



# **Merging CRSP and Compustat**

Luis Palacios

# **Agenda**

- 1. Comparing CRSP and Compustat Universe.**
- 2. Merging using CUSIP**
  - a. Understanding CUSIP codes.**
  - b. Preparing data.**
  - c. Merging example.**
- 3. CRSP-Compustat Merge (CCM) product**
  - a. Understanding CCM linking file**
  - b. Merging example.**
- 4. CCM web query**

# 1. Comparing CRSP and Compustat

- CRSP covers stock market data on major stock exchanges (NYSE, AMEX, NASDAQ).
- CRSP main identifiers: PERMCO and PERMNO
- CRSP can have multiple securities for each firm.
- CRSP includes data after IPO.
- Data since 1925.
- Compustat covers accounting data for public & private companies
- Compustat Identifier: GVKEY
- Special Compustat records: Red-Herring Companies.
- In general, Compustat requires that a firm has some number of years of history as public company before including it in the dataset.
- Data since 1950.

## 2. Merging using CUSIPs

### 2a. Which CRSP and Compustat files use?

#### CRSP (monthly) files:

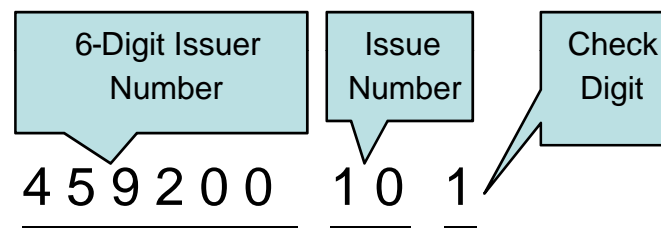
**STOCKNAMES:** PERMCO, PERMNO, **CUSIP**, ...security identifiers...(no time series)

#### Compustat (annual) files:

**NAMESANN:** GVKEY, **CNUM** (First 6-digit of Cusip), **CIC** (7<sup>th</sup> to 9<sup>th</sup> CUSIP digits), Company name (header), ... firm identifiers (no time series) .

## 2a. ... Understanding CUSIPs ...

Example using CUSIP for IBM:



**Compustat :**

CNUM=459200, CIC=101

Header Variables

**CRSP:**

CUSIP=45920010

Header Variable.

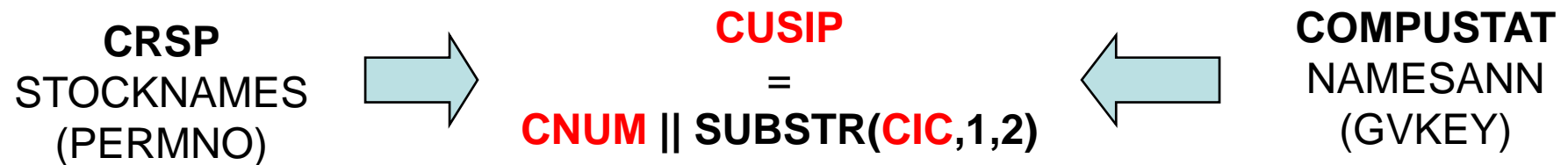
NCUSIP=45920010

Historical Variable.

---

## 2a. ... Two merging methods using CUSIPs.

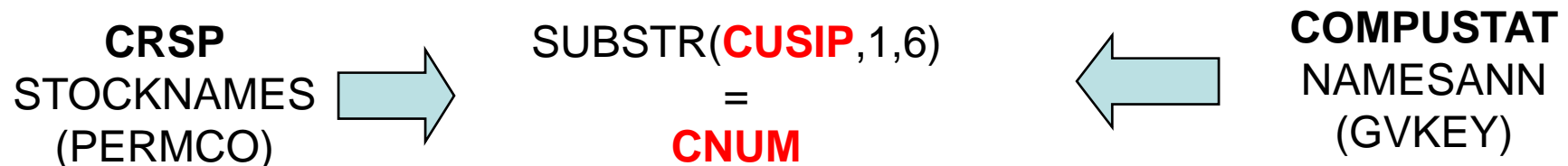
At security level:



PERMNO-GVKEY match. It matches with the security that has price in Compustat.

---

At firm level:



PERMCO-GVKEY match. It still needs to come up with the 'main' PERMNO for each PERMCO.

## 2b. Preparing Data

### ➤ From Compustat's NAMESANN file

```
proc sort data= comp.namesann (keep=gvkey cnum) out=comp1 nodupkey;  
by cnum;  
run;
```

- Comp1 is a file with unique gvkey cnum;

### ➤ From CRSP's STOCKNAMES file

```
data crsp1;  
set crsp.msfnames (keep = cusip permco);  
cnum = substr(cusip,1,6);  
run;
```

```
proc sort data=crsp1 nodupkey;  
by cnum;  
run;
```

\* File crsp1 contains unique cusip permco;

## 2c. Full Join

```
proc sql;  
  
create table ccm2  
as select *  
from comp1 as a full join crsp1 as b  
on a.cnum=b.cnum;  
  
quit;  
  
proc print data=ccm2 (obs=10);  
var cusip6 gvkey permco;  
run;
```



## 2c. Merging Example

CUSIP6	GVKEY	PERMCO
000021	23052	29189
000032	1000	23369
000165	1001	6398
000209	13592	9160
000352	1002	22159
000354	1003	6672
00035P	128178	.
000360	21542	10817
000361	1004	20000
000370	1005	11
000375	210418	41444
000400	30222	.
000573	9570	7252
000722	11730	.
000736	1006	.
000742	.	13
000752	29395	12673
000770	21435	.
000771	1412	7360
000774	1007	20

Not in CRSP

Not in CRSP

Not in CRSP

Not in CRSP

Not in Compustat

Not in CRSP

## **2c. How good is merging with CUSIPs?**

- From a set of 25,315 GVKEYs (comp.namesann file), **81** percent have an equivalent PERMCO in CRSP.
- From a set of 25,611 PERMCOs (comp.stocknames file), **80** percent have an equivalent GVKEY in Compustat.

## **What to do with the unmatched cases?**

Matching by company names, tickers, SPEDIS function,  
FORUMS@WRDS

# **3. CRSP-Compustat Merge (CCM)**

**3a. Understanding CCM linking file**

**3b. Preparing Compustat Data for Merging.**

**3c. Merging Example.**

## 3a. Understanding CCM linking file

The linking file is **CRSP.CSTLINK2**

These are its variables:

GVKEY	=	Compustat's Identifier
NPERMNO	=	CRSP PERMNO
NPERMCO	=	CRSP PERMCO
LINKTYPE	=	Link Type Code
LINKFLAG	=	Link Flag
LINKDT	=	First Effective Date of Link
LINKENDDT	=	Last Effective Date of Link
USEDFLAG	=	(1=single match)

### 3.a. ... What is the LINKTYPE Code ?

linktype Code	Description
LU	Link is specified, no additional detail is provided.
LC	Standard link available to COMPUSTAT company data.
LD	A link is provided to an issue or a company. However, the link is a duplicate. A CRSP link is already available between other COMPUSTAT gvkeys and the CRSP data. Consolidated COMPUSTAT records are typically assigned a Link Type Code of LD. Care must be taken to avoid double-counting if using one of these links.
LF	The COMPUSTAT gvkey consists of pre-FASB reporting. A link is provided to an issue in CRSP. However, a CRSP link is already provided to this issue from a different gvkey with standard reporting. Care must be taken to avoid double-counting if using one of these links.
LN	Has a direct link through CUSIP, but COMPUSTAT record has no prices.
LO	Links to trading company, but no linking trading issues exists.
LS	The COMPUSTAT record directly links to a CRSP issue. The company may have other issues, but data for those issues links directly to other COMPUSTAT gvkeys. Care must be taken to avoid double-counting data if using PERMCO to link.
LX	Links to foreign company with an issue trading on US exchanges.
NE	No link. All prices refer to minor or non-US exchanges, or predate CRSP coverage of AMEX or Nasdaq data., or are outside the CRSP coverage of major exchanges (e.g. foreign issues traded on Nasdaq.)
NU	No link. Further information not yet available.
NP	No link. No COMPUSTAT prices and no indicator of major exchanges.
NR	No link. Research completed (Closed).
NX	No link found. CRSP record exists, but no COMPUSTAT prices are found on a US exchange. Research pending.

LINKTYPES = {NE,NU,NP,NR,NX} mean NO link !

### 3.a. ... What is the LINKFLAG Code?

- A 3-character flag.
- The first position refers to the last CRSP annual release.
- The second position refers to the last CRSP monthly release,
- The third position refers to the last CRSP quarterly release.

linkflag Code	Description
B	PERMNO Resides on both CRSP Monthly and Daily Databases
D	PERMNO Resides on a CRSP Daily Database
M	PERMNO Resides on a CRSP Monthly Database
X	PERMNO Does Not Reside on a CRSP Subscriber Database

Example:

Linkflag = “XXX” means that this PERMNO is not in the daily neither the monthly CRSP database.

### 3.a. ... LINKDT and LINKENDDT code

LINKDT:

-Date marking the first effective date of the link.

LINKENDDT:

- Date marking the last date where the link is valid. If the link is still, LINKENDDT is set to **.E**

### 3.a. ... Example of the “cstlink2” file

GVKEY	NPERMNO	NPERMCO	LINKTYPE	LINKFLAG	LINKDT	LINKENDDT	USEDFLAG
1076	10517	5674	LU	BBB	19810401	19921231	1
1076	78049	5674	LU	BBB	19930101	E	1
1241	.	.	NU	XXX	19800101	19840731	0
1241	11608	7161	LU	BBB	19840801	19890331	1
6066	12490	20990	LC	BBB	19500101	E	1
145327	.	.	NR	XXX	19990701	E	0
1415	54332	20070	LD	BBB	19840101	19840531	0
1416	54332	20070	LC	BBB	19840101	19911209	1



### 3.a. .. How good is merging with CCM?

- From a set of 25,315 GVKEYs (comp.namesann file), **87** percent have an equivalent PERMCO in CRSP.
- From a set of 25,611 PERMCOs (comp.stocknames file), **84** percent have an equivalent GVKEY in Compustat.

## 3b. Preparing Compustat Data

- Because this Compustat file is shown in “Fiscal years”, we will need first to create a “Calendar” dates for merging with CRSP.
- Specifically, we create the first and last day of each fiscal year.
- In this example, we will merge Compustat and CRSP data for only three firms: Dell, IBM and Microsoft.
- We will use the annual Compustat data (the comp.compann file).

## 3.b. Preparing Compustat data for merging

```
data comp22;
set comp.compann (keep = gvkey smb1 yeara fyr data6 data181
    data25 data199);
where smb1 in ("IBM", "DELL", "MSFT");
* calendar year calculation;
if fyr>0;
if fyr <= 5 then cyear = yeara + 1;
else cyear=yeara;
* begin and end dates for fiscal year;
date1= MDY(fyr,1,cyear);
endfyr= intnx('month',date1,0,'end'); format endfyr date9.;
begfyr= intnx('month',endfyr, -11,'beg');format begfyr
    date9.;
* create a book-to-market ratio;
btm = (Data6-Data181)/(data25*data199);
if btm >0;
* keep only relevant variables;
keep gvkey smb1 begfyr endfyr btm;
run;
```

### 3.b. File with Compustat data before merging (subset for DELL data)

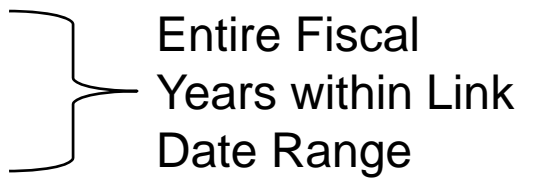
<b>SMBL</b>	<b>GVKEY</b>	<b>BEGFYR</b>	<b>ENDFYR</b>	<b>BTM</b>
<b>DELL</b>	14489	1-Feb-88	31-Jan-89	0.42
<b>DELL</b>	14489	1-Feb-89	31-Jan-90	0.92
<b>DELL</b>	14489	1-Feb-90	31-Jan-91	0.26
<b>DELL</b>	14489	1-Feb-91	31-Jan-92	0.36
<b>DELL</b>	14489	1-Feb-92	31-Jan-93	0.22
<b>DELL</b>	14489	1-Feb-93	31-Jan-94	0.56
<b>DELL</b>	14489	1-Feb-94	31-Jan-95	0.39
<b>DELL</b>	14489	1-Feb-95	31-Jan-96	0.38
<b>DELL</b>	14489	1-Feb-96	31-Jan-97	0.09
<b>DELL</b>	14489	1-Feb-97	31-Jan-98	0.04
<b>DELL</b>	14489	1-Feb-98	31-Jan-99	0.02
<b>DELL</b>	14489	1-Feb-99	31-Jan-00	0.05
<b>DELL</b>	14489	1-Feb-00	31-Jan-01	0.08
<b>DELL</b>	14489	1-Feb-01	31-Jan-02	0.07
<b>DELL</b>	14489	1-Feb-02	31-Jan-03	0.08
<b>DELL</b>	14489	1-Feb-03	31-Jan-04	0.07
<b>DELL</b>	14489	1-Feb-04	31-Jan-05	0.06

## 3c. Merging Example using CCM

- First, merge Compustat data with the linking table (crsp.cstlink2);
- Second, add CRSP data;
- The CRSP data is the monthly data file (crsp.msf).
- I will present two different ways to add CRSP data depending on the dates.

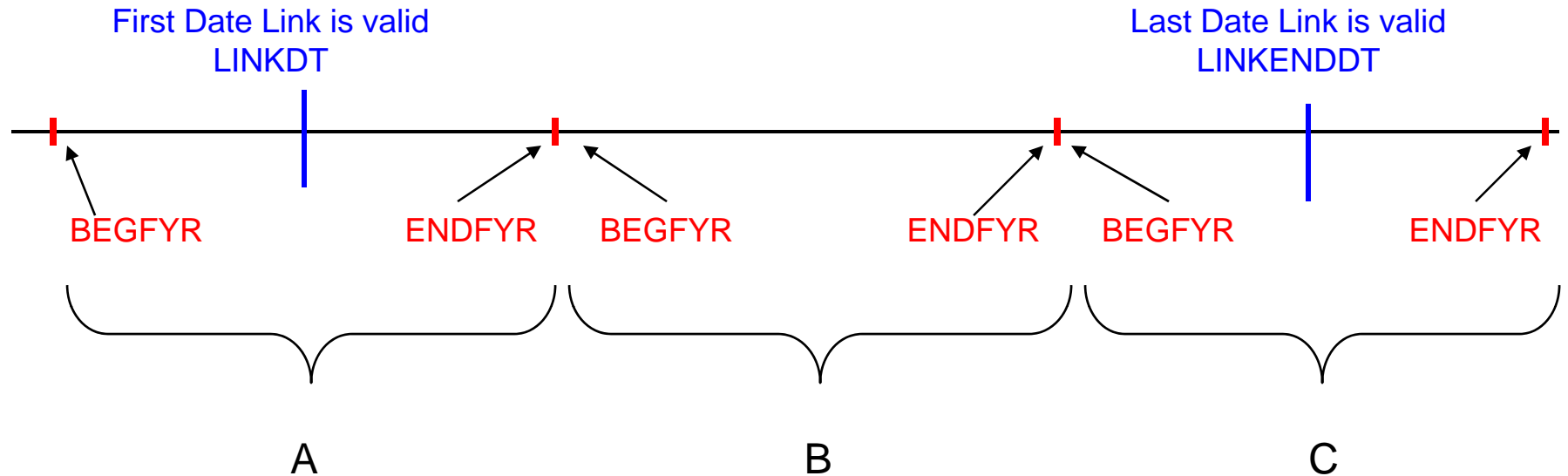
### 3.c. First, Merging Compustat data with CCM linking table

```
proc sql;  
create table mydata  
as select *  
  
from comp22 as a, crsp.cstlink2 as b  
  
where a.gvkey = b.gvkey and  
  
b.LINKTYPE in ( "LU" , "LC" , "LD" , "LF" , "LN" , "LO" , "LS" , "LX" ) and  
  
b.usedflag=1 and  
  
(b.LINKDT<=a.BEGFYR) and  
(a.ENDFYR<=b.LINKENDDT or b.LINKENDDT=.E) ;  
  
quit;
```



Entire Fiscal  
Years within Link  
Date Range

### 3.c. Fiscal Years and Link Date Range



- Entire fiscal period must be within link date range: Observation B
- Fiscal period end date must be within link date range: Observations A and B
- Any part of fiscal period is within link date range: Observations A, B and C


### 3.c. File with Compustat data and CRSP identifiers (subset for DELL data)

SMBL	GVKEY	NPERMNO	NPERMCO	ENDFYR	BTM
DELL	14489	11081	9833	31-Jan-90	0.92
DELL	14489	11081	9833	31-Jan-91	0.26
DELL	14489	11081	9833	31-Jan-92	0.36
DELL	14489	11081	9833	31-Jan-93	0.22
DELL	14489	11081	9833	31-Jan-94	0.56
DELL	14489	11081	9833	31-Jan-95	0.39
DELL	14489	11081	9833	31-Jan-96	0.38
DELL	14489	11081	9833	31-Jan-97	0.09
DELL	14489	11081	9833	31-Jan-98	0.04
DELL	14489	11081	9833	31-Jan-99	0.02
DELL	14489	11081	9833	31-Jan-00	0.05
DELL	14489	11081	9833	31-Jan-01	0.08
DELL	14489	11081	9833	31-Jan-02	0.07
DELL	14489	11081	9833	31-Jan-03	0.08
DELL	14489	11081	9833	31-Jan-04	0.07
DELL	14489	11081	9833	31-Jan-05	0.06



### 3.c. Adding CRSP data and Final dataset

```
proc sql;  
create table mydata2  
as select *  
  
from mydata as a, crsp.msf as b  
  
where a.npermno = b.permno and  
  
month(a.endfyr) = month(b.date) and  
year(a.endfyr) = year(b.date);  
  
quit;
```



Simple case  
when the  
match is at the  
**end** of the  
fiscal year.

### 3.c. Final merged set: Compustat + CRSP data (subset for DELL data)

SMBL	GVKEY	ENDFYR	PERMNO	PERMCO	DATE	PRC	RET
DELL	14489	31-Jan-90	11081	9833	19900131	4.63	-0.16
DELL	14489	31-Jan-91	11081	9833	19910131	22.63	0.22
DELL	14489	31-Jan-92	11081	9833	19920131	31.88	0.24
DELL	14489	31-Jan-93	11081	9833	19930129	46.25	-0.04
DELL	14489	31-Jan-94	11081	9833	19940131	22.00	-0.03
DELL	14489	31-Jan-95	11081	9833	19950131	42.63	0.04
DELL	14489	31-Jan-96	11081	9833	19960131	27.38	-0.21
DELL	14489	31-Jan-97	11081	9833	19970131	66.13	0.24
DELL	14489	31-Jan-98	11081	9833	19980130	99.44	0.18
DELL	14489	31-Jan-99	11081	9833	19990129	100.00	0.37
DELL	14489	31-Jan-00	11081	9833	20000131	38.44	-0.25
DELL	14489	31-Jan-01	11081	9833	20010131	26.13	0.50
DELL	14489	31-Jan-02	11081	9833	20020131	27.49	0.01
DELL	14489	31-Jan-03	11081	9833	20030131	23.86	-0.11
DELL	14489	31-Jan-04	11081	9833	20040130	33.44	-0.02
DELL	14489	31-Jan-05	11081	9833	20050131	41.76	-0.01

### 3.c. Alternative way of Adding CRSP data

```
proc sql;  
create table mydata3 as select *  
  
from mydata as a, crsp.msf as b  
  
where a.npermno = b.permno and  
  
intck('month',a.endfyr,b.date)  
between 3 and 14;  
  
quit;
```

Match Accounting data, with fiscal yearends in month 't', with CRSP data ('returns') of month 't+3' to month 't+14' (12 months).

# Final merged dataset for alternative way of adding CRSP data (subset for DELL data)

SMBL	GVKEY	PERMNO	PERMCO	ENDFYR	DATE	BTM	RET
DELL	14489	11081	9833	31-Jan-90	19900430	0.92	0.11
DELL	14489	11081	9833	31-Jan-90	19900531	0.92	0.29
DELL	14489	11081	9833	31-Jan-90	19900629	0.92	0.15
DELL	14489	11081	9833	31-Jan-90	19900731	0.92	-0.07
DELL	14489	11081	9833	31-Jan-90	19900831	0.92	0.00
DELL	14489	11081	9833	31-Jan-90	19900928	0.92	-0.26
DELL	14489	11081	9833	31-Jan-90	19901031	0.92	0.21
DELL	14489	11081	9833	31-Jan-90	19901130	0.92	0.24
DELL	14489	11081	9833	31-Jan-90	19901231	0.92	0.41
DELL	14489	11081	9833	31-Jan-90	19910131	0.92	0.22
DELL	14489	11081	9833	31-Jan-90	19910228	0.92	0.12
DELL	14489	11081	9833	31-Jan-90	19910328	0.92	0.13
DELL	14489	11081	9833	31-Jan-91	19910430	0.26	-0.18
DELL	14489	11081	9833	31-Jan-91	19910531	0.26	0.06
DELL	14489	11081	9833	31-Jan-91	19910628	0.26	-0.01
DELL	14489	11081	9833	31-Jan-91	19910731	0.26	0.17
DELL	14489	11081	9833	31-Jan-91	19910830	0.26	0.13
DELL	14489	11081	9833	31-Jan-91	19910930	0.26	0.02
DELL	14489	11081	9833	31-Jan-91	19911031	0.26	-0.25
DELL	14489	11081	9833	31-Jan-91	19911129	0.26	-0.06
DELL	14489	11081	9833	31-Jan-91	19911231	0.26	0.09
DELL	14489	11081	9833	31-Jan-91	19920131	0.26	0.24
DELL	14489	11081	9833	31-Jan-91	19920228	0.26	0.09
DELL	14489	11081	9833	31-Jan-91	19920331	0.26	0.04
DELL	14489	11081	9833	31-Jan-92	19920430	0.36	0.08

## 4. CCM web query

- The web query handles many of the merging issues discussed above.
- It provides Compustat data with CRSP identifiers.
- It has the options to:
  - Select data using GVKEY, PERMNO or other identifiers.
  - Ability to choose unique links.
  - Handle Fiscal period and link dates ranges.

# Summary

- **CRSP and Compustat Universe**
- **Merging using CUSIP**
  - a. **CUSIP codes.**
  - b. **Preparing data.**
  - c. **Merging example.**
- **CRSP-Compustat Merge (CCM) product**
  - a. **Understanding CCM linking file**
  - b. **Merging example.**
- **CCM web query**