitty-gritty, detail-orientated world of fiancial data analysis.

Comments on the previous draft of Jannen and Pham

I think that this draft is in good shape. Below are lots of little suggestions. But, big picture, this is exactly the sort of work that I hoped students would be producing by the end of the class.

1) Abstract is very good. I would not mention Table I and Table IV specifically. Get into those details later. I would say 1998 to 2007. No need to mention specific months. Typo "their their". Make sure to proof-read. Have a friend look over stuff closely. Give another sentence on recency. Just how "strong" an indicator is it? Give a specific number.

2) Make the font bigger (12 pt?) and double space. It is a little hard to read. Also, this will allow you to place the figures and tables more nicely. Do so.

3) Jumping ahead, Table 3 seems to be missing lots of rows. (I was looking at this to find the number that you should use in the abstract.)

4) I would not use the phrase "search space" in the introduction. Too computer sciency! "universe" instead.

5) The first paragraph of the Introduction is perfect.

6) The second paragraph comes too fast. I think that you need a paragraph after the first and before the second which, more slowly, introduces the reader to the world of portfolios, momentum investing, 52 week high and so on. Maybe putting the current 3rd paragraph on data as the 2nd paragraph would help?

7) Given that you cite all your papers (correctly) as Marshall and Cahan (2005) or whatever, you should not *also* use the [MC05] type citations. Model your citations on the papers we have read. Just use the Marshall and Cahan (2005) type approach. Feel free to hack this by not having the [MC05] type citations appear in the text, even if you use them to have the right entry appear in the bibliography.

8) "capitalizes" should be "capitalizes on"; "stocks' prices" should be "stock prices".

9) In the last paragraph of the introduction, you should give a sentence or two of more detail along with actual numbers. Something like "We break stocks into three categories based on how many days since the 52-week high . . . . stocks in the first category perform twice as well as . . ." Or give the actual percentages.

You data and methods section is not nearly as good.

Most importantly, you need to describe exactly what you/we do in exactly the way that you do it. Consider the opening sentence:

"Our data consists of 2937 securities, traded between 1998-01-31 and 2007-11-30. From the data, we select the largest 1500 stocks in each given year. All portfolios are formed strictly from a portion of those largest stocks."

That's not true! You don't begin with 2,937 securities (from where? how chosen?) and then, out of that convenience sample, selecting the 1,500 largest. Instead, you go to December 31 in each year from 1997 through 2006 and determine the 1500 largest cap US stocks. That forms the core of the universe. Then you look at the 106 (or whatever) month ends from date X to date Y and blah, blah, blah.

You need to describe precisely where the data comes from and how it is created. What was the date of your first portfolio? Hint: It is not 1998-01-31.

The first two paragraphs need to be rewritten. Just describe step by step what happens. And provide much more detail on the data, as we discussed in class. What month had the fewest number of stocks. Which had the most? How many industries are there? And so on. This all needs to be about three times longer. You need to make the reader very comfortable with your data.

The rest of the section is OK. You do a good job describing the specifics of the different strategies.

I don't think that your last paragraph belongs here. Save it for the extension section. At this point, the reader does not care about your extension. He just wants to know about your data and the methods you use to replicate the other results. In fact, it looks like you just copy/pasted that paragraph in two places! Not good.

You only have two references.

Minor comments on 3rd paragraph of Data and Methods:

"industry that" should be "industry to which" delete "the ratio"

typo "motht-1"

"go long on" should be "go long"

I would just delete the sentence on self-financing. It is awkward and you don't really need to explain the concept in the paper.

This is a brief note on sum of the implausible results in your figure 1. I think that you have either done something wrong in calculating 6 month forward returns or are constructing portfolios in a suspect way.

First, here is what I see for 6 month returns, using the current version of all the code in the functions.R file

```
> x <- grab.data(symbols = secref$symbol, years = 1998:2001)
> dim(x)
[1] 2233504       8
> y.1 <- calc.returns(x, d.before = 0, d.after = 182)
> dim(y.1)
[1] 108296      3
> names(y.1)
[1] "symbol"      "ret.0.182.d" "date"
> summary(y.1$ret.0.182.d)
     Min.    1st Qu.     Median       Mean    3rd Qu.       Max.       NA's
 -0.99800   -0.16110    0.02831    0.14010    0.23720  499.00000 3083.00000
>
```

And that looks OK to me. Note that I am using all securities. Note that I am focusing on just those dates in the first three years, which I think are the source of the problem. Note that I am using all rows (not throwing out month/symbol combinations because they are not in top 1500 or don't have a 52 week high or whatever.

Ahhh, now I see better! Although almost all this data is fine, one company is a huge problem, as you pointed out. Consider:

```
> subset(y.1, ret.0.182.d > 300)
     symbol ret.0.182.d       date
83444 3STTCE    362.6364 1998-06-30
85696 3STTCE    306.6923 1998-07-31
90207 3STTCE    475.1905 1998-09-30
92465 3STTCE    499.0000 1998-10-31
>
```

The key is the 300 here is not 300% (which can happen), but 30,000%, which is ridiculous. There are also some other suspect names. Consider:

```
> unique(subset(y.1, ret.0.182.d > 10)$symbol)
 [1] "IRC"    "BMTS"   "INRG"   "3DIGL"  "CMRCQ"  "CRA"    "CRGN"   "CTIC"
```

```
 [9] "IMMU"    "MEDX"    "MYGN"    "SYNC"    "TIBX"    "ORTL"    "3CBHDE" "3MFNF"
[17] "3STTCE" "CHTM"    "GNET.1" "AMTD"    "ETFC"    "LDIG"    "NITE"    "IPI."
>
```

Although it is possible for a stock to go up ten times on a six month period, it is highly unusual, especially for a stock that is already in the top 1500. If you are already a $5 billion market cap company, then it usually takes much longer than 6 months to grow to $50 billion.

How many of these names are actually in your universe when you are forming portfolios? List them out in an e-mail to the list and we can look into them.

I realize that this is a bother, but this is exactly how professionals conduct research. The data are always a mess, filled with lies.

Wow. This is really interesting.

In retrospect, I should have devoted more class time to an example like this, to illustrate the process by which we look for a deal with bad data. My bad.

But it is not too late! We can still look for a deal with bad data. Let's take a tour with 3STTCE

```
> subset(y.1, ret.0.182.d > 300)
    symbol ret.0.182.d       date
83444 3STTCE    362.6364 1998-06-30
85696 3STTCE    306.6923 1998-07-31
90207 3STTCE    475.1905 1998-09-30
92465 3STTCE    499.0000 1998-10-31
> subset(secref, symbol == "3STTCE")
         id symbol              name m.sec m.ind
1753 02976801 3STTCE STRATOSPHERE CORP   CND HOTEL
> subset(yearly, symbol == "3STTCE")
          id symbol year    cap.usd top.1500
1753  02976801 3STTCE 1998 1167860000     TRUE
4836  02976801 3STTCE 1999   42630000    FALSE
7919  02976801 3STTCE 2000   85006250    FALSE
11002 02976801 3STTCE 2001   86478000    FALSE
14085 02976801 3STTCE 2002   89320000    FALSE
17168 02976801 3STTCE 2003         NA    FALSE
20251 02976801 3STTCE 2004         NA    FALSE
23334 02976801 3STTCE 2005         NA    FALSE
26417 02976801 3STTCE 2006         NA    FALSE
29500 02976801 3STTCE 2007         NA    FALSE
>
```

Hmmm. Was Stratosphere really a billion dollar company on Dec 31, 1997. Perhaps. Consider this 10-Q

http://www.secinfo.com/dRc22.9ub.htm

Note that we only need to look at the company for 1998 since it is FALSE in yearly thereafter.

Now, it would take a lot more work to determine whether this is "good" data. There are at least two potential problems. First, was their cap really that high at the end of 1997? That is tough to know for sure and easy to make a mistake in calculating. Second, even if it really belonged in the universe for 1998, was it actually getting the returns that we see.

```
> data(daily.1998)
> temp <- subset(daily.1998, symbol == "3STTCE")
> dim(temp)
[1] 207   8
> summary(temp)
     id               symbol              v.date             price.unadj
 Length:207         Length:207         Min.   :1998-01-02   Min.   : 0.0450
 Class :character   Class :character   1st Qu.:1998-03-18   1st Qu.: 0.0680
 Mode  :character   Mode  :character   Median :1998-06-02   Median : 0.0800
                                       Mean   :1998-06-02   Mean   : 0.5042
                                       3rd Qu.:1998-08-13   3rd Qu.: 0.1295
                                       Max.   :1998-12-29   Max.   :25.5000
     price          volume.unadj         volume             tret
 Min.   : 0.0450   Min.   :    100   Min.   :    100   Min.   : -0.40741
 1st Qu.: 0.0680   1st Qu.:  59950   1st Qu.:  59950   1st Qu.: -0.05985
 Median : 0.0800   Median : 108100   Median : 108100   Median :  0.00000
 Mean   : 0.5042   Mean   : 224796   Mean   : 224796   Mean   :  0.81244
 3rd Qu.: 0.1295   3rd Qu.: 261000   3rd Qu.: 261000   3rd Qu.:  0.07418
 Max.   :25.5000   Max.   :1896500   Max.   :1896500   Max.   :165.66667
>
```

Surely looks to me like this was mostly trading as a many stock and that it then did a reverse split of some sort. That can easily mess up return calculations.

But, there is really just a single day problem here.

```
> head(sort(temp$tret, dec = TRUE))
[1] 165.6666667   0.8000000   0.5714286   0.4166667   0.3600000   0.3600000

> subset(temp, tret > 1)
             id symbol     v.date price.unadj price volume.unadj volume
409370 02976801 3STTCE 1998-11-24          10    10          500    500
           tret
409370 165.6667
>
```

I think that this one day return is the source of all the problems with this name. Delete that one day (which is almost certainly false data) and everything is OK.

I bet that most of the other problems are very similar. Consider the largest one day returns.

```
> head(sort(x$tret, dec = TRUE), 20)
 [1] 165.666667 159.000000  36.500000  19.000000   9.571429   9.000000
 [7]   7.428571   5.064516   4.500000   4.460317   3.333333   3.248000
[13]   3.000000   3.000000   3.000000   3.000000   2.571429   2.529851
[19]   2.500000   2.500000
>
```

I recommend just deleting any *one day* return that is greater than 1 (meaning 100%). Almost all of these are data errors.

But make sure that you describe the fact that you are doing that in the .Rnw and why! You should also mention this in the paper.

I have some issues with the caption on table 2.

1) Are you using one month forward returns or 6 month? You need to make that clear. I assume that it is one month.

2) One easy way to make the table more clear is just to expand the labels at the far left. (You have lots of room.) Instead of just "Winner", how about "Industry Winner" and then "Industry Neutral" and so on. Then, on the second column, have "52-Week-High Winner" and so on.

3) Are you sure that your dates are right? How can your first date be 1998-01-31? At the very least you need 6 month prior returns, if not 52 week highs. I think that your first portfolio is formed 1998-12-31.

4) You need some info in the table about average number of stocks each month with data. Obviously, you go into much more detail (?) about this in the main write up, but you need something here.

Similar with Table 3, we need some more detail on the caption. What is the exact breakdown in terms of frequency that you are using to split stocks into three groups. How many stocks do you have in a given month on average?

You need a conclusion, obviously.

In terms of the data and run time, it is fine to just save out your main results and load() them up when you want to "recompile" the Sweave. In other words, you do not need to re-run every calculation every time you add a couple sentences to the Sweave and want to see how the pdf looks.

The key is that your .Rnw still has the code to calculate everything, albeit commented out. Instead, it just goes:

load("mydata.Rdata") at the key point (which probably just loads the key dataframe z with everything (52 week highs, 6 month past returns, et cetera) already done.

Then the code that runs backtest, builds the tables and so on should only take a few seconds and can be redone every time you run Sweave().

Let's focus on these data problems. For now, I'll just look at 1998 to 2000 since I suspect that most of the problems are there.

```
> x <- grab.data(symbols = secref$symbol, year = 1998:2000)

> dim(x)
[1] 1686436       8
> head(x)
        id symbol price.unadj   price volume.unadj volume         tret
1 00100401    AIR     28.0000 28.0000        76500  76500 -0.004444444
2 00100401    AIR     28.1250 28.1250        43700  43700 -0.006622517
3 00100401    AIR     28.3125 28.3125        19400  19400 -0.015217391
4 00100401    AIR     28.7500 28.7500       152100 152100  0.010989011
5 00100401    AIR     28.4375 28.4375        46100  46100  0.006637168
6 00100401    AIR     28.2500 28.2500        91600  91600 -0.004405286
        date
1 1998-04-20
2 1998-04-17
3 1998-04-16
4 1998-04-15
5 1998-04-14
6 1998-04-13
> x <- x[order(x$symbol, x$date),]
> head(x)
            id symbol price.unadj  price volume.unadj volume         tret
41337 00281201  0491B      45.750 45.750        89800  89800  0.000000000
41336 00281201  0491B      46.000 46.000       226400 226400  0.005464481
41335 00281201  0491B      45.938 45.938       192200 192200 -0.001347826
41334 00281201  0491B      44.750 44.750       203900 203900 -0.025860943
41333 00281201  0491B      45.438 45.438       296700 296700  0.015374302
41332 00281201  0491B      43.500 43.500       162700 162700 -0.042651525
            date
41337 1998-01-02
41336 1998-01-05
41335 1998-01-06
41334 1998-01-07
41333 1998-01-08
41332 1998-01-09
>
```

We will find it easier to have things sorted like this. And I also want the row numbers to be correct.

```
> row.names(x) <- 1:nrow(x)
> options(width = 100)
> head(x)
        id symbol price.unadj  price volume.unadj volume         tret       date
1 00281201  0491B      45.750 45.750        89800  89800  0.000000000 1998-01-02
2 00281201  0491B      46.000 46.000       226400 226400  0.005464481 1998-01-05
3 00281201  0491B      45.938 45.938       192200 192200 -0.001347826 1998-01-06
```

```
4 00281201  0491B     44.750 44.750      203900 203900 -0.025860943 1998-01-07
5 00281201  0491B     45.438 45.438      296700 296700  0.015374302 1998-01-08
6 00281201  0491B     43.500 43.500      162700 162700 -0.042651525 1998-01-09
>
```

Now, let's find the bad rows. I am deeply suspicious of any stock that is up more than 100% in a day. Unfortunately, that happens 70 times!

```
> dim(subset(x, tret > 1))
[1] 70  8
>
```

One simple solution (done by almost all the papers that we have read) is to get rid of companies that have a price that is too low. After all, very few portfolios are run a stocks that are less than $5. So, any stock with an *unadjusted* price (the price in the newspaper at the time) of less than 5 is likely to be either a) a data error or b) a stock that we would not have want to invest in. This is only 3% of the sample.

```
> sum(x$price.unadj < 5)/length(x$price.unadj)
[1] 0.03328736
>
```

How many does that screen take care of among our problem stocks?

```
> dim(subset(x, tret > 1 & price.unadj > 5))
[1] 18  8
>
```

52 out of 70! Not bad. So, I would recommend dropping all stock/month rows (only at the point that you have z and *not* before) that have an price.unadj less than 5. Make sure you describe that procedure (or whatever you do) in the Data section.

More later.

Continuing with this example, let's look more closely at the remaining 18 problem rows.

```
> subset(x, tret > 1 & price.unadj > 5)
          id symbol price.unadj      price volume.unadj   volume
   tret       date
24290   01531701  3MFNF     6.00000    6.00000        200      200
36.500000 1999-05-07
30465   02976801 3STTCE    10.00000   10.00000        500      500
165.666667 1998-11-24
30552   02976801 3STTCE    47.00000   47.00000        500      500
1.238095 2000-03-29
138422  12145301   APNT    33.00000   33.00000    13925000 13925000
1.129032 1999-09-10
340729  00294601   CHTM  1400.00000 1400.00000          3        3
159.000000 1998-12-08
```

```
522277  03045001   EGCO     8.00000    8.00000         1300     1300
1.666667 1999-08-19
522286  03045001   EGCO     6.25000    6.25000          100      100
1.083333 1999-09-17
660427  12432101   GLDN    32.00000   32.00000       930000   930000
1.639175 1999-12-31
746595  06629301  HSM.1    28.56250   28.56250     11598700 11598700
1.135514 2000-05-12
799857  00376901   INRG    46.25000   46.25000     17406100 17406100
9.571429 2000-09-22
846504  12081801   JWEB    66.75000   66.75000     29068600 29068600
1.301724 1999-12-21
889028  02322701   LDIG    29.56250   29.56250      6851400  6851400
2.529851 1999-04-06
994105  06150701   MMWW    33.32812   33.32812     19497300 19497300
1.083008 2000-03-22
1050757 01361401    NDB    31.50000   31.50000      3747300  3747300
1.250000 1998-12-29
1091978 06168501   NUAN     8.59375    8.59375     21928200 21928200
1.083333 1999-11-11
1187799 02389701   PLAT    24.06250   24.06250     42027300 42027300
1.436709 1999-03-29
1192448 12560401   PLUG    79.00000   79.00000      5946300  5946300
1.179310 2000-01-07
1322391 06398401   SCAI    79.87500   39.93750      3544600  7089200
1.505882 1999-04-13
>
```

Some of these have suspiciously low volume. Others are probably not in the top 1500 for the appropriate year. How many are, like 3STTCE, just coming out of a weird corporate action?

To see this, let's use the row numbers we know to grab out those rows and the ones before.

```
> keys <- as.numeric(row.names(subset(x, tret > 1 & price.unadj > 5)))
> keys
 [1]   24290   30465   30552  138422  340729  522277  522286  660427
746595  799857  846504  889028
[13]  994105 1050757 1091978 1187799 1192448 1322391
>
```

Let me intersperse my comments:

```
> x[sort(c(keys, keys - 1)),]
           id symbol price.unadj      price volume.unadj   volume
    tret       date
24289   01531701  3MFNF     0.16000    0.16000      1183700  1183700
0.03225806 1999-04-23
24290   01531701  3MFNF     6.00000    6.00000          200      200
36.50000000 1999-05-07
```

Data problem! How can the stock be up 36 times on May 7 if it did not trade on May 6 or even any day after April 23.

30464   02976801 3STTCE   0.06000   0.06000   106300   106300
0.07142857 1998-10-20
30465   02976801 3STTCE   10.00000   10.00000   500   500
165.66666667 1998-11-24

Same problem we noted before. Again, we (obviously!) don't have the time to go through everyone of these by hand. We need a rule for dealing with all/most of them. But this kind of painstaking detailed grunt work is a big part of actual research. I am a bad professor for not introducing it to you in week 2. But better late than never!

30551   02976801 3STTCE   21.00000   21.00000   800   800
0.00000000 2000-03-28
30552   02976801 3STTCE   47.00000   47.00000   500   500
1.23809524 2000-03-29

At least here we have trading the day before. So, perhaps this data is OK. The key is to look to see if the change in price (which has been adjusted for splits) is consistent with the tret column. And, here is looks OK. If 3STTCE closed at $21 on March 28 and then went up to $47 on March 29, then, sure enough, it was up 123% that day. So, this seems OK. But recall that 3STTCE was not even in the universe by 2000, so this is a data point that doesn't matter.

138421 12145301   APNT   15.50000   15.50000   67900   67900
-0.02362205 1999-09-09
138422 12145301   APNT   33.00000   33.00000   13925000 13925000
1.12903226 1999-09-10

Again, this seems OK. The change in price is consistent with the return.

340728 03091401 CHSWQ   0.68750   0.68750   328500   328500
0.00000000 2000-04-03
340729 00294601   CHTM 1400.00000 1400.00000   3   3
159.00000000 1998-12-08

Again, it seems like every return above 2 is an error. CHTM first traded on Dec 8 and then didn't trade for a month. That's just junk.

```
> head(subset(x, symbol == "CHTM"))
           id symbol price.unadj price volume.unadj volume
tret       date
340729 00294601   CHTM        1400  1400            3      3
159.000000000 1998-12-08
340730 00294601   CHTM        1345  1345            2      2
-0.039285714 1999-01-13
340731 00294601   CHTM        1345  1345            3      3
0.000000000 1999-03-19
340732 00294601   CHTM        1480  1480            3      3
0.100371747 1999-03-30
340733 00294601   CHTM        1485  1485            2      2
0.003378378 1999-06-04
```

```
340734 00294601   CHTM       1485  1485          6      6
0.000000000 1999-06-15
>
```

The next example, the data is OK:

```
522276  03045001   EGCO    3.00000    3.00000        2000     2000
-0.62500000 1999-08-17
522277  03045001   EGCO    8.00000    8.00000        1300     1300
1.66666667 1999-08-19
522285  03045001   EGCO    3.00000    3.00000         300      300
-0.50000000 1999-09-15
522286  03045001   EGCO    6.25000    6.25000         100      100
1.08333333 1999-09-17
```

but, again, we are saved because 1999 is not a top1500 year for this stock.

```
> subset(yearly, symbol == "EGCO")
         id symbol year   cap.usd top.1500
1805  03045001   EGCO 1998        NA    FALSE
4888  03045001   EGCO 1999        NA    FALSE
7971  03045001   EGCO 2000        NA    FALSE
11054 03045001   EGCO 2001  16140600    FALSE
14137 03045001   EGCO 2002 575268750    FALSE
17220 03045001   EGCO 2003 129253500    FALSE
20303 03045001   EGCO 2004  42237510    FALSE
23386 03045001   EGCO 2005   5619315    FALSE
26469 03045001   EGCO 2006  13499780    FALSE
29552 03045001   EGCO 2007   6232320    FALSE
>
```

Matter of fact, EGCO is never in the top1500. Hmmmm. How did he get in the dataset? Fortunately, we don't have to worry about that since he doesn't matter.

And so on. I'll leave the rest of these to you to consider:

```
660426  12432101   GLDN    12.12500   12.12500       10000    10000
0.01041667 1999-12-30
660427  12432101   GLDN    32.00000   32.00000      930000   930000
1.63917526 1999-12-31
746594  06629301  HSM.1    13.37500   13.37500      166700   166700
0.02884615 2000-05-11
746595  06629301  HSM.1    28.56250   28.56250    11598700 11598700
1.13551402 2000-05-12
799856  11148801   INKT    17.87500   17.87500    14887000 14887000
-0.10903427 2000-12-29
799857  00376901   INRG    46.25000   46.25000    17406100 17406100
9.57142857 2000-09-22
846503  12081801   JWEB    29.00000   29.00000     9334700  9334700
0.77099237 1999-12-20
846504  12081801   JWEB    66.75000   66.75000    29068600 29068600
1.30172414 1999-12-21
889027  02322701   LDIG     8.37500    8.37500      605400   605400
```

```
0.52272727 1999-04-05
889028  02322701   LDIG    29.56250   29.56250     6851400  6851400
2.52985075 1999-04-06
994104  06150701   MMWW    16.00000   16.00000      678300   678300
-0.01538462 2000-03-21
994105  06150701   MMWW    33.32812   33.32812    19497300 19497300
1.08300781 2000-03-22
1050756 01361401    NDB    14.00000   14.00000      235900   235900
-0.14503817 1998-12-28
1050757 01361401    NDB    31.50000   31.50000     3747300  3747300
1.25000000 1998-12-29
1091977 06168501   NUAN     4.12500    4.12500     6686600  6686600
1.12903226 1999-11-10
1091978 06168501   NUAN     8.59375    8.59375    21928200 21928200
1.08333333 1999-11-11
1187798 02389701   PLAT     9.87500    9.87500     1188100  1188100
0.03267974 1999-03-26
1187799 02389701   PLAT    24.06250   24.06250    42027300 42027300
1.43670886 1999-03-29
1192447 12560401   PLUG    36.25000   36.25000     1652500  1652500
0.22881356 2000-01-06
1192448 12560401   PLUG    79.00000   79.00000     5946300  5946300
1.17931034 2000-01-07
1322390 06398401   SCAI    31.87500   15.93750      836500  1673000
0.16438356 1999-04-12
1322391 06398401   SCAI    79.87500   39.93750     3544600  7089200
1.50588235 1999-04-13
>
```

Summary: trets above 2 are almost always a mistake. trets below are often correct, even if they are unusually. My advice:

1) After calc.returns (when you have daily returns for all stocks, all days), substitute a tret of 0 for any tret > 2. This is almost guaranteed to make the data more accurate.

2) After you have made z, throw out any rows with an unadjusted stock price (for that day) of less than 5. Papers do this all the time, including, I think, the ones that you are replicating.

I bet that these two changes will make your month-by-month results much more plausible.

On the new abstract:

Abstract is good. Are you sure that you a replicating "comparisons"? I think that you are replicating the different momentum strategies discussed in George and Hwang (2004).

"its" should be "their".

Check your spelling here and elsewhere: "forcasting" should be "forecasting".

"in 6 month horizons" should be "over 6 month horizons"

Not sure that the last sentence makes sense. But it is good to have some specific numbers. You want something more like: Portfolios formed using stocks which are within 3 months of their 52 week high have a spread of 8%, twice as much as the spread for portfolios created from stocks that are more than 9 months from their 52 week high. (Or whatever).

I thought that I gave you feedback on Data/Methods. Have you gone through all 15 e-mails?

:-)

I did not give you much feedback on the Results and Extensions sections because they were still quite rough. Lots of typos, incomplete sentences and so on. But here are some brief thoughts on the latest draft:

Results:

1) "between One the first row" huh?

2) The usage of "the JT's individual stock momentum" is awkward. You can have the "the" or you can have the apostrophe s. But both together are weird. Either "the JT individual stock momentum" (my recommendation) or "JT's individual stock momentum".

3) Didn't I review some of this already? Things like (Figure.1)? Anyway, clean up the obvious issues and I will look more closely at the next draft, if you like.

Extensions:

Let me focus on substance. Note that this is still a mess with the first two paragraphs largely identical. Fix all the stuff that you can fix yourself. Try reading the paper out loud to your partner. Once those issues are fixed, I can focus on issues that are my comparative advantage.

Are you really modifying their ranking methodology? I don't think so. But until I see the rest of Table 3, I can't tell.

Instead, I *think* that you are using the same methodology (score is a function of 52 week high and current price) but you are then segregating the universe into three parts on the basis of how recent the 52 week high is. And then you are pointing out that the score (same score that they use) is much more powerful in the more recent portion of the universe. 52 week high is useless as a forecast of future 6 month returns if the high was more than 6 months ago. (Not sure if you find that or something like it).

Ahhhh. I had misunderstood what you were doing, which means that you want to emphasize the point more clearly in the caption, as well as the abstract, introduction and conclusion.

I *thought* that you were decomposing the universe of all stocks into three categories in terms of distance from 52 week high and then using that as a by.var in measuring the actual 52 week high strategy. (Which still might be interesting to do.) This would lead to a table with three rows (and perhaps an overall as well for reference), showing that, for stocks whose 52 week high was in the last 3 months (that 1/3 of the universe), the GH spread is 6% (the spread between stocks that are near their high and stocks that are from from their high), for those whose 52 week high was between 3 and 6 months ago, the GH spread is 3% and so on.

But instead (and just as interesting), you are saying "Forget GH. Look at our cool new idea." That is a fine approach and, of course, you just need one line to tell the overall story.

But every smart reader is going to look at that row and say, "Wait a second! Isn't this the same as GH? I bet that companies that are near their 52 week high in terms of recency (your new idea) are also near the high in terms of price (as in GH). And the opposite for those far away."

So, to answer that reader, you want some sort of pairwise.

Comments on previous draft of Fraulo and Nguyen

Much better! Although there is still work to be done, you guys are obviously working very hard.

What is very cool about this process, and the tools that you are using, is that it is now very easy for me and the other participants to check your work. For example, all I need to do is download your .Rnw file and type:

```
> Stangle("c:/Documents and Settings/David Kane/Desktop/Rstuff/draft2")
Writing to file draft2.R
>
```

(Obviously, others need to give a path that works for them.)

Now, I can just load the draft2.R code (either all at once or step by step) and I can replicate your results exactly. I am doing that now and may have comments later.

I encourage all our participants to do the same.

Consider this code chunk:

```
y.1 <- calc.returns(x, d.before = 30, d.after = 0, actual = TRUE)
y.2 <- calc.returns(x, d.before = 0,   d.after = 30, actual = TRUE)
```

```
z <- merge(y.1, y.2)

## Calculate 6 month previous and forward returns. Bring data into z
y.3 <- calc.returns(x, d.before = 182, d.after = 0, actual = TRUE)
y.4 <- calc.returns(x, d.before = 0, d.after = 182, actual = TRUE)
z$ret.182.0.d <- y.3$ret.182.0.d[match(paste(z$
symbol, z$date),
paste(y.3$symbol, y.3$date))]
z$ret.0.182.d <- y.4$ret.0.182.d[match(paste(z$symbol, z$date),
paste(y.4$symbol, y.4$date))]

## Remove the first 6-months and last 6-months of data to account for
calc.returns function problem
z <- z[z$date >= "1998-06-30" & z$date <= "2007-06-30",]
```

1) Is there any reason why you merge the one month return data but match in the 6 month data? This isn't wrong per se, but I was curious. And, whatever your reason, you want to explain it in your .Rnw file.

The effect is that you have a row for every stock that has a one month forward and previous return, even if those stocks are missing either (or both) six month returns. Again, this is not wrong, but it does require you to use na.rm = TRUE later when you are calculating industry returns over a six month window (which you do).

Hmmmm. But now that I run your data, I see that I have not thought this through. Consider:

```
> summary(z)
   symbol              date              ret.30.0.d
 Length:257848      Min.   :1998-01-30   Min.   :-1.000000
 Class :character   1st Qu.:2000-05-31   1st Qu.:-0.056724
 Mode  :character   Median :2002-11-29   Median : 0.007928
                    Mean   :2002-11-29   Mean   : 0.015144
                    3rd Qu.:2005-04-29   3rd Qu.: 0.074951
                    Max.   :2007-11-30   Max.   : 9.833333
  ret.0.30.d         ret.182.0.d        ret.0.182.d
 Min.   :-1.000000   Min.   :-1.00000   Min.   :-1.00000
 1st Qu.:-0.057797   1st Qu.:-0.11027   1st Qu.:-0.11868
 Median : 0.006394   Median : 0.04952   Median : 0.04161
 Mean   : 0.012814   Mean   : 0.09821   Mean   : 0.08777
 3rd Qu.: 0.071925   3rd Qu.: 0.22316   3rd Qu.: 0.21515
 Max.   : 9.000000   Max.   :31.09877   Max.   :31.09877
>
```

There are no missing values for 30 day returns (which we expect because of the merge) but nor are their missing values for 182 day. And, now that I think of it, the answer is obvious: calc.returns does not require x days of data when it calculates x day returns. In fact, all it requires is 1 day of data! Does that seem problematic to anyone? It does to me. But, obviously, you should not worry about it for now.

2) I am surprised that your date removal code works

```
z <- z[z$date >= "1998-06-30" & z$date <= "2007-06-30",]
```

In previous versions of R this would have failed because "1998-06-30" is a character not a date. You used to have to convert it by as.Date("1998-06-30"). But, R now handles this automatically:

```
> dim(z)
[1] 257848      6
> z <- z[z$date >= "1998-06-30" & z$date <= "2007-06-30",]
> dim(z)
[1] 237136      6
> summary(z$date)
        Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
"1998-06-30" "2000-08-31" "2002-11-29" "2002-12-04" "2005-02-28" "2007-06-29"
> class(z$date)
[1] "Date"
>
```

which is pretty cool. This exercise is teaching me a bunch of stuff too!

Here are more thoughts on your code.

1) As I think about it, perhaps the usage of calc.returns as it is now is OK. Again, the appropriate metaphor is a time machine (and to imagine that you are actually running a portfolio with real money). You go back to June 30, 2002, you buy/short a bunch of stocks that you intend to hold for 180 days. Does it matter to you if some of them go bankrupt or are bought out by another company in just 2 days or 20 or 90? No! You are still stuck with the gains/losses for those positions even if those stocks don't last the full 182 days. So, calc.returns is doing what it should and all your results are fine. You get the same answer whether you merge in or match in the 182 returns.

2) Notice how much longer this match:

```
z$price <- x$price.unadj[match(paste(z$
symbol, z$date),
paste(x$symbol, x$date))]
```

than almost all the other matches? Exercise for our participants: Why does that happen? I am pretty sure that the reason is that R has trouble with converting things from dates, which is what it needs to do with the paste command. See the error that I get here:

```
> z$temp <- as.character(z$date)
> x$temp <- as.character(x$date)
Error: cannot allocate vector of size 61.9 Mb
```

Of course, this probably doesn't cause an error for you since you have a more powerful machine. So, I need to do a total hack.

```
> temp1 <- as.character(x$date[1:2000000])
> temp2 <- as.character(x$date[2000001:length(x$date)])
> x$temp <- c(temp1, temp2)
> z$price <- x$price.unadj[match(paste(z$symbol, z$temp), paste(x$symbol, x$temp))]
```

but then the last command runs almost instantaneously because temp is a character vector rather than a date. Anyway, after the last subset for price and top1500, I have:

```
> dim(z)
[1] 153358     12
> summary(z)
   symbol              year                date              ret.30.0.d
 Length:153358      Length:153358      Min.   :1998-06-30   Min.   :-0.81883
 Class :character   Class :character   1st Qu.:2000-09-29   1st Qu.:-0.04286
 Mode  :character   Mode  :character   Median :2003-01-31   Median : 0.01344
                                       Mean   :2003-01-12   Mean   : 0.02212
                                       3rd Qu.:2005-04-29   3rd Qu.: 0.07356
                                       Max.   :2007-06-29   Max.   : 6.94702
  ret.0.30.d         ret.182.0.d         ret.0.182.d          industry
 Min.   :-0.82304   Min.   :-0.94481   Min.   :-0.99574   Length:153358
 1st Qu.:-0.04578   1st Qu.:-0.06315   1st Qu.:-0.07084   Class :character
 Median : 0.01083   Median : 0.07618   Median : 0.06416   Mode  :character
 Mean   : 0.01737   Mean   : 0.13612   Mean   : 0.10643
 3rd Qu.: 0.06957   3rd Qu.: 0.23762   3rd Qu.: 0.21649
 Max.   : 5.11728   Max.   :31.09877   Max.   :31.09877
    price                id               cap.usd          top.1500
 Min.   :     5.01   Length:153358      Min.   :5.782e+08   Mode:logical
 1st Qu.:    22.91   Class :character   1st Qu.:1.493e+09   TRUE:153358
 Median :    33.30   Mode  :character   Median :2.692e+09   NA's:0
 Mean   :    95.73                      Mean   :9.655e+09
 3rd Qu.:    47.25                      3rd Qu.:6.797e+09
 Max.   :110050.00                      Max.   :6.024e+11
>
```

which looks reasonable. Note that year is character rather than numeric or integer (as it is in yearly). That's because format returns a character.

```
z$year <- format(z$date, "%Y")
```

R automatically converts the year in yearly to character in your merge.

```
z <- merge(z, yearly)
```

Again, that all works fine, but you need to be careful of this sort of stuff. I would have forced year to be integer in z.

```
z$year <- as.integer(format(z$date, "%Y"))
```

Anyway, in summary, all your code looks fine so far and seems to be doing what you expect it to do.

Well done!

You have made some good progress here. Again, I agree that your data is correct (or, at least, I can't find any mistakes) so you can focus on your write up and then extensions. I will break up my comments into a series of e-mails.

1) If you wanted to have an extension until Wednesday morning start of classes, I would be willing to grant one. Let me know. (Same applies to Bill/Vincent, but I sense that they are closer to finishing.) Your choice. I would feel guilty if you missed the Super Bowl to work on the paper!

2) This version of the abstract is much better. But

a) Do not use the MG abbreviation here. You can repeat their full names/date once or twice.

b) Check over the abstract again and again. Ten times more people (like the Williams econ department when I invite them to your presentation) will read the abstract then will read the paper. It should be perfect.

c) "returns that" should be "returns than".

d) Your third sentence should be your second and your second your third. First you explain what industry momentum strategies are and then you explain how they compare to individual stock strategies. Your first sentence is just about perfect. After it, I would go with "Industry momentum strategies buy stocks from . . ." You don't need to preface that with "MG also demonstrate". The same with the next sentence. We all know that you are talking about MG since you tell us that at the beginning.

3) See my previous comments on the first sentence of the Introduction. It still needs work. But my comments are the same.

4) Although the Introduction is not as important as the Abstract, it is the second most important part. Make sure that your sentences are precise. "this phenomenon, such as George and Hwang (2004)" is *gibberish.* The academic paper George and Hwang (2004) has nothing to do with phenomenon or characteristics or even, directly, the momentum effect. The reference here it to one of the "attempts", but that occurs so far earlier in the sentence that the connection can not be made by the reader.

----
1) I like the way that you out both the backtest article and the Ihaka R article in your references. (Bill and Vincent *must* do the same.) But do you site the articles anywhere in this draft? Not that I can see. I would site the Ihaka article in the footnote on the first page as "The code which replicates the results of this paper was written in R (Ihaka and Gentleman (1996) and is available from the authors." or something like that. It is good to mention R explicitly and good to be

clear that you will provide code if asked. You could site the backtest article in your results section, mentioning (in a fottnote) that you use the backtest package and then citing the article. Bill and Vincent should do the same.

2) I think that you also fail to use in the text several other references. Fix that.

3) Fix the x-axis date labels to figure 1. Did you try the trick that I sent out to the list? Since you have fewer months, you might need to tweak the length of the index.

Getting back to the Introduction

4) "to 2007" should be "through 2007" to make clear that all of 2007 is in the data.

5) You mention "equal sized groups" but, in the methods, you talk about 30/40/30 groups. Which is it?

6) Fix the last sentence of the "Our data" paragraph. You should explicitly say, "We term these strategies (6, 6) and (1, 1), respectively."

7) It is good (read: required) to cite a specific number in your Introduction, as you do at the end of the next paragraph with 0.014, but you need to make clear exactly what that means. It is worth a whole sentence. Does 0.014 mean 1.4% per month? I think it does, but that is a clearer way of saying it. It is also helpful to put that same number in an annual context, "1.4% per month or almost 17% per year, ignoring transaction costs."

It is nice to have an annual number, which is often easier for people to understand. And, although we did not talk about this much, we need to at least mention one time that we are ignoring transaction costs.

8) The last paragraph is a bit of a mess. It is good to end with a discussion of your extension, so the beginning is fine. But the last 2 sentences (starting "It is also interesting . . .") do not belong here. They belong in the introduction, but a paragraph or two before. You want to end with a bang, with the specific result of your extension.

Write as tightly as possible. (Better: Write tightly.)

Consider the opening sentence of Data and Methods: Instead of the 6 words "We use our data that consists of", how about just 3: "Our data includes". Insead if "that span," use "over". And so on.

Note that you are not really covering a "wider range of industries." The "range" is the same. You have a more detailed, or at least more numerous, collection of industries. If you want to make this comparison, tell us exactly how many they had.

Do not use "top 1500" in quotes, much less mention the variable name. Readers do not care what your variable names are. And top1500 is not particularly standard, which is why you don't see it in other papers.

second paragraph of D&M

"between July 1998 and June 2007 . . " I think you need to mention the month here so that you are clear that you are not testing on the whole year. (Although you form your first portfolio on midnight on June 30, 1998, it only uses returns for July and after to measure performance.)

"If data is unavailable . . . " is a bad sentence. What dates are you talking about? Don't you have data for every month? Aren't there 109 month ends in that range? I suppose that you are talking about return measurements, but even that is unclear. I have a sense of what you are trying to say. The trick is that you need to say it slowly and clearly. Just this info (missing returns) is worth an entire *paragraph*. Give us all the messy details. Go step by step. What happens to a company that is only two months old? What are its 6 month previous returns? Take the time and the space to be clear.

The next sentence "However, this is ..." is similar awkward. Read it aloud. Is it clear? You are, of course, talking about a very different concept here, data that is missing for everyone, not just for a single company. And, again, the way to handle this is to spend several sentences on just this concept. Walk us through things slowly. "Since our daily returns only start on January 3, 1998, the first date that we have six months prior return for companies is June 30, 1998" and so on.

I realize that just these two things (and a couple of others) will add another page to the article. But that is a good thing. You have a lot to explain.

Much of this applies to Bill and Vincent. Your write up about your methods is not quite as detailed as the actual code itself, but it is close.

Still on Data and Methods:

1) The sentence about 200% returns and $5 price is probably accurate, but is certainly too quick. Do you remove a stock if it has a one month return of 200%? No! You keep those. You only return daily returns that are that big. (And the reason you remove them --- and you should say this --- is because they are almost all data errors. In fact, because they are data errors, it is not clear that you need to mention it here. Your call.

But the key is that this is a process that occurs with the *daily* data, before you calculate the longer returns. That is a separate discussion (paragraph) from the discussion of what rows you remove from z.

And so your description of $5 is wrong to. If a stock is priced at $4 on June 15, is it removed? No! The only prices that matter are the prices on the last day of the month. You need a whole paragraph devoted to this. Describe each detail. Do it slowly. You do discuss the top 1500 a bit, but even here it is not clear. You need sentences like, "In order to be included in a portfolio for a given date, say June 30, 1999, a stock must have been in the top 1,500 by market capitalization as of December 31 of the previous year, December 31m 1998 in this case. Also, the stock must trade above $5 on the last trade date before month end. In this case, the last trade date was June 28 (just made that up) because June 29 and 30 were on the week-end."

Blah, blah, blah. Slowly and thoroughly.

"Industries are formed monthly?" No!!! A stock is assigned to an industry and that assignment is fixed for the entire ten years.

Try to get Table 1 to print on page 3 (first page after the abstract). It is nice to space the tables and figures throughout the text. One way to do that is to put the table much earlier in the .Rnw. (I *think* that no matter how early you put it, it won't appear on the first page.)

Your extension is an interesting idea and I like how you provide a discussion of the underlying investor behavior. But be very careful that you describe things accurately. You write that people (really "investors") act "irrationally." That is OK, but irrationality is fairly strong language, as if investors just got out of the loony bin. Is that fair? Do the papers you read use that language? (They may have, but I think less strong descriptions like "mistaken" or "sub-optimally" are more the norm.

You use the phrase "as their returns get higher." First, that's awkward phrasing. Second, is that really what you mean? If a stock has moved between $5 and $20 over the last year, and I bought it at $6 while it is now at $10, does that mean that I am about to act irrationally (or, at least, make a mistake) since my returns are higher and or have gone higher? I don't think GH make that claim. In fact, note that distance away from the 52-week low seems to play no part in their results. Instead, I see their story as much more driven by nearness to the 52-week high and, in particular, the reluctance of investors to buy at this point, even in the face of good news, because the stock is "too high." So, good news has happened but the stock price has not adjusted. So, I should buy now (even if I don't know of any good news in particular) since, on average, there is unrecognized good news in stocks near their 52 week-high, news that is eventually incorporated in the price.

I am not saying that this story is true, but it seems close to what GH are claiming than any phrase involving "their returns get higher".

But the important thing is that it is good for you to discuss this underlying economic behavior. My main suggestion is to do so at greater length and greater care. Study how this discussion is handled in all the papers in the syllabus.

Table 1 has come a long very nice. Note that "69industries" is a typo.

Figure 1 needs better x-axis. Try my suggestion. It also needs a full description. At least a paragraph of detail on data, portfolio construction and so on.

Why not also have a Figure 2 which would be identical but use (6, 6)? I think that would be nice. The caption would be identical, but would also need a sentence that highlighted how the overlapping holding periods meant that returns were correlated among portfolios that were formed around the same time.

Table 2 has some weird characters in the caption. Note that you need to be careful about units. I *think* that you want both results given in *per month* units, so that the two columns are directly comparable. But, right now, the 1, 1 give 1.4% per month, while the 6, 6 gives 12% per six months. Instead, you should divide all those numbers by 6, so that 6, 6 reports 2% per month. The facts are the same, but this saves the reader from doing the math in his head. He wants to compare 1.4 to 2.0 directly. The caption should make it clear that you have done this. Same applies for Table 3.

You should also note in the main text (but not in the caption) how these results compare to JT and GH. That is of great interest to the reader. You report that 1,1 for individual stocks produces a zero spread. Does JT report that? I think so (and it is due to one month reversal, as documented by Jagedeesh 1993, which you should cite). How about the 1% *per month* for 6, 6 on individual stocks? Is that about the same as JT? I think so. If so, tell us that!

Table 3 also looks good. And so don't forget to explain every little detail in the text. I should be able to read the text (while ignoring the tables) and still get everything of importance. In particular, how similar are your results to MG? We don't really care about the levels for winner and loser (that is impacted by the overall market over this time). We care about all 6 spread numbers. How similar are these to MG? Discuss each one. That will take two paragraphs, but it is the heart of your replication.

Last thoughts.

1) To the extent you (or Vincent/Bill) want till Wednesday morning, I would be happy to review and provide feedback on another draft.

2) "This process is only theoretical . . . " is a pretty horrible sentence. Read it aloud. For *and* against?

3) typo "to when"

4) I like the whole discussion of larger industries == more efficient market.

5) Again, the first paragraph of the conclusion is good. See my previous comments. The second paragraph should be cut. Although you should *mention* the data (large cap, 1998 --- 2007, US), the conclusion is much shorter than the introduction (and longer than the abstract). Moreover, it is much less focused on your data and much more on your results and your extensions.

Otherwise, good draft. I think that you have made a huge amount of progress and I look forward to seeing what comes next.