

Working with the TAQ data

Joel Hasbrouck

March 26, 2004

1. Introduction

The TAQ (Trade and Quote) database is the primary source of historical trade and quote data for US equities. It is most conveniently available to academic subscribers via the Wharton Research Data System (WRDS). Using the WRDS web interface, you can download subsets of the data as Excel spreadsheets.

This document will walk you through the process of building an Excel spreadsheet with trades and quotes. As an example, I'll use Consolidated Graphics, Inc. (ticker CGX) on March 13 and 14, 2003. The final spreadsheet is [CGX.xls](#).

TAQ has two main parts: CT and CQ.

a. Trades

The consolidated trade (CT) database covers transactions. It reports time, price, volume and reporting exchange for all trades. It also notes error conditions, special settlement terms, etc.

b. Quotes

The consolidated quote (CQ) database covers quotes: time, bid, ask and quoting exchange. It also notes special quote modes (like “trading halted”, “fast markets, etc.)

2. WRDS access procedures

As an example of how to extract the data into Excel spreadsheets, we'll take Consolidated Graphics, Inc. (ticker CGX) on March 13 and 14, 2003. The WRDS web page is <http://wrds.wharton.upenn.edu/>. Follow the link to “members login” and use username “tmf2001”. Your password will be given to you separately.

a. The first screen after the login lists the available databases:

Click on “NYSE TAQ”, near the bottom of the menu

b. The next screen has the menu that allows the user to select “Consolidated Trades” or “Consolidated Quotes”:

We'll do trades first.

c. The interface screen for consolidated trades looks like this:

[de and Quote Database >](#)

TAQ - Consolidated Trades

Data Query | [Documentation](#) | [Data Manuals](#)

Step One: Date Range

Beginning Mar 13 2003

Ending Mar 14 2003

* Data is available from January 1993 through November 2003.

Step Two: Time Range

Filter observations by timestamp

Beginning 09 : 30 : 00

Ending 16 : 00 : 00

* Default to observations within normal trading hours

Step Three: Search

Search By SYMBOL

Choose 1 Of 3 Methods

1. Company Codes

[Code Lookup](#)

2. File Containing Company Codes

3. Entire Database

At step one, we fill in the dates; we ignore step two; at step three, we put in the ticker symbol CGX.

d. Next, we scroll down to fill in the variables:

Filter observations by timestamp
 Beginning : :
 Ending : :
* Default to observations within normal trading hours

Step Three: Search

Search By

Choose 1 Of 3 Methods

1. Company Codes
 [Code Lookup](#)

2. File Containing Company Codes

3. Entire Database

Step Four: Variables

Actual trade Price per Share Combined "G" Rule 127 & Stopped Stock Trade Indicator
 Correction Indicator Sale Condition Exchange on which the Quote Occured
 Number of Shares Traded

Step Five: Output

Output [Format](#)

Compression Type

E-Mail Address (Optional)

At step four, click on all variables except the “G” flag. At step five, specify “comma-delimited text”. (This is easiest for Excel to import.) Then press “submit request”.

e. The system will respond with a screen that looks like this:

The screenshot shows a web browser window with the following content:

Data Request Summary

Data Request ID	221500035
Libraries/Data Sets	taq/ct /
Frequency/Date Range	intraday / 13Mar2003 - 14Mar2003
Search Variable	SYMBOL
Input Codes 1 item(s)	CGX
Conditional Statements	n/a
Output format/Compression	csv /
Variables Selected	PRICE CORR COND EX SIZE
Extra Variables and Parameters Selected	

Your request is being processed. When finished, the output will be found at <http://wrds.wharton.upenn.edu/output/221500035.html?>

This page will refresh every 10 seconds until the output appears.

If the output is not displayed...

- Check your web browser preferences to ensure that cached data is compared to the network **every time**.
- Send e-mail to WRDS Technical Support at wrds-support@wharton.upenn.edu.

Please note that the output will remain on the system for 48 hours.

f. After a few moments (on minutes, depending on how busy the system is), you'll see a screen:

Data Request Summary

Data Request ID	221500035
Libraries/Data Sets	taq/ct /
Frequency/Date Range	intraday / 13Mar2003 - 14Mar2003
Search Variable	SYMBOL
Input Codes 1 item(s)	CGX
Conditional Statements	n/a
Output format/Compression	csv /
Variables Selected	PRICE CORR COND EX SIZE
Extra Variables and Parameters Selected	

Your output is complete. Click on the link below to open the output file.

[221500035.csv](#) (39 KB)

Download instructions
 Netscape users... Shift-click
 Internet Explorer users... Right-click and select "Save Target As..."

Follow the download instructions.

g. Opening the file in Excel should give you something like this:

	A	B	C	D	E	F	G	H	I	J
1	symbol	date	time	price	size	corr	cond	ex		
2	CGX	13-Mar-03	9:32:55	22.81	100	0		N		
3	CGX	13-Mar-03	9:38:30	22.83	300	0		N		
4	CGX	13-Mar-03	9:48:02	22.82	100	0		N		
5	CGX	13-Mar-03	9:50:33	22.83	100	0		N		
6	CGX	13-Mar-03	9:50:49	22.83	100	0		N		
7	CGX	13-Mar-03	9:58:44	22.83	100	0		N		
8	CGX	13-Mar-03	10:00:14	22.82	400	0	E	N		
9	CGX	13-Mar-03	10:05:28	22.83	300	0		N		
10	CGX	13-Mar-03	10:11:22	22.82	100	0		N		
11	CGX	13-Mar-03	10:36:38	22.82	100	0		N		

I'll discuss the data in more detail below. But for the moment, let's return to the TAQ menu and go to "Consolidated Quotes"

h. The consolidated quote menu looks like this:

TAQ
//NYSE Trade and Quote Database >

[Consolidated Trades](#)
[Consolidated Quotes](#)
[NBBO](#)
[NBBO+Trades](#)

TAQ - Consolidated Quotes

Data Query | [Documentation](#) | [Data Manuals](#)

Step One: Date Range

Beginning Mar 13 2003
 Ending Mar 14 2003
 * Data is available from January 1993 through November 2003.

Step Two: Time Range

Filter observations by timestamp

Beginning 09 : 30 : 00
 Ending 16 : 00 : 00
 * Default to observations within normal trading hours

Step Three: Search

Search By SYMBOL

Choose 1 Of 3 Methods

1. Company Codes
 CGX [Code Lookup](#)

2. File Containing Company Codes
 [Browse...](#)

3. Entire Database

Step Four: Variables

Bid Price Offer Price Bid Size in Number of Round Lots Offer Size in Number
 Quote Condition Exchange on which the Quote Occurred
 Nasdaq market marker for each NASD Quote

As with the trade menu, fill in the date range, ignore the time range, and fill in the ticker symbol CGX. Then scroll down.

i. The bottom of the screen is:

The screenshot shows a web browser window with a toolbar containing 'Search Web', 'Search Site', 'PageRank', and 'Options'. The main content area is divided into several sections:

- Filter observations by timestamp:** Includes 'Beginning' (09:30:00) and 'Ending' (16:00:00) time pickers. A note states: '* Default to observations within normal trading hours'.
- Step Three: Search:** Features a 'Search By' dropdown set to 'SYMBOL'. Below it, 'Choose 1 Of 3 Methods' are listed:
 - 1. Company Codes:** Selected. Includes a text input with 'CGX' and a 'Code Lookup' link.
 - 2. File Containing Company Codes:** Includes a text input and a 'Browse...' button.
 - 3. Entire Database:** Unselected.
- Step Four: Variables:** A list of checkboxes, all of which are checked:
 - Bid Price
 - Offer Price
 - Bid Size in Number of Round Lots
 - Offer Size in Number of Round Lots
 - Quote Condition
 - Exchange on which the Quote Occured
 - Nasdaq market marker for each NASD Quote
- Step Five: Output:** Includes an 'Output Format' dropdown set to 'comma-delimited text (*.csv)', a 'Compression Type' dropdown set to '<none>', and an 'E-Mail Address (Optional)' text input field.

At the bottom of the form are two buttons: 'Submit Request' and 'Reset'.

At step 4, select all variables (you have to click on each one individually). At step five, set the output to comma-delimited text; submit the request and wait for the response.

j. After a few minutes:

Data Request Summary - Microsoft Internet Explorer provided by Verizon Onl

File Edit View Favorites Tools Help

Address <http://wrds.wharton.upenn.edu/output/221500609.html?>

Google e close manipulation Search Web Search Site PageRank

Data Request Summary

Data Request ID	221500609
Libraries/Data Sets	taq/cq /
Frequency/Date Range	intraday / 13Mar2003 - 14Mar2003
Search Variable	SYMBOL
Input Codes 1 item(s)	CGX
Conditional Statements	n/a
Output format/Compression	csv /
Variables Selected	BID OFR BIDSIZ OFRSIZ MODE EX MMID
Extra Variables and Parameters Selected	

Your request is being processed. When finished, the output will be found at <http://wrds.wharton.upenn.edu/output/221500609.html?>

This page will refresh every 10 seconds until the output appears.

If the output is not displayed...

- Check your web browser preferences to ensure that cached data is compared to the network **every time**.
- Send e-mail to WRDS Technical Support at wrds-support@wharton.upenn.edu.

Please note that the output will remain on the system for 48 hours.

Follow the download instructions.

k. When the file is read into Excel, it looks like this:

Microsoft Excel - 221500609.csv

File Edit View Insert Format Tools Data Window Help Adobe PDF

100% Arial 10

	A	B	C	D	E	F	G	H	I	J
	symbol	date	time	bid	ofr	bidsiz	ofrsiz	mode	ex	mmid
2	CGX	13-Mar-03	8:00:08	0	0	0	0	12	P	
3	CGX	13-Mar-03	8:04:39	0.01	0	1	0	12	P	
4	CGX	13-Mar-03	8:18:26	0.01	45.6	1	1	12	P	
5	CGX	13-Mar-03	8:30:07	0	0	0	0	12	T	TRIM
6	CGX	13-Mar-03	8:30:07	0	0	0	0	12	T	BRUT
7	CGX	13-Mar-03	9:32:56	0.01	91.2	1	1	12	T	BRUT
8	CGX	13-Mar-03	9:32:56	0.01	91.2	1	1	12	T	CAES
9	CGX	13-Mar-03	9:33:01	22.7	22.9	2	2	10	N	
10	CGX	13-Mar-03	9:33:02	21.7	23.9	1	1	12	M	
11	CGX	13-Mar-03	9:33:03	22.1	23.1	1	1	12	X	
12	CGX	13-Mar-03	9:33:03	22.11	23	10	3	12	T	TRIM
13	CGX	13-Mar-03	9:33:03	22.11	23	10	3	12	T	CAES
14	CGX	13-Mar-03	9:33:07	22.7	22.89	2	2	12	N	
15	CGX	13-Mar-03	9:33:08	22.1	23.09	1	1	12	X	
16	CGX	13-Mar-03	9:33:08	21.7	23.89	1	1	12	M	
17	CGX	13-Mar-03	9:33:08	22.71	22.89	2	2	12	N	

3. Sources

TAQ is a transcript of the real-time data flowing from the Consolidated Quote System (CQS) and the Consolidated Trade System (CTS). These are systems run by the Consolidated Tape Association (CTA), an industry consortium that, for the present at least, functions as the primary aggregator of trade and quote data from diverse market centers. The quotes you'd be getting from an E-broker or (delayed) from other web source originate from CTS/CQS.

The primary documentation is the TAQ user guide, available at <http://wrds.wharton.upenn.edu/ds/taq/manuals/>. A copy of the manual is also on my web site ([click here](#)). TAQ is marketed by the NYSE for commercial and academic use.

4. The Consolidated Quote (CQ) file

a. General remarks

- ▶ The CQ file covers most activity in major US market centers. It does not cover non-US market centers, nor does it cover activity between 6pm and 7am (?).
- ▶ A record in the CQ file represents a quote update originating in one of the market centers.
- ▶ The record generally establishes the best bid and offer (BBO) prevailing in the market center.
- ▶ Normally, a quote from a market center is regarded as firm and valid until superceded by a new quote from the center.
- ▶ An exception occurs when the market center wishes to indicate that it is withdrawing or canceling its quote and is not posting another quote.
- ▶ At present this is indicated by posting a bid of one penny and an offer that is approximately twice the previous day's close.

- ▶ A CQ record does not establish a comprehensive market-wide “national” best bid and offer (NBBO).
- ▶ If we want to know what the NBBO was at a particular instant, we need to look back and determine the most recent bids and offers posted by all market centers. Then we take the highest bid and the lowest offer.
- ▶ The CQ file does not necessarily indicate the most timely data available to market participants.

For example, when the best bid on the Island ECN changes, Island’s subscribers are notified by an electronic link directly from Island. Island also sends the update to the CTA, which then disseminates it further. This latter route is more circuitous and presumably slower.

b. Discussion of CGX on March 13, 2003 (a “normal” day)

Here are the first records in the CGX data:

symbol	date	time	bid	ofr	bidsiz	ofrsiz	mode	ex	mmid
CGX	13-Mar-03	8:00:08	0.00	0.00	0	0	12	P	
CGX	13-Mar-03	8:04:39	0.01	0.00	1	0	12	P	
CGX	13-Mar-03	8:18:26	0.01	45.60	1	1	12	P	
CGX	13-Mar-03	8:30:07	0.00	0.00	0	0	12	T	TRIM
CGX	13-Mar-03	8:30:07	0.00	0.00	0	0	12	T	BRUT
CGX	13-Mar-03	9:32:56	0.01	91.20	1	1	12	T	BRUT
CGX	13-Mar-03	9:32:56	0.01	91.20	1	1	12	T	CAES
CGX	13-Mar-03	9:33:01	22.70	22.90	2	2	10	N	
CGX	13-Mar-03	9:33:02	21.70	23.90	1	1	12	M	
CGX	13-Mar-03	9:33:03	22.10	23.10	1	1	12	X	
CGX	13-Mar-03	9:33:03	22.11	23.00	10	3	12	T	TRIM
CGX	13-Mar-03	9:33:03	22.11	23.00	10	3	12	T	CAES
CGX	13-Mar-03	9:33:07	22.70	22.89	2	2	12	N	
CGX	13-Mar-03	9:33:08	22.10	23.09	1	1	12	X	
CGX	13-Mar-03	9:33:08	21.70	23.89	1	1	12	M	
CGX	13-Mar-03	9:33:23	22.71	22.89	2	2	12	N	
CGX	13-Mar-03	9:33:24	22.11	23.09	1	1	12	X	
CGX	13-Mar-03	9:33:24	21.71	23.89	1	1	12	M	

Here are some things to note:

- ▶ Although CGX is an “NYSE-listed” stock, it trades on other exchanges.

- ▶ “P” refers to the Pacific Exchange. The Pacific Exchange used to be a floor-based market, but it merged with Archipelago (and ECN). The “P” quotes are the best bid and offer from Archipelago’s book. When an ECN has nothing on the bid (buy) side of its book, it posts a bid of \$0.01; when there’s nothing on the ask (sell) side, it posts a an offer of roughly twice the current stock price.
- ▶ When a Nasdaq market-maker ID (MMID) appears, it refers to a Nasdaq dealer (TRIM= “Trimark”) or ECN (“BRUT”).
- ▶ CAES is the acronym for “Computer Assisted Execution System”. This is a Nasdaq system that allows its members to quote NYSE-listed stocks. The CAES quote generally represents the best Nasdaq quotes in the issue.
- ▶ Exchange symbol “X” refers to Philadelphia; “M” refers to the Chicago (formerly the Midwest) Stock Exchange. See the TAQ documentation for a full list of the exchange codes.
- ▶ The bid and offer are in dollars per share. “Bidsiz” and “Ofrsiz” are the amounts at the bid and ask (in 100-share round lots). For example, at 9:33:03, Trimark is bidding \$22.11 for 1,000 shares, with 300 shares offered at \$23.00.
- ▶ A “normal” quote has mode 12; The NYSE uses mode 10 to denote its opening quotes. See the TAQ documentation for other modes.

c. Discussion of CGX on March 14, 2003: An unusual day

Before the opening on March 14, there had apparently been a negative news announcement. Here are the initial quote records:

time	bid	ofr	bidsiz	ofrsiz	mode	ex	mmid
8:00:13	0.00	22.61	0	8	12	P	
8:17:03	0.00	22.60	0	10	12	P	
8:27:59	0.00	22.00	0	5	12	P	
8:29:25	0.00	21.99	0	10	12	P	
8:30:07	0.00	0.00	0	0	12	T	TRIM
8:30:07	0.00	0.00	0	0	12	T	BRUT
8:50:25	0.01	21.99	1	10	12	P	
9:04:38	0.01	21.00	1	10	12	P	
9:08:13	0.01	20.99	1	10	12	P	
9:20:26	0.01	22.00	1	5	12	P	
9:20:37	0.01	21.99	1	10	12	P	
9:29:35	0.01	22.61	1	13	12	P	
9:32:48	0.00	0.00	0	0	7	N	
9:32:50	0.00	0.00	0	0	7	X	
9:33:07	20.00	22.00	0	0	7	N	
9:35:01	0.01	22.50	1	10	12	P	
9:38:19	0.01	22.49	1	10	12	P	
9:44:20	18.00	20.00	0	0	7	N	
9:46:22	0.01	22.61	1	13	12	P	
9:50:43	0.01	22.61	1	14	12	P	
9:51:05	17.00	19.00	0	0	7	N	
9:51:21	0.01	22.61	1	24	12	P	
9:51:27	0.01	22.61	1	14	12	P	
9:51:30	0.01	22.59	1	10	12	P	
9:56:58	0.01	22.00	1	20	12	P	
10:00:41	0.01	22.59	1	10	12	P	
10:15:24	15.00	18.00	0	0	7	N	
10:17:27	15.01	22.59	2	10	12	P	
10:29:08	0.00	0.00	0	0	29	N	
10:29:18	15.01	19.87	2	10	12	P	
10:29:24	16.50	16.51	2	2	10	N	

- ▶ Prior to 9:30, the state of the market is “ask without”: the Pacific is posting offers close to the previous day’s close, but there is nothing on the bid.
- ▶ At 9:32:48, New York posts a zero quote with mode “7”. This is an “order imbalance” trading halt. It is issued when the specialist has orders piled up on one side of the market (in this case, “sell”). A full list of quote modes is given in the TAQ documentation..
- ▶ Over the next hour, New York issues successive mode “7” quotes. The bid and ask generally indicate the range where the specialist thinks CGX will open. These quotes are, in a way, invitations for buyers to submit orders.

- ▶ Finally, at 10:29:24, the stock opens. The opening bid and offer are down about five dollars from the previous day's close.

5. The Consolidated Trade (CT) file

a. General remarks

All US equity trades must be reported to the Consolidated Tape Association as soon as possible, but in any event, no later than 90 seconds after the trade has happened.

b. The trades of CGX on March 13, 2003 (a normal day)

Here is the top of the Excel spreadsheet:

symbol	date	time	price	size	corr	cond ex
CGX	13-Mar-03	9:32:55	22.81	100	0	N
CGX	13-Mar-03	9:38:30	22.83	300	0	N
CGX	13-Mar-03	9:48:02	22.82	100	0	N
CGX	13-Mar-03	9:50:33	22.83	100	0	N
CGX	13-Mar-03	9:50:49	22.83	100	0	N
CGX	13-Mar-03	9:58:44	22.83	100	0	N
CGX	13-Mar-03	10:00:14	22.82	400	0 E	N
CGX	13-Mar-03	10:05:28	22.83	300	0	N
CGX	13-Mar-03	10:11:22	22.82	100	0	N
CGX	13-Mar-03	10:36:38	22.82	100	0	N
CGX	13-Mar-03	10:40:22	22.81	300	0	N
CGX	13-Mar-03	10:40:23	22.81	600	0	N
CGX	13-Mar-03	10:41:39	22.79	200	0	N

This is fairly straightforward. The only complication arises in the “cond” field. The TAQ documentation does not specify the meaning of “E”. (As a precaution, ignore any trades for which the correction field *corr* is nonzero.)

c. The trades of CGX on March 14, 2003 (an unusual day)

symbol	date	time	price	size	corr	cond	ex
CGX	14-Mar-03	9:37:57	22.61	2000	0		T
CGX	14-Mar-03	9:48:50	22.61	1000	0		T
CGX	14-Mar-03	9:56:02	22.61	5000	0		T
CGX	14-Mar-03	10:29:16	16.50	46800	0	O	N
CGX	14-Mar-03	10:29:21	16.50	1200	0		N
CGX	14-Mar-03	10:29:31	16.51	200	0	E	N
CGX	14-Mar-03	10:29:35	16.51	500	0	Z	T
CGX	14-Mar-03	10:29:38	16.51	400	0		T
CGX	14-Mar-03	10:29:47	16.51	100	0		T
CGX	14-Mar-03	10:30:47	16.53	1000	0		T
CGX	14-Mar-03	10:30:54	16.50	500	0	E	N
CGX	14-Mar-03	10:30:56	16.51	200	0	E	N
CGX	14-Mar-03	10:30:59	16.50	1000	0	E	N
CGX	14-Mar-03	10:31:00	16.50	1000	0	E	N
CGX	14-Mar-03	10:31:00	16.50	400	0	E	N
CGX	14-Mar-03	10:31:01	16.51	500	0		T
CGX	14-Mar-03	10:31:01	16.50	300	0	E	N
CGX	14-Mar-03	10:31:03	16.50	1000	0	E	N
CGX	14-Mar-03	10:31:05	16.50	300	0	E	N
CGX	14-Mar-03	10:31:07	16.50	5500	0		N

- ▶ First note that there are trades prior to the NYSE open on Nasdaq, indicated by exchange symbol T. (And how would you feel if you'd bought 5,000 shares at \$22.61?)
- ▶ The NYSE open is marked with an "O" condition. The 46,800 size is an aggregate amount, representing all buyers and sellers.

6. Some things to look at

a. Computing the spread

The spread is simply the offer price (ask) less the bid. I've made a copy of the CQ worksheet named "Quotes by exchange", and computed the spread in column K.

b. Computing the bid-ask spread, by exchange.

Next, I sort the data by exchange, and within exchange, by (descending spread). The purpose here is to group all of the quotes for a particular exchange, and within each

exchange, to isolate the ones that are either zero or extremely large. These are meaningless quotes. You can then plot the spreads, compute average spreads, etc.

c. The NBBO

The national best bid and offer are the highest bid and lowest offer in the market at a point in time. Within Excel it is difficult to construct the NBBO in a programmatic way, because it involves keeping track, over time, of what each venue is quoting.

d. Signing the trades (advanced)

When a trade occurs on the bid price, it's reasonable to infer that the initiator of the trade (the active side of the trade) was a buyer. When a trade occurs on the offer, we infer that the active side was a seller. Sometimes it's useful or interesting to characterize a trade as "buy" or "sell".

To do this, it's cleanest to work with NYSE quotes only, as these tend to be the most representative and error-free. I've copied the NYSE quotes to a worksheet labeled "NYSE quotes" and sorted them by date and time. I've also made a copy of the CT data in a worksheet labeled "Signed trades".

For a given trade, we need to know what were the prevailing bid and offer at the time of the trade. We can determine this using Excel's VLOOKUP function, applied to the "NYSE quotes" data. In columns I and J of "Signed trades", I've put in the lookup code. In column K, I put the difference between the transaction price and the quote midpoint (the average of the bid and ask).

If the column-K difference is negative, column L reads "Sell"; if the difference is positive, column M reads "Buy".