Marvel Character Appearances

Market Basket Analysis

Whitney King

Contents

Introduction

This document will walk through the process of cleaning up a set of data that pertains to Marvel character appearances in comic books over time, as well as formatting the data for use in a market basket style data mining model looking at the frequencies various characters appear together. It will include both a how-to for the data cleanup and mining model, as well as a written section where I analyze the findings in the mining model created with the clean data. Additionally, I'll include a few SQL scripts to explore the data from the mining model.

Raw Data

The raw data we are using for this project is found in a flat file available to the public located on the Universitat De Les Illes Balears Math Department website (named porgat.txt), was obtained from the Marvel Chronology Project, and originally used for the Marvel Character Social Graph.

The data is a file that contains all of the information we will be importing into a new SQL database. In addition, there are two other flat files, Character Vertexes and Comic Book Vertexes which help us make sense of what the information in the main flat file means. At first glance this is an overwhelming document and it can be difficult to make sense of what it going on, so let's spend a moment dissecting it.

We know we are looking at data pertaining to Marvel character appearances in comic books based on the previous application of the data, so it will need to have at least three categories of information to be useful as database tables: **Characters**, **Comics** and **Appearances**. The first row of information in the text file contains **`*Vertices 19428 6486`**.

The rest of the list *appears* to be divided into two columns, one for an ID, and one for a name. When we look for the number **6486** in this data file, we can see it is a transitional point in the data where the second column restarts from the end of the alphabet to the beginning of the alphabet:

```
6486 "ZONE"
6487 "AA2 35"
```

From **1 – 6486**, the second column appears to contain hero names, however from **6487** onward, it appears the second column switches to a random abbreviation. If we compare the first set of data in this list (items **1 – 6486**) to the information in the *Character Vertexes* file, we can determine that each item in this first section is a shortened name of a comic book *character*. The Character Vertexes file contains a longer version of these character names. We will use this data for our **Marvel Characters** table, which should have 6486 characters in it assuming we treat the number column as an index.

As stated previously, from **6487** onwards the second column changes to random abbreviations. If we look for index **19428** (our other known vertex), we can see that the format of the data in this file changes completely beyond that point and this appears to be the end of the intended section. If we compare the data in the second column of this section to the information in the *Comic Book Vertexes*, we can determine that **6487 - 19428** are short form names of published comic books. This data is what we will use for our **Comic Names** table, which will have 12,941 comics in it (the end of the comic index minus the beginning of the comic index).

After item **19428**, we come to **`*Edgeslist`**, which restarts an index count at **1** and ends at **6487** (our character indexes). Each character index then has up to 15 columns of data per line pertaining to indexes in the range of **6487 – 19428** (our comic book indexes). This data is going to take some handling in Excel before we can use it but we can at least determine that this is a list of Character IDs and the Comic IDs in which they appear, and will make up our **Appearances** table. Due to the amount of work needed to clean this section of data, we don't have a total count of appearances expected in this table yet, but it's safe to assume with 6,486 characters, and 12,941 comics books this is a large number.

When all is said and done, we've just completed the first big task in figuring out how to clean and work with this data. We can take the information we've just learned, and create three new flat files, one pertaining to each section of the data.

In this project (and in the companion documents), these newly created text files are named the same way they're named in Figure 1, and they contain the information specified within each section of the original file. Once a little more cleanup is completed, these are the files that will be directly imported into a new SQL database for our model.

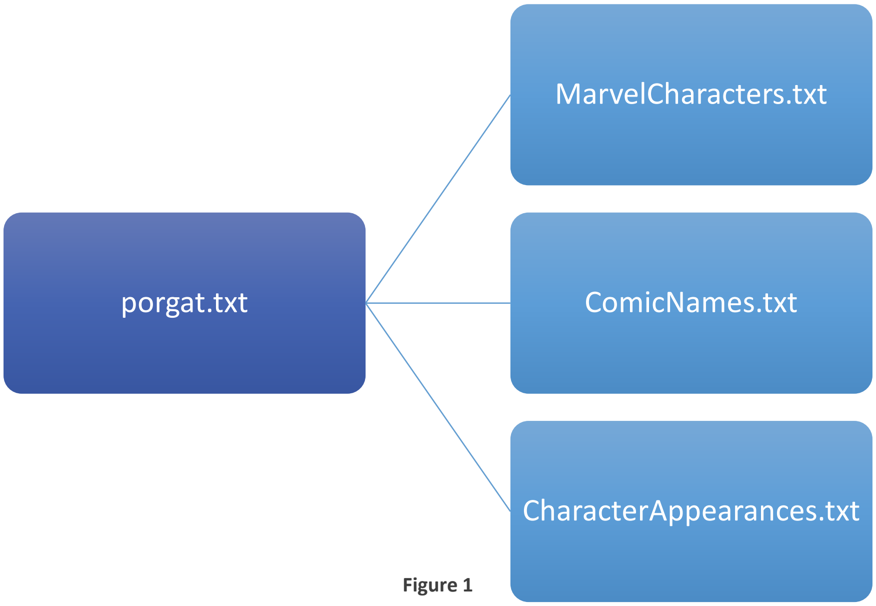


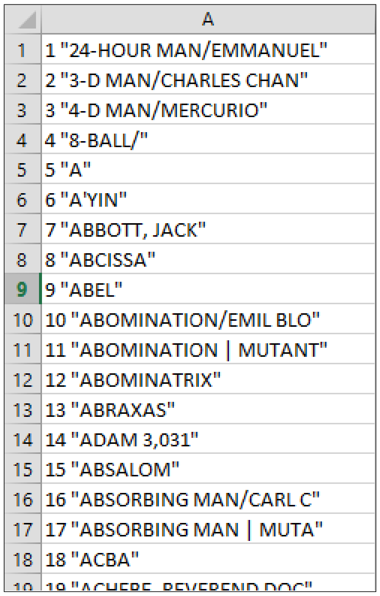
Figure 1

Data Cleanup

Now the real fun can begin! There are few distinct observations to investigate and correct (or begin correcting if we can) pertaining to each category of data that was just dissected. To make this part easier to follow, we'll break it down document by document.

MarvelCharacters.txt

Let's take a closer look at the information in this file. We're going to want to begin thinking of each column of data in this text file as a column in a SQL table. The first column (which is the index column), we will refer to as **`CharID`**. The second column is the character name, so we will call this column **`CharName`**. It's difficult to work with data in bulk in a plain text file, so we'll paste the data into Excel, and delimit the pasted data using **"** (quotation mark) as the delimiter which should parse the data out into two separate columns for us. We use the quotation mark since that is what is surrounding the data in each field of the **`CharName`** column – using it as the delimiter will not only get rid of the quotation marks, it will parse **`CharID`** and **`CharName`** into two different columns.

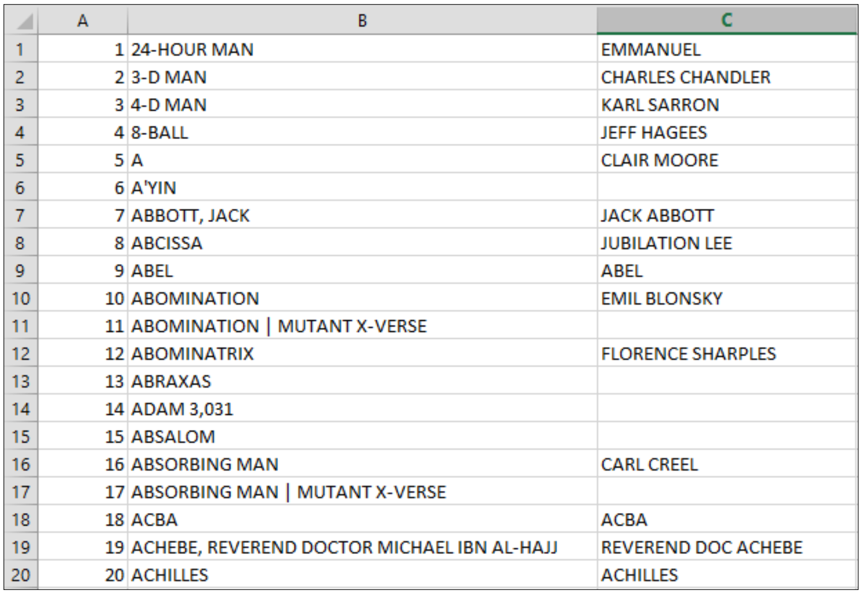


	A
1	1 "24-HOUR MAN/EMMANUEL"
2	2 "3-D MAN/CHARLES CHAN"
3	3 "4-D MAN/MERCURIO"
4	4 "8-BALL/"
5	5 "A"
6	6 "A'YIN"
7	7 "ABBOTT, JACK"
8	8 "ABCISSA"
9	9 "ABEL"
10	10 "ABOMINATION/EMIL BLO"
11	11 "ABOMINATION | MUTANT"
12	12 "ABOMINATRIX"
13	13 "ABRAXAS"
14	14 "ADAM 3,031"
15	15 "ABSALOM"
16	16 "ABSORBING MAN/CARL C"
17	17 "ABSORBING MAN | MUTA"
18	18 "ACBA"
19	19 "ACHERE, REVEREND DOC"

Example of raw Character data pasted into Excel, prior to modifications

Since this data was copied directly from the source file, the names in the **`CharName`** column are shortened, and don't include the full name of the Marvel characters we want to investigate. This is annoying, but not a work stoppage. Luckily, the *Character Vertexes* text file we used to compare this data with earlier *does* have the full name s of each character in the second column! After opening that file in Excel (this time using **: (colon)** as the delimiter since there are no quotation marks in this file), we can copy the data from the second column of Character Vertexes over the data in the **`CharName`** column of **MarvelCharacters.txt**. Now we have a **`CharName`** column that actually contains full character names.

There is one further step to take. The **CharName** column contains two names for many characters; both the super-hero/villain name, as well as the real name. While this is fine of a market basket analysis, it's nice to be able to see this information parsed out into separate fields. Using "*Text to Column*" in Excel on the **CharID** column, and with **/ (forward slash)** as the delimiter will parse the super name from the real name, and add a third column we'll name **CharRealName** to our dataset. Finally, we use a formula that will change the structure of the **CharRealName** column from **[LastName, FirstName]** to [**FirstName LastName]**. After this final change, the format of **MarvelCharacters.txt** should roughly match the format of the file included in the project files, which is the final format for this table that we will import into SQL.



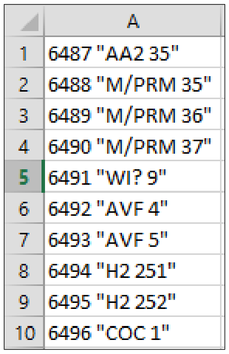
	A	B	C
1	1	24-HOUR MAN	EMMANUEL
2	2	3-D MAN	CHARLES CHANDLER
3	3	4-D MAN	KARL SARRON
4	4	8-BALL	JEFF HAGEES
5	5	A	CLAIR MOORE
6	6	A'YIN	
7	7	ABBOTT, JACK	JACK ABBOTT
8	8	ABCISSA	JUBILATION LEE
9	9	ABEL	ABEL
10	10	ABOMINATION	EMIL BLONSKY
11	11	ABOMINATION | MUTANT X-VERSE	
12	12	ABOMINATRIX	FLORENCE SHARPLES
13	13	ABRAXAS	
14	14	ADAM 3,031	
15	15	ABSALOM	
16	16	ABSORBING MAN	CARL CREEL
17	17	ABSORBING MAN | MUTANT X-VERSE	
18	18	ACBA	ACBA
19	19	ACHEBE, REVEREND DOCTOR MICHAEL IBN AL-HAJJ	REVEREND DOC ACHEBE
20	20	ACHILLES	ACHILLES

Example of MarvelCharacters.txt after cleanup

__Callouts:__ I did do a little bit of additional cleanup, as well as looked up a few more character real names in the Wikia Marvel Database for the third column while testing this data, so it may not match exactly to what it would look like generating the file using these exact steps. Additionally, since I saved the file as a text file from Excel, I lost the formula I used to change the name formatting of the real name column. There is an immense amount to discuss in this document, so I am keeping the steps for the data cleanup as concise as possible while including the pre-cleaned files in the submission.

ComicNames.txt

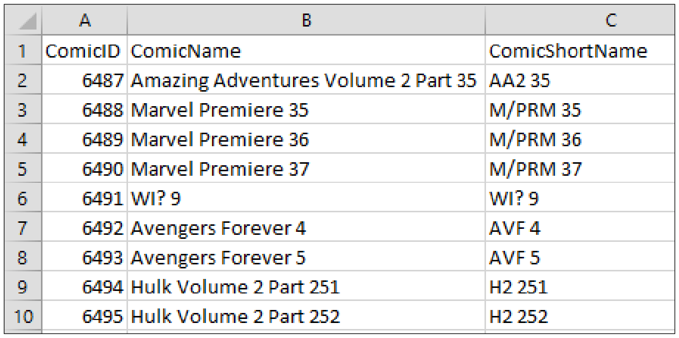
We have a very similar cleanup process for comic names as we did with character names. It starts with the same basic process of pasting the raw data into Excel, and using "*Text to Column*" with a **quotation mark** delimiter to parse the data into two columns. The *Comic Vertexes* document does not contain information about the long names of comic books, so that's not much help to us in the way the *Character Vertexes* document was for the last file. At this point, we now have **ComicID** (which begins with index 6487) and **ComicShortName** as our two columns.



	A
1	6487 "AA2 35"
2	6488 "M/PRM 35"
3	6489 "M/PRM 36"
4	6490 "M/PRM 37"
5	6491 "WI? 9"
6	6492 "AVF 4"
7	6493 "AVF 5"
8	6494 "H2 251"
9	6495 "H2 252"
10	6496 "COC 1"

Before cleanup

__Callouts__: I used the Marvel Chronology Project and Marvel Database Wikia to look up these shortened comic book name abbreviations, and created a third column in this document for __ComicName__. I had a really hard time making much sense of the data without this additional column, since the abbreviations are not very descriptive of what the comic series actually is. Needless to say, I didn't get through every comic name since there are so many of them, so some comics don't have a long form name. Most of this work was done by copying the short name column and using "find and replace".

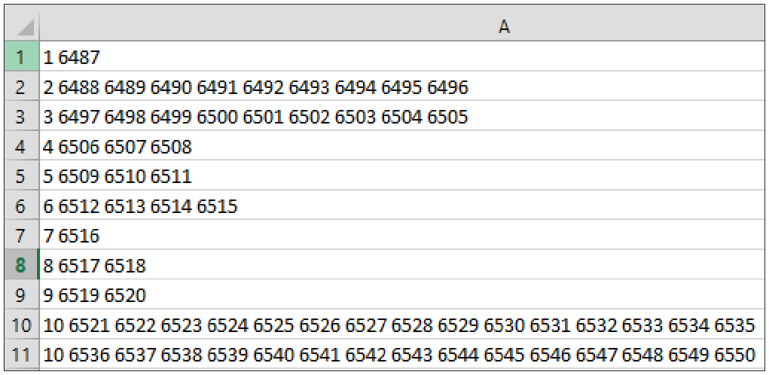


	A	B	C
1	ComicID	ComicName	ComicShortName
2	6487	Amazing Adventures Volume 2 Part 35	AA2 35
3	6488	Marvel Premiere 35	M/PRM 35
4	6489	Marvel Premiere 36	M/PRM 36
5	6490	Marvel Premiere 37	M/PRM 37
6	6491	WI? 9	WI? 9
7	6492	Avengers Forever 4	AVF 4
8	6493	Avengers Forever 5	AVF 5
9	6494	Hulk Volume 2 Part 251	H2 251
10	6495	Hulk Volume 2 Part 252	H2 252

After cleanup and addition of third column

CharacterAppearances.txt

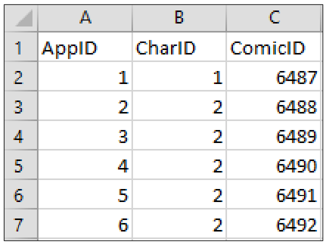
This was by far the most difficult file of the three tables to create and cleanup. The vast majority of this cleanup work was done with formulas in Excel using **CharacterAppearances.xlsx** (included with project files) to get **one comic appearance per character per line**. The data is presented as one character per line, with each line containing up to 15 comics they appeared in, delimited by spaces. Each character may have more than one line if they have more than 15 appearances.



	A
1	1 6487
2	2 6488 6489 6490 6491 6492 6493 6494 6495 6496
3	3 6497 6498 6499 6500 6501 6502 6503 6504 6505
4	4 6506 6507 6508
5	5 6509 6510 6511
6	6 6512 6513 6514 6515
7	7 6516
8	8 6517 6518
9	9 6519 6520
10	10 6521 6522 6523 6524 6525 6526 6527 6528 6529 6530 6531 6532 6533 6534 6535
11	10 6536 6537 6538 6539 6540 6541 6542 6543 6544 6545 6546 6547 6548 6549 6550

Before cleanup

First, we parse out the character indexes from the comic indexes and give them their own column. Then we use a formula to take the comic indexes and remove the list of comics one by one so we have access to each comic index individually (this sounds more complicated than it is). We have 15 columns with these reduction formulas, which feed data into an additional 14 columns with formulas that will splice in the character index and the last comic from the reduced list of comic appearances. Any of the values from these clean columns that contain two numbers (**`CharID`** and **`ComicID`**), were then copy/pasted to a new sheet in the workbook so they were all in the same column, and then parsed one more time to separate the **`CharID`** and **`ComicID`** into two columns. Finally, one more sheet was made where these two columns were copy/pasted, and one additional column was added to count each appearance from one upwards and give each row a unique identifier (**`AppID`**). When everything was completed, the final count of appearances in this table is ***91,952***. After the final list with **`AppID`** included was made, the data was finally copied into a flat file like the first two sets of data named **CharacterAppearances.txt**.



	A	B	C
1	AppID	CharID	ComicID
2	1	1	6487
3	2	2	6488
4	3	2	6489
5	4	2	6490
6	5	2	6491
7	6	2	6492

After cleanup

Callouts: *Since the work to parse this data was so complicated, it's been retained as an Excel file so it can be looked at more closely if so desired. There were a lot of various formulas used, and I felt that the majority of that explanation was outside the scope of this document.*

Data Limitations

There are a few limitations on the dataset we've just cleaned up, but nothing that should prevent a market basket analysis from being done on this data. We can think of each comic book as a basket of items, and each character as an item in the basket. The limitations we have are all "nice to have" features, but we'll quickly address them. Much of this desirable information is out there on the Chronology Project or the Marvel Database Wikia, it's just not parsed or compiled in a manner that makes it immediately useful for our purposes.

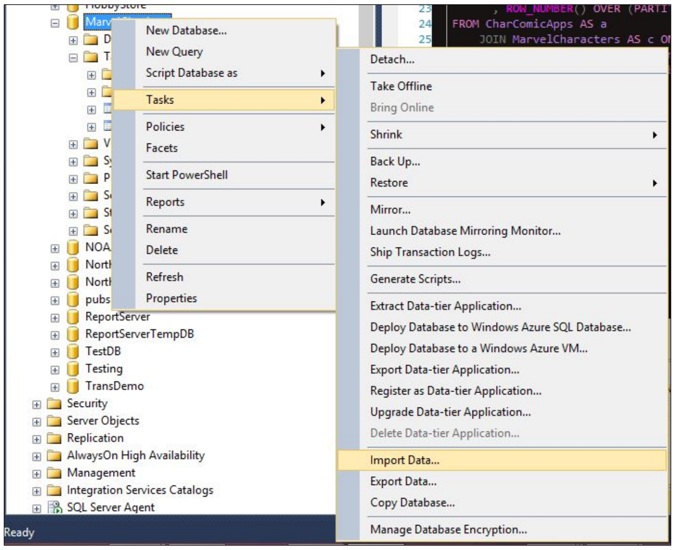
- **Lack of facts about each character;** knowing things like gender or specific universe each character is would add new dimension to the data, and allow other types of analysis to be done on this data, such as decision trees.
- **Lack of comic publication dates;** this limits our exploration of the data
- **Lack of long names for many comics and characters;** while I've added some of this information and done a lot of legwork to make things readable and to be more specific about things like comic names, there are so many of them that I just didn't get to them all. At least major things should be covered.
- **Lack of date of last publication or up to date data;** once again, dates aren't needed for this specific type of analysis, but they would certainly add an interesting dynamic to the overall quality and malleability of the data in this set.

- We can certainly assume there are a lot of more recent publications since Marvel is still going strong (the last date from this set is either 2002 or 2011 based on my research, but I haven't located a firm date unfortunately).

Creating the Database

The majority of the rest of this document will be focused around the creation and analysis of the data mining model based on the data we just cleaned, however the creation of the SQL database warrants mention. In creating this project, flat files were prepped to be imported into a new database named `MarvelCharApps` that was created with simple SQL scripts. Once created, the flat files were imported into individual tables using **Tasks -> Import Data…** three separate times. No primary or foreign keys are necessary to analyze the data in this database. Once the flat files are imported, we will have three tables: `MarvelCharacters`, `ComicNames`, and `CharComicApps`.

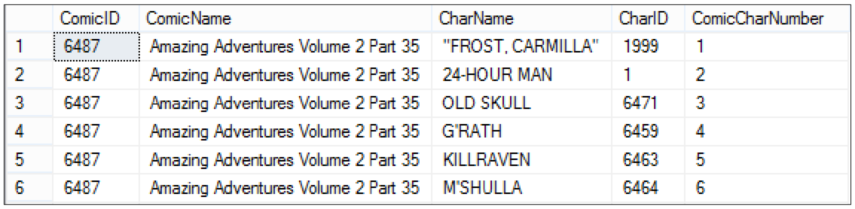
In addition to the flat files included with this document, there are some SQL script files and .mdf/.ldf files that can be used to load this database on another SQL server.



```
23     , ROW_NUMBER() OVER (PARTI
24   FROM CharComicApps AS a
25       JOIN MarvelCharacters AS c ON
```

Accuracy Check

Before delving into the data using an analysis services project, we should quickly verify the data in the database actually works as expected. Included in the **MarvelAppearances.sql** file, is the following script to display each comic, alongside a list of which characters and how many appeared in each comic book. If this script works, we can proceed towards creating the mining model:



	ComicID	ComicName	CharName	CharID	ComicCharNumber
1	6487	Amazing Adventures Volume 2 Part 35	"FROST, CARMILLA"	1999	1
2	6487	Amazing Adventures Volume 2 Part 35	24-HOUR MAN	1	2
3	6487	Amazing Adventures Volume 2 Part 35	OLD SKULL	6471	3
4	6487	Amazing Adventures Volume 2 Part 35	G'RATH	6459	4
5	6487	Amazing Adventures Volume 2 Part 35	KILLRAVEN	6463	5
6	6487	Amazing Adventures Volume 2 Part 35	M'SHULLA	6464	6

Script results

```
--CREATE VIEW CharacterAppearances AS
       SELECT  n.ComicID
             , n.ComicName
             , c.CharName
             , c.CharID
             , ROW_NUMBER() OVER (PARTITION BY n.ComicName ORDER BY n.ComicName) AS ComicCharNumber
       FROM CharComicApps AS a
              JOIN MarvelCharacters AS c ON a.CharID = c.CharID
              JOIN ComicNames AS n ON a.ComicID = n.ComicID
       ORDER BY n.ComicID, ComicCharNumber;
```

Creating Views for Analysis

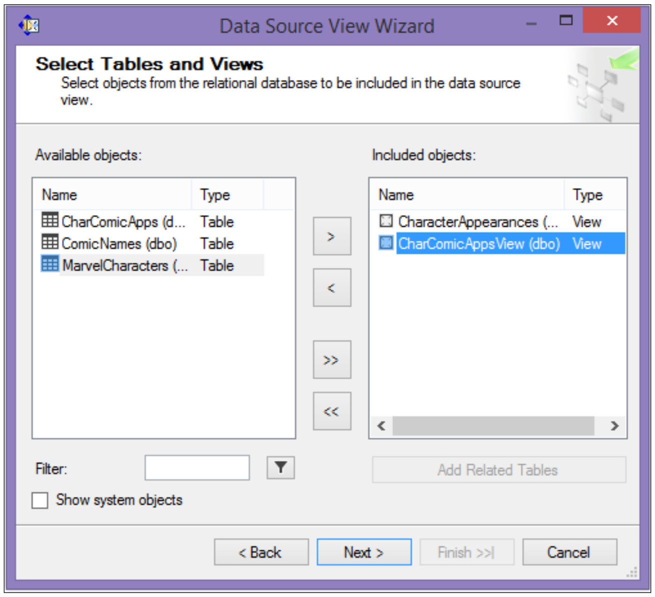
Using the above script from **MarvelAppearances.sql** as a foundation, we'll create two views in our `MarvelCharApps` database that will be used in the **Data Source View** of the mining model we're going to create. There are two scripts in this file that have the "`CREATE VIEW … AS`" line of the view commented out (so we can simply use the select statement if we choose). Not all the columns in each view are used in the model, however these views are both readable and functional for our purposes.

We'll create two slightly different views of the data called `CharacterAppearances` and `CharComicAppsView`. Since the first view is already written in the above section, only the second view is included below:

```
CREATE VIEW CharComicAppsView AS
SELECT  n.ComicID
      , n.ComicName
      , c.CharName
      , c.CharID
FROM CharComicApps AS a
       JOIN MarvelCharacters AS c ON a.CharID = c.CharID
       JOIN ComicNames AS n ON a.ComicID = n.ComicID
```

Creating the Mining Model

After opening a new analysis services project in Visual Studio and setting up a new data source using our MarvelCharApps database, we'll create a new **Data Source View** using the two views we made in SQL Server as the included objects.



Data Source View Wizard

Select Tables and Views
Select objects from the relational database to be included in the data source view.

Available objects:

Name	Type
CharComicApps (d...	Table
ComicNames (dbo)	Table
MarvelCharacters (...	Table

Included objects:

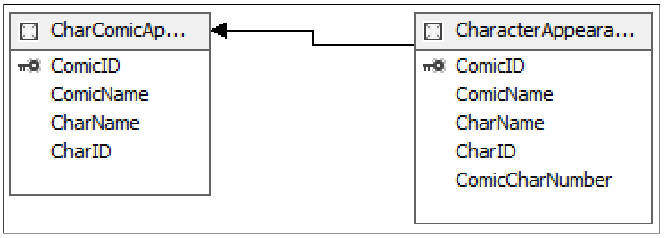
Name	Type
CharacterAppearances (...	View
CharComicAppsView (dbo)	View

Filter:

☐ Show system objects

Data Source View Creation

Once the new data source view has loaded, drag ComicID from the CharacterAppearances table over to ComicID in the CharComicAppsView table to create a relationship between the two tables.



CharComicAp...
- ComicID
- ComicName
- CharName
- CharID

CharacterAppeara...
- ComicID
- ComicName
- CharName
- CharID
- ComicCharNumber

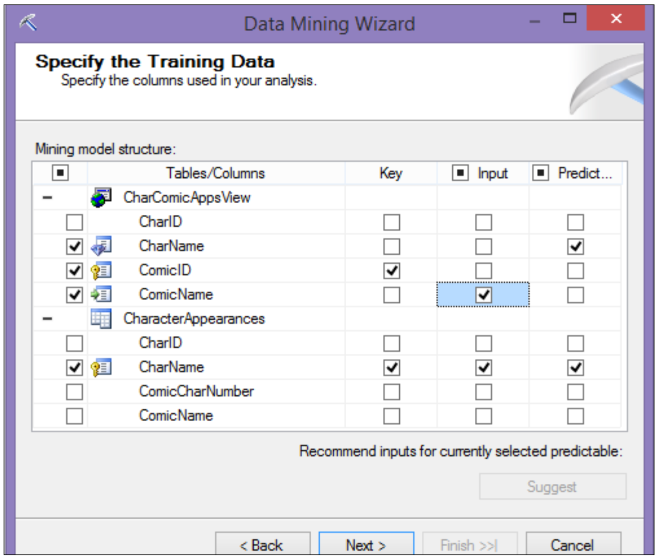
Data Source View with Related Tables

Next, create a new mining model using:

- Microsoft Association Rules
- The newly created data source view
- CharComicAppsView as the case table
- CharacterAppearances as the nested table

For the columns used in the training data, select:

- **CharComicAppsView**
  - o CharName as the prediction (we want to know which characters show up together frequently in the same comics)
  - o ComicID as the key (comics being analogous to a shopping basket)



Data Mining Wizard

Specify the Training Data
Specify the columns used in your analysis.

Mining model structure:

■	Tables/Columns	Key	■ Input	■ Predict...
−	CharComicAppsView			
☐	CharID	☐	☐	☐
☑	CharName	☐	☐	☑
☑	ComicID	☑	☐	☐
☑	ComicName	☐	☑	☐
−	CharacterAppearances			
☐	CharID	☐	☐	☐
☑	CharName	☑	☑	☑
☐	ComicCharNumber	☐	☐	☐
☐	ComicName	☐	☐	☐

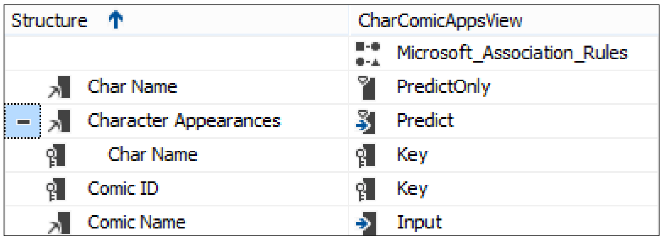
Recommend inputs for currently selected predictable:

Mining Model Association Rule Data Structure

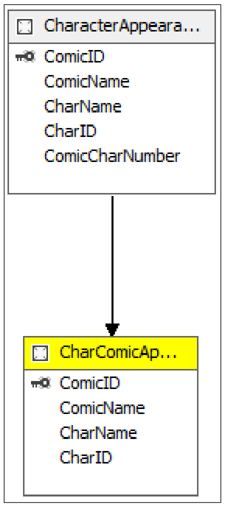
    - `ComicName` as the input (this is what we're grouping the character on)
- **`CharacterAppearances`**
    - `CharName` as the key, prediction and input

When asked about reserving data for testing, choose 0%, on the next screen allow drill through for the data then choose a suitable name, and finally select finish to create the mining structure.

On the **Mining Models** tab of the new structure we created, the setup should look like the image below:



Structure ↑	CharComicAppsView
	Microsoft_Association_Rules
Char Name	PredictOnly
− Character Appearances	Predict
Char Name	Key
Comic ID	Key
Comic Name	Input



CharacterAppeara...
ComicID
ComicName
CharName
CharID
ComicCharNumber

CharComicAp...
ComicID
ComicName
CharName
CharID

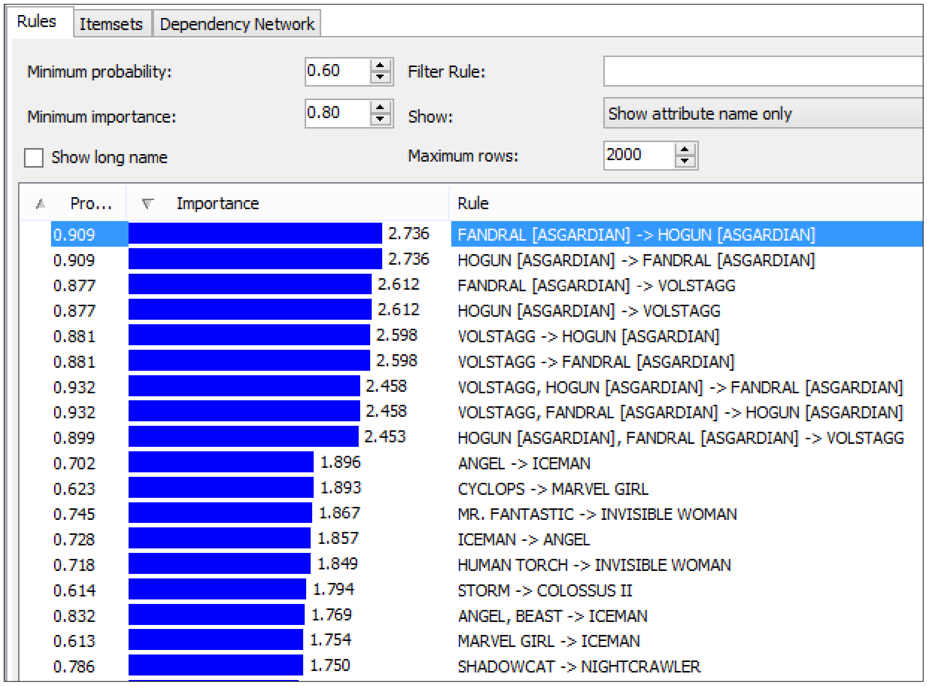
Completed Mining Structure

Setup the Mining Model Viewer

Select the **Mining Model Viewer** tab to be prompted to build and deploy the mining model. Once the model has been built and deployed, you will finally be presented with a list of the association rules for the Marvel characters appearing in comic books in our dataset.

The settings used in the analysis of these results on the **Rules** tab of the mining model are as follows:

- **Minimum Probability**: 0.60
    - This is sets of characters that have more than a 60% chance of appearing together
- **Minimum Importance**: 0.80
    - This is the "trustworthiness" of the rule, so we want it relatively high as a minimum
- **Show**: Show Attribute Name Only
    - This is much easier to read without the word "exists" all over the screen
- **Sort**: Importance
    - This will move probability to second level sort
    - Gives us a view of the rules with both high probability and high importance at the top



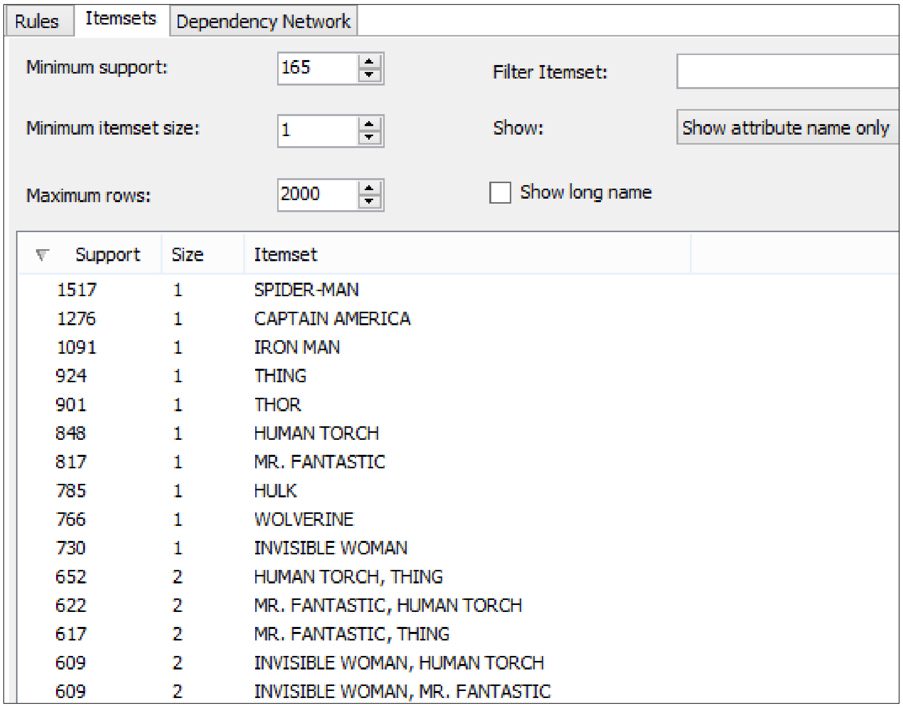
Minimum probability: 0.60
Minimum importance: 0.80
□ Show long name
Filter Rule:
Show: Show attribute name only
Maximum rows: 2000

Pro...	Importance		Rule
0.909		2.736	FANDRAL [ASGARDIAN] -> HOGUN [ASGARDIAN]
0.909		2.736	HOGUN [ASGARDIAN] -> FANDRAL [ASGARDIAN]
0.877		2.612	FANDRAL [ASGARDIAN] -> VOLSTAGG
0.877		2.612	HOGUN [ASGARDIAN] -> VOLSTAGG
0.881		2.598	VOLSTAGG -> HOGUN [ASGARDIAN]
0.881		2.598	VOLSTAGG -> FANDRAL [ASGARDIAN]
0.932		2.458	VOLSTAGG, HOGUN [ASGARDIAN] -> FANDRAL [ASGARDIAN]
0.932		2.458	VOLSTAGG, FANDRAL [ASGARDIAN] -> HOGUN [ASGARDIAN]
0.899		2.453	HOGUN [ASGARDIAN], FANDRAL [ASGARDIAN] -> VOLSTAGG
0.702		1.896	ANGEL -> ICEMAN
0.623		1.893	CYCLOPS -> MARVEL GIRL
0.745		1.867	MR. FANTASTIC -> INVISIBLE WOMAN
0.728		1.857	ICEMAN -> ANGEL
0.718		1.849	HUMAN TORCH -> INVISIBLE WOMAN
0.614		1.794	STORM -> COLOSSUS II
0.832		1.769	ANGEL, BEAST -> ICEMAN
0.613		1.754	MARVEL GIRL -> ICEMAN
0.786		1.750	SHADOWCAT -> NIGHTCRAWLER

Top results for Rules of the processed mining model

The settings used in the analysis of these results on the **Itemsets** tab are as follows:

- **Minimum Support:** 165
    - This is the minimum number of total appearances in comics for each itemset
- **Minimum Itemset Size:** 1
    - We will change this number to 2 and then 3 as part of the data analysis, but we'll start with 1 so we can see the most frequently occurring characters overall, as well as frequently occurring *sets* of characters
- **Show**: Show Attribute Name Only
    - This is much easier to read without the word "exists" all over the screen



Minimum support: 165

Filter Itemset:

Minimum itemset size: 1

Show: Show attribute name only

Maximum rows: 2000

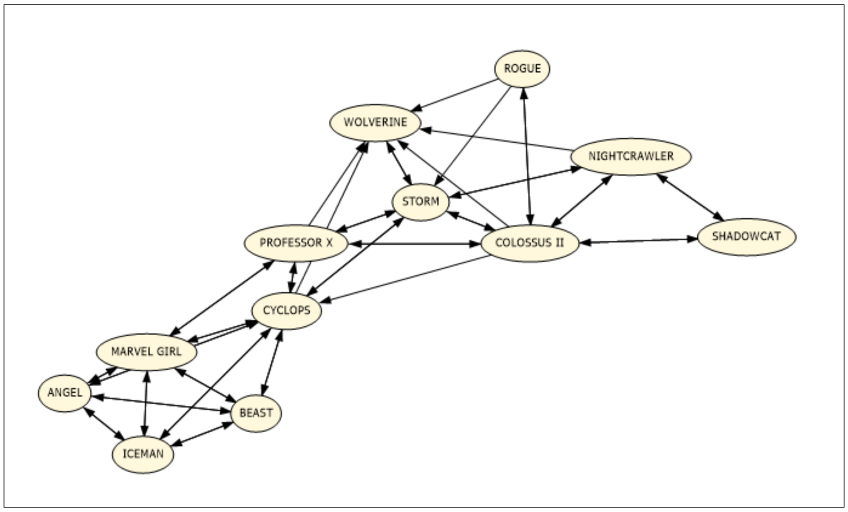
□ Show long name

Support	Size	Itemset
1517	1	SPIDER-MAN
1276	1	CAPTAIN AMERICA
1091	1	IRON MAN
924	1	THING
901	1	THOR
848	1	HUMAN TORCH
817	1	MR. FANTASTIC
785	1	HULK
766	1	WOLVERINE
730	1	INVISIBLE WOMAN
652	2	HUMAN TORCH, THING
622	2	MR. FANTASTIC, HUMAN TORCH
617	2	MR. FANTASTIC, THING
609	2	INVISIBLE WOMAN, HUMAN TORCH
609	2	INVISIBLE WOMAN, MR. FANTASTIC

Top results for Itemsets in the processed mining model

The settings used in the analysis of these results on the **Dependency Network** tab are as follows:

- **Show**: Show Attribute Name Only
    - This is much easier to read without the word "exists" all over the screen



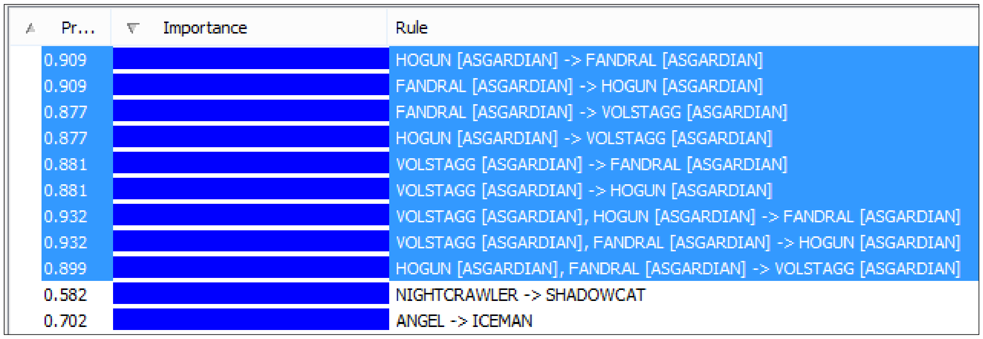
Partial view of the Dependency Network in the mining model: X-Men

Mining Model Analysis

Even starting with quick glances at each tab of the mining model viewer, there is some really cool information that surfaces about the relationships between Marvel characters as they appear in comic books. Diving into this data analysis it's important to keep in mind that this isn't a completely up to date set of information, as well as the fact that we're not seeing absolutely every relationship between every comic book character; just the most frequently occurring ones (characters appearing less than 165 times are not included). With that said, there's a lot to be learned here.

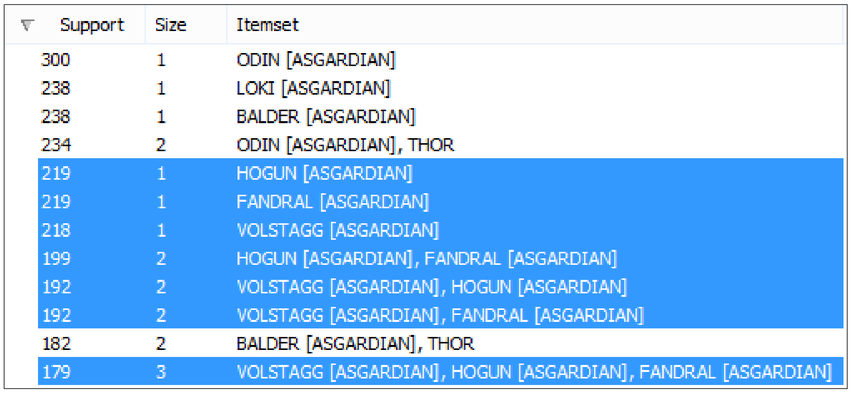
The Warriors Three

When the item grouping on the **Rules** tab is completed in the manner described in the setup steps, the top 9 rules all pertain to various permutations of groupings of the same three characters in the same comic books: **Fandral**, **Hogun** and **Volstagg**. **This may come as a surprise since there are many more well-known characters in the Marvel universe, however it's important to remember that the strength of a rule on this tab doesn't necessarily mean these characters are the most frequently occurring in the universe; it means they're the most likely characters to appear *together*.**



Pr...	Importance	Rule
0.909	█	HOGUN [ASGARDIAN] -> FANDRAL [ASGARDIAN]
0.909	█	FANDRAL [ASGARDIAN] -> HOGUN [ASGARDIAN]
0.877	█	FANDRAL [ASGARDIAN] -> VOLSTAGG [ASGARDIAN]
0.877	█	HOGUN [ASGARDIAN] -> VOLSTAGG [ASGARDIAN]
0.881	█	VOLSTAGG [ASGARDIAN] -> FANDRAL [ASGARDIAN]
0.881	█	VOLSTAGG [ASGARDIAN] -> HOGUN [ASGARDIAN]
0.932	█	VOLSTAGG [ASGARDIAN], HOGUN [ASGARDIAN] -> FANDRAL [ASGARDIAN]
0.932	█	VOLSTAGG [ASGARDIAN], FANDRAL [ASGARDIAN] -> HOGUN [ASGARDIAN]
0.899	█	HOGUN [ASGARDIAN], FANDRAL [ASGARDIAN] -> VOLSTAGG [ASGARDIAN]
0.582	█	NIGHTCRAWLER -> SHADOWCAT
0.702	█	ANGEL -> ICEMAN

"Warriors Three" – Dominating the top of the rule list

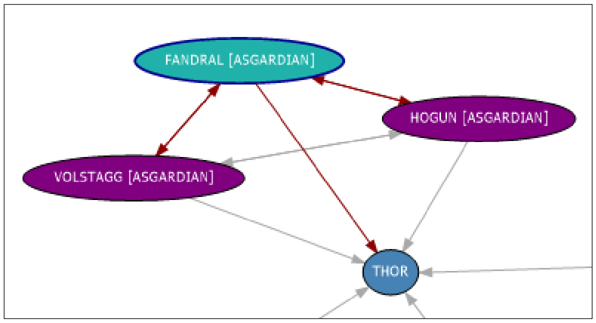


Support	Size	Itemset
300	1	ODIN [ASGARDIAN]
238	1	LOKI [ASGARDIAN]
238	1	BALDER [ASGARDIAN]
234	2	ODIN [ASGARDIAN], THOR
219	1	HOGUN [ASGARDIAN]
219	1	FANDRAL [ASGARDIAN]
218	1	VOLSTAGG [ASGARDIAN]
199	2	HOGUN [ASGARDIAN], FANDRAL [ASGARDIAN]
192	2	VOLSTAGG [ASGARDIAN], HOGUN [ASGARDIAN]
192	2	VOLSTAGG [ASGARDIAN], FANDRAL [ASGARDIAN]
182	2	BALDER [ASGARDIAN], THOR
179	3	VOLSTAGG [ASGARDIAN], HOGUN [ASGARDIAN], FANDRAL [ASGARDIAN]

Total appearances of Volstagg, Hogun and Fandral

If you're unfamiliar with the primary Marvel Universe (Earth-616), these three characters are Asgardian warriors and allies of Thor known as the Warriors Three. Knowing that context behind these characters alone definitely explains why they're seen together so frequently with 179 out of about 219 appearances each.

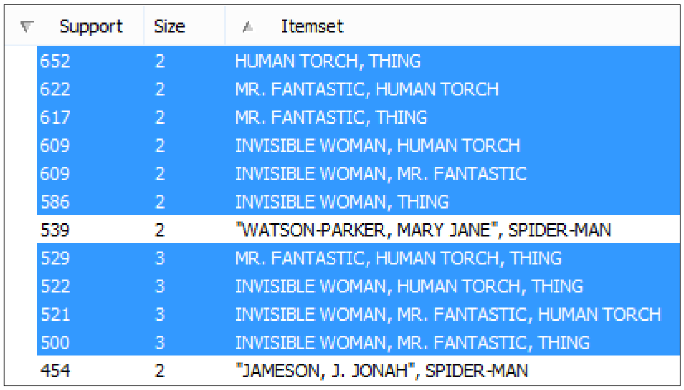
Doing a search with **[ASGARDIAN]** as the **filter** on the **Itemsets** tab will show us the number of appearances total for each of these characters (in addition to a few more Asgardians that appear frequently in Marvel comics). We find that none of these three characters take the top spot for Asgardian appearances by themselves, but when the appearances are thought of as a unit of the Warriors Three and not just single characters, they are the most frequently appearing Asgardians. In addition to knowing these characters are likely to appear together, **if we investigate the Dependency Network we can confirm that these characters appearing (or appearing together) will also predict Thor appearing (and having the most loyal sidekicks).** This view confirms the Warrior Three's strong supporting friendship to this primary and more well-known Marvel hero, and their ties to each other.



Volstagg/Hogun/Fandral predict both ways, they all predict Thor

The Fantastic Four

If we return to the **Itemsets** tab, clear the filter from the previous section and switch the **Minimum Itemset Size** to **2**, we will begin to see another trend emerge. Suddenly the character sets with the most support include four different characters. **Mister Fantastic**, the **Invisible Woman**, the **Human Torch**, and the **Thing**; better known as the **Fantastic Four**, one of Marvel's most iconic groups of superheroes. **Every single one of these characters appears alongside one of the other characters from the Fantastic Four in at least 500 comics.**

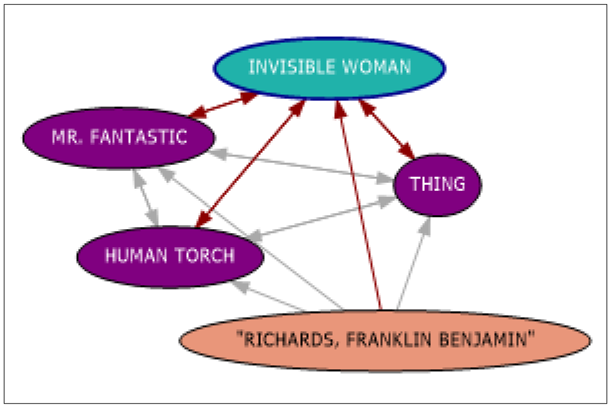


Support	Size	Itemset
652	2	HUMAN TORCH, THING
622	2	MR. FANTASTIC, HUMAN TORCH
617	2	MR. FANTASTIC, THING
609	2	INVISIBLE WOMAN, HUMAN TORCH
609	2	INVISIBLE WOMAN, MR. FANTASTIC
586	2	INVISIBLE WOMAN, THING
539	2	"WATSON-PARKER, MARY JANE", SPIDER-MAN
529	3	MR. FANTASTIC, HUMAN TORCH, THING
522	3	INVISIBLE WOMAN, HUMAN TORCH, THING
521	3	INVISIBLE WOMAN, MR. FANTASTIC, HUMAN TORCH
500	3	INVISIBLE WOMAN, MR. FANTASTIC, THING
454	2	"JAMESON, J. JONAH", SPIDER-MAN

Fantastic Four emerges on the Itemsets tab

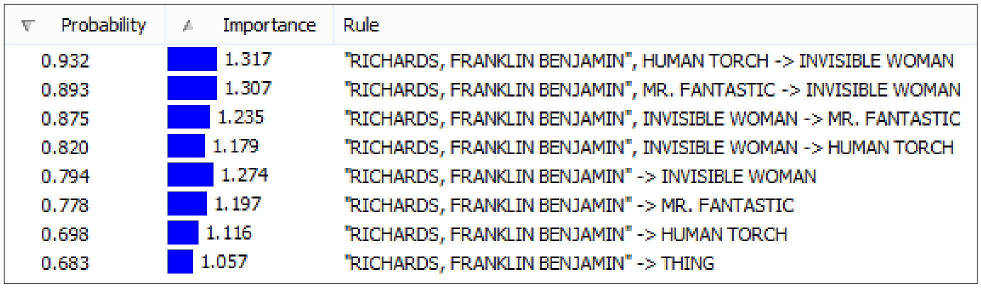
Not only are the Fantastic Four frequently occurring with one another, they're frequently occurring *period*, making up a distinct social grouping within the Marvel Universe that can be seen if we investigate further on the Dependency Network tab. Each one of these four characters predicts the appearance of the others in the squad.

Additionally, we see a fifth character appear in this network… one that predicts the appearance of the Fantastic Four (not the other way around). **Franklin Benjamin Richards**, the son of **Mister Fantastic** (a.k.a. Reed Richards) and the **Invisible Woman** (a.k.a. Sue Storm). Knowing **Mister Fantastic** and the **Invisible Woman** are married and have a son as a major plot point in their story arc certainly explains the fifth person cramping the Fantastic Four's style.



Fantastic Four and son

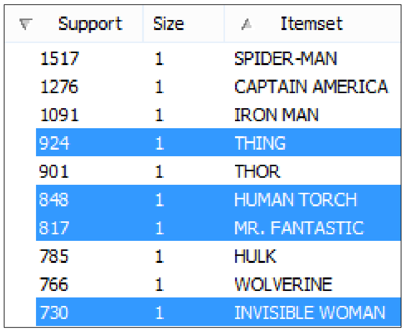
In fact, if we look at the **Rules** tab and sort the results based on **probability** first instead of **importance**, we'll see one of the strongest bonds possible emerge towards the top; one between mother and son. Filter the rules results using **Richards**, and not only does this relationship stand out, the theme of family resounds in the way the rules are formed. The **Invisible Woman** and **Mister Fantastic** are married and have a son. This can be seen. The **Human Torch** is the **Invisible Woman**'s brother and **Franklin**'s uncle. This can be seen. Rule after rule, we can now see something that wasn't obvious in the data prior to executing this model: **family**, and how it makes up the framework of the stories that built the Fantastic Four universe.



Probability	Importance	Rule
0.932	1.317	"RICHARDS, FRANKLIN BENJAMIN", HUMAN TORCH -> INVISIBLE WOMAN
0.893	1.307	"RICHARDS, FRANKLIN BENJAMIN", MR. FANTASTIC -> INVISIBLE WOMAN
0.875	1.235	"RICHARDS, FRANKLIN BENJAMIN", INVISIBLE WOMAN -> MR. FANTASTIC
0.820	1.179	"RICHARDS, FRANKLIN BENJAMIN", INVISIBLE WOMAN -> HUMAN TORCH
0.794	1.274	"RICHARDS, FRANKLIN BENJAMIN" -> INVISIBLE WOMAN
0.778	1.197	"RICHARDS, FRANKLIN BENJAMIN" -> MR. FANTASTIC
0.698	1.116	"RICHARDS, FRANKLIN BENJAMIN" -> HUMAN TORCH
0.683	1.057	"RICHARDS, FRANKLIN BENJAMIN" -> THING

Fantastic Four Association Rules

If we return to the **Itemsets** tab and set the **Minimum Itemset Size** back to **1** so we can see overall appearances for each individual character, we will also find that **all four members of the Fantastic Four are in the top ten most frequently appearing characters overall**. Not only that, but **the Invisible Woman is the most frequently occurring female character in the entire Marvel universe! Now that truly is fantastic!**



Support	Size	Itemset
1517	1	SPIDER-MAN
1276	1	CAPTAIN AMERICA
1091	1	IRON MAN
924	1	THING
901	1	THOR
848	1	HUMAN TORCH
817	1	MR. FANTASTIC
785	1	HULK
766	1	WOLVERINE
730	1	INVISIBLE WOMAN

Top Ten Characters by Appearances

Social Groupings

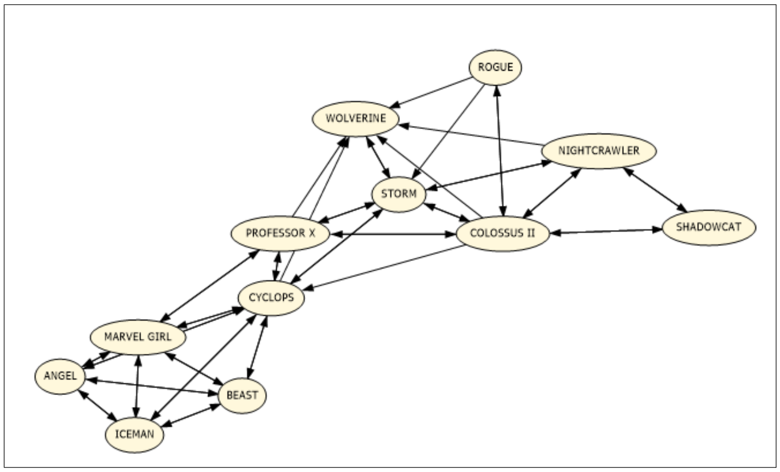
The **Dependency Network** for the entire mining model ended up being so interesting that it warrants some analysis in and of itself. Zooming out on the whole network when it is limited to characters with more than 165 appearances gives us a view with **8 distinct groups of characters**. Based on the structure of each of these groups, we can give each one a name based on their primary super hero or network of superheroes:

- Spider-Man
- The Fantastic Four
- The Avengers
- Thor
- X-Men
- Daredevil
- The Hulk
- Doctor Strange

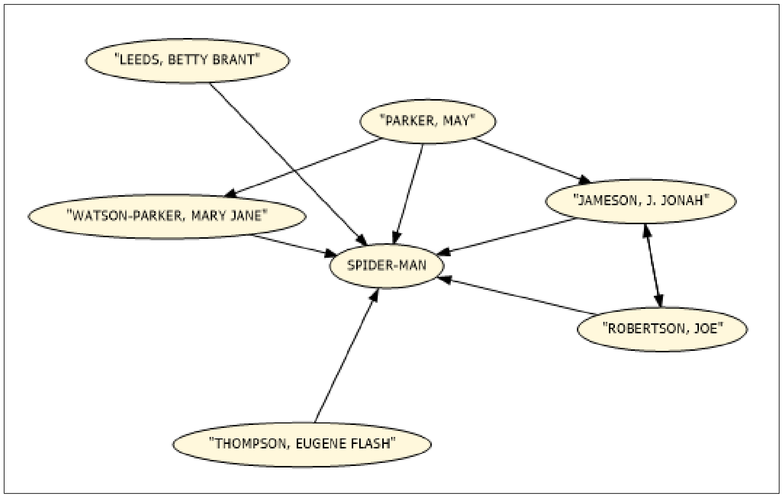
There are crossover comic books where these separate groups of superheroes collide, but either they don't appear together enough, or the characters from the universes that crossover appear so infrequently that they didn't factor into the relationships we're seeing here. This grouping is good because it allows us the ability to distinguish these groups (not all of which are always on Earth-616) from one another at a quick glance.

As a manual testing and validation of method for the data in the dependency network, a second SQL query file is included called **MarvelQuery.sql**. There are a handful of various scripts to play around with included in this file, however there are some pertinent examples like all appearances of the **Invisible Woman**:

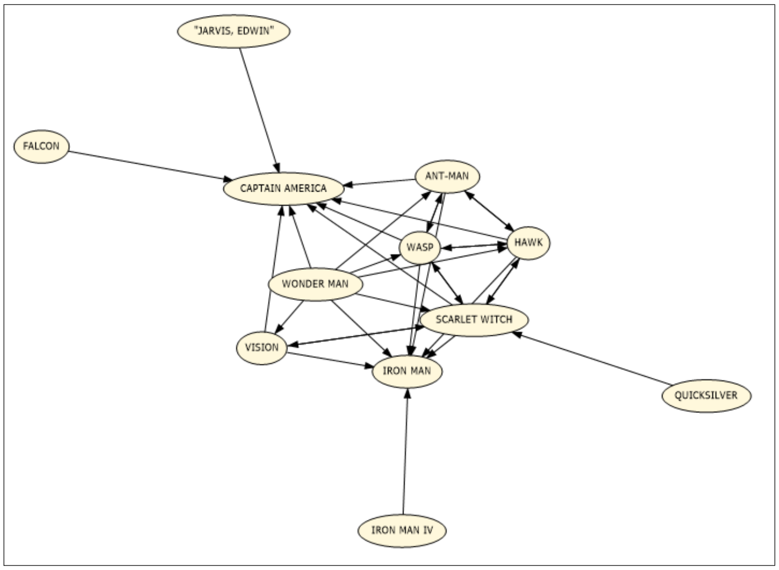
```
--Appearances of the Invisible Woman
SELECT  AppID
      , c.CharName
         , n.ComicName
FROM CharComicApps AS a
       JOIN MarvelCharacters AS c ON a.CharID
= c.CharID
       JOIN ComicNames AS n ON a.ComicID =
n.ComicID
  WHERE c.CharName LIKE 'INVISIBLE WOMAN'
  ORDER BY ComicName;
```



X-Men Network



Spider-Man Network



Avengers Network

Conclusion

The analysis done in this document is just scratching the surface of what you can begin to look at and find with this data. With more work and time spent adding and cleaning up the information, there is a world of possibility that could be explored here, including applying different types of mining algorithms to the data.

Any specific details about individual characters and the Marvel universe that weren't part of the Chronology Project's dataset were obtained through the Wikia Marvel Database. The rest of the resources cited in the bibliography were called out in the text of this document when they were used.

This project was done in a manner that didn't include a hypothesis about what was going to be found after running the mining model, and the analysis was done strictly based on the results. Frankly, going into this I wasn't even sure I was going to be able to pull this off after weighing the initial state of the data against my skill level. With that said, it did turn out as I'd hoped, and both the database by itself and market basket mining model could be used to help answer a myriad of interesting questions about the social happenings in the Marvel Universe.

Bibliography

Chang, K., Turner, T., & Braswell, J. (2015, May 28). Marvel Universe Social Graph. Retrieved June 9, 2015, from http://exposedata.com/marvel/

Chappell, R., Bourcier, P., & Jensen, D. (2015, February 1). Marvel Chronology Project - Main. Retrieved June 9, 2015, from http://www.chronologyproject.com/

Miro, J., Rosselló, C., & Alberich, R. (2002, February 1). Social characteristics of the Marvel Universe. Retrieved June 9, 2015, from http://bioinfo.uib.es/~joemiro/marvel.html

Various Contributors. (2015, June 9). Marvel Database. Retrieved June 11, 2015, from http://marvel.wikia.com/Main_Page