SNA Class Project - Marvel Universe Revisited

Rick Riccelli

April 23, 2013

Goal

To apply metrics and learnings from the Coursera Social Network Analysis class thru application to a real-world network (and to have a little fun doing it).

Data Criteria and Selection

The assignment statement allowed discretion in selecting or creating a dataset for analysis. A Google search provided pointers to a variety of existing data sets from multiple online sources. Selection criteria were established as follows:

- Data accessible, familiar and interesting
- Not too large for available computing resource (gen 2 core i5-based laptop)
- Previous work available for guidance, comparison, reference

After scanning the options available from multiple social network data archives (Gephi, InfoChimp, Columbia, others), the Marvel Comics collaboration network was selected.

This dataset was available in Gephi-readable format and clearly met the "familiar and interesting" criteria. It was generated from the Marvel Chronology Project (http://www.chronologyproject.com) database, in which all characters appearing in each of approximately 13,000 issues of Marvel comic books were identified, resulting in a bipartite-structured (characters and issues) collaboration network. This network was the focus of previous work by Alberich et al 2002 and Gleiser 2007 and was used by the winners of a Gephi visualization competition.

Although smaller (19,291 nodes and 96,519 edges) than many other datasets available online, the Marvel Comics collaboration network would prove to be on the upper end of the "not too large" criteria, at least relative to Gephi's capabilities in the hands of a novice user.

In addition to fulfilling the above selection criteria, the Marvel Comics collaboration network also provided an opportunity to:

- Try "projecting" a bipartite network as Lada mentioned during Week 4 Video B
- Take a more detailed look at characteristics of collaboration networks

Guidance and Data from Previous Work

Approach and methods used in this project were influenced and guided by previous work on the Marvel Comics collaboration network by Alberich et al and Gleiser. (Abbreviated

abstracts given below.)

- R. Alberich, J. Miro-Julia, F. Rosello; “Marvel Universe looks almost like a real social network”; arXiv:cond-mat/0202174v1 [cond-mat.dis-nn] 11 Feb 2002

  Alberich presents a detailed discussion of the Marvel comics characters and universe, calculates and discusses some network metrics and compares them to an Erdos-Renyi random model, concluding that the Marvel Universe collaboration network “is clearly not a random network, and that it has most but not all of the characteristics of ‘real-life’ collaboration networks”

- P.M. Gleiser; “How to become a superhero”; arXiv:0708.2410v1 [physics.soc-ph] 17 Aug 2007

  Gleiser’s analysis of degree correlations reveals disassortative degree mixing, different from most ‘real-world’ social networks. He then studies the system as a weighted network to find and characterize communities. Through the analysis of the community structure and the clustering as a function of degree, it is shown that the network presents a hierarchical structure.

Additional explanations, background and metrics data on various collaboration networks was drawn the from Newman and Ramasco articles below.

- M.E.J. Newman; “The structure of scientific collaboration networks”; PNAS Jan.16, 2001, vol 98, no 2; pp 404-409

- M.E.J. Newman; “The structure and function of complex networks”; SIAM Review, vol 45, no 2, pp 167–256

- J.J. Ramasco, S.N. Dorogovtsev, R. Pastor-Satorras; “Self-organization of collaboration networks; arXiv:cond-mat/0403438v2 [cond-mat.stat-mech] 15 Sept 2004

Project Roadmap

- Initial review of dataset in bipartite form, compare prior metrics
- Convert Marvel Universe data to unipartite
- Analyze metrics
- Compare to E-R random graph
- Compare MU metrics to other collaboration networks
- Visualization

Marvel Comics Bipartite Network

The starting dataset for study of the Marvel Comics collaboration network was downloaded from InfoChimp, which provided tabular character-issue edge data as shown in Figure 1 for

the bipartite graph in .csv format, easily processed and imported into Gephi and Excel.

Character	Issue
FROST, CARMILLA	AA2 35
WASP/JANET VAN DYNE	AVF 4
CAPTAIN AMERICA	AVF 4
HAWK	AVF 4
ANT-MAN/DR. HENRY J.	AVF 4
CAPTAIN AMERICA	AVF 5
HAWK	AVF 5
ANT-MAN/DR. HENRY J.	AVF 5
WASP/JANET VAN DYNE	AVF 5
BANNER, BETTY ROSS T	H2 251
HULK/DR. ROBERT BRUC	H2 251

Figure 1 - Excerpt of input dataset for Marvel bipartite network, listing every character appearing in each issue. Issues are indicated by an abbreviation of the issue's title and number.

Using Gephi with the Multi-Mode Network Transformation plugin and Excel, analysis of the Marvel bipartite network dataset provided network metrics as follows:

Network type	**Directed**
Nodes N	19,291
- Character nodes	6,467 (33.5%)
- Issue nodes	12,824 (66.5%)
Edges m	96,519
Issues/character	15
Characters/issue	7.5
Average degree z	5.003
Maximum degree	1,625
Avg path length APL	4.47
Giant Component	19,230 (99.7%)
Diameter	11

Figure 2 - Marvel Comics bipartite network metrics. 'Characters/issue' and 'Issues/character' calculated using Excel. All other metrics produced using Gephi.

Some metrics (e.g. counts of character and issue nodes) may be slightly affected due to the data glitches discussed below. However, since less than 0.1% of the total 19,291 nodes may be mis-classified, the impact does not appear to be significant.

The Bipartite Zone

The Gephi visualization in Figure 3a illustrates the bipartite form of the network. In this view, all the 'issue' nodes appear in white and the 'character' nodes in red. Relative node size is reflective of node degree (white: characters/issue, red: issues/character). Note the much larger number of issue nodes depicted, visually confirming the counts in preceding

data tables.

The Force Atlas 2 layout algorithm positioned the nodes relative to each other based on the number of edges between them. The pale red fog in the background is composed of all the overlapping edges directed from characters to issues. Glimpses of grey background can be seen between the slightly less dense edges in the lower right corner of the image.

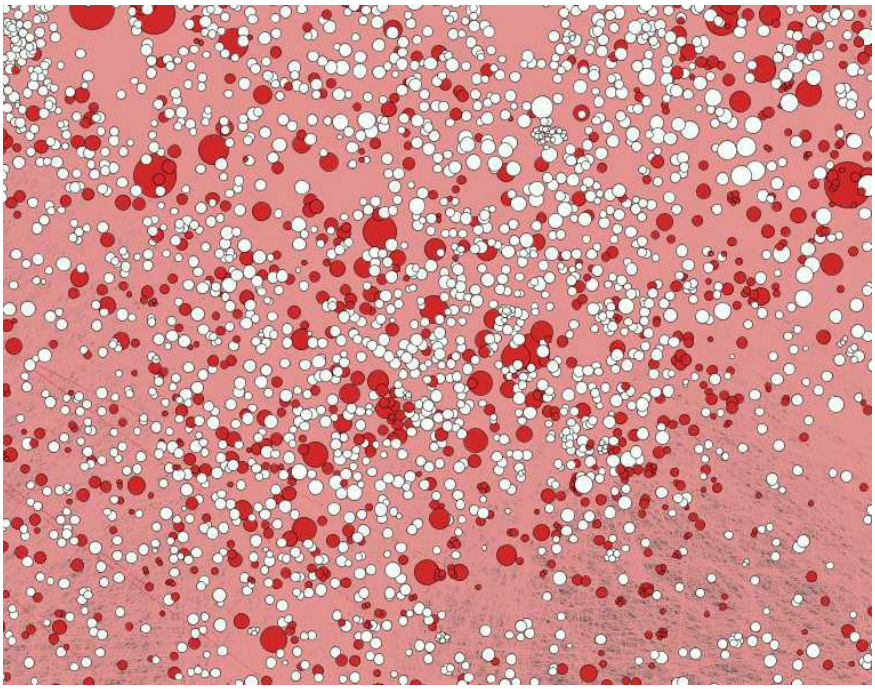

Figure 3a – Visualization of Marvel Comics bipartite network depicting the network's bipartite structure. Issue nodes are white, character nodes are red. Relative node size indicates the number of connections to each node (white: characters/issue, red: issues/character).

Superhero Collaboration Visualized

A second Gephi visualization of the Marvel Comics network bipartite form is shown in Figure 3b. The primary image is a close-up of the denser part of the network for enhanced visual differentiation. The inset is a distant view of the entire network with a red outline indicating the position of the primary view.

The Force Atlas 2 layout produced "limbs" consisting of many overlapping edges which extend radially from center frame at 12:00, 5:00 and 7:00 o'clock positions, connecting the dense center to other more distant groups of character nodes. The darker background shows through the more sparsely populated areas at 9:00, 2:00 and 6:00.

Nodes, representing the characters and issues, are colored and sized based on degree using a medium fillet spline mapping, improving differentiation between low and high degree nodes. This combination was chosen so that low-degree 'Issue' nodes will be recognizable as such by appearing to all be roughly the same color and small size (the maximum issue

node degree is 111). Larger degree nodes representing ‘character’ nodes will stand out due to their contrasting colors and larger size (the maximum character node degree is 1625) .

In the bipartite structure, edges connect characters to the issues in which they collaborate. In Figure 3b, Gephi’s “mixed” edge color option is used, in which edge color is an average of the source and target node colors, resulting in bluer edges in regions where many high degree characters collaborate. However, the overall color of the edge “fog” in Figure 3b is dominated by the reds and yellows of lower degree nodes, which are much more prevalent in the network.

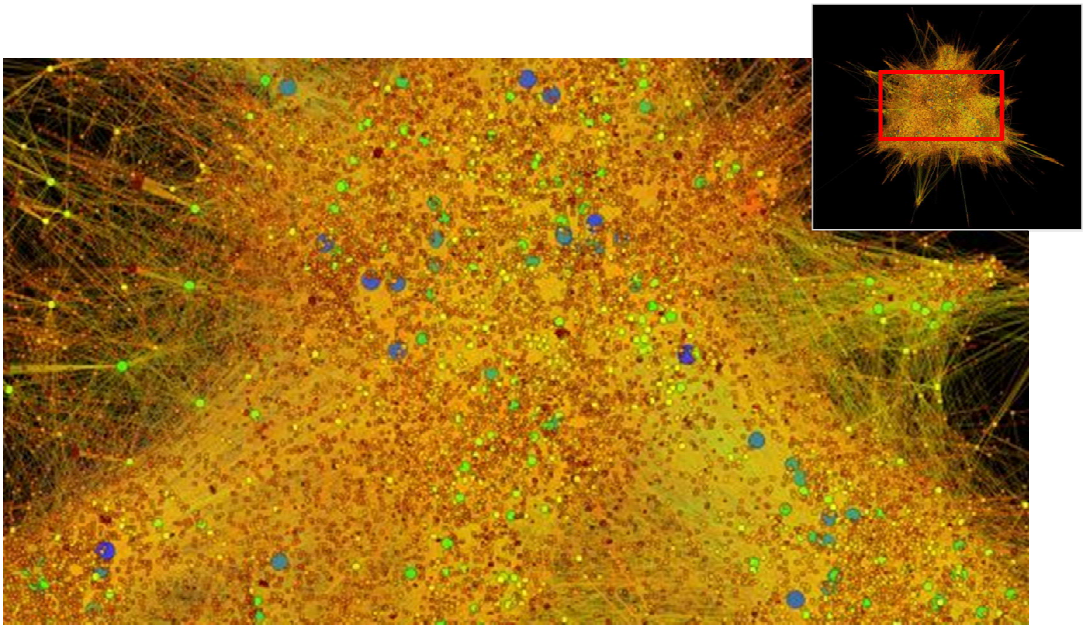

Figure 3b – Visualization of Marvel Comics bipartite network. Larger high degree character nodes stand out in blues and green, while nodes representing lower degree characters and issues nodes are represented by shades of red and orange. Edges are colored by averaging the colors of the nodes they connect.

Consistency of Results

As a starting point for discussion of this effort, metrics from the Alberich study of the Marvel Comics network are compared to those from this study in Figure 4. With the exception of minor variances in node counts and calculations dependent on them, results are very similar (as expected). Some of the the variances are probably explainable by differences in the input datasets used as discussed in the Kryptonite sections below. Those highlighted in yellow appear to have been measured on the projected unipartite network.

	Alberich et al	This Study
Network mode	Bipartite	Bipartite
Network type	Directed	Directed
Nodes N	19,428	19,291
- Character nodes	6,486 (33.4%)	6,467 (33.5%)
- Issue nodes	12,942 (66.6%)	12,824 (66.5%)
Edges m	96,642	96,519
Issues/character	14.9	15
Characters/issue	7.47	7.5
Average degree z	51.88	5.003
Maximum degree	1933	1,625
Avg path length APL	2.63	4.47
Giant Component	6,449 (99.4%)	19,230 (99.7%)
Diameter	5	11

Figure 4 - Comparison of Marvel Comics network metrics from previous study (Alberich et al) to those of current study. Metrics highlighted in yellow appear to reflect a unipartite graph structure.

Kryptonite in the Data

During the loading and analysis of the data, it was noted that the Gephi Multi-mode plugin incorrectly categorized the dataset as not bipartite, although Gephi's statistics and visualization functions were still usable. It was further noted that the Multi-mode plugin's 'node color' function, intended to distinguish between the two types of nodes in the bipartite structure, identified several nodes incorrectly.

Using Excel, the network dataset was also formatted for input to R/iGraph in bipartite form, again requiring identification of the two node types. Using the formatted data as input, R/iGraph's 'create bipartite graph' function also failed.

Alternate sources of the Marvel Comic network data (in various other file types and data formats) were obtained and tested, producing the same results described above, leading to the assumption that these difficulties are due to a few errors in the source data.

Another difficulty encountered was that the Multi-mode plug-in's function to 'project' a unipartite structure from the bipartite structure, consistently aborted reporting a 'memory exceeded' condition. This could either be another consequence of the presumed data errors described above, or signal that this function's computational requirements for this size dataset exceed Gephi's processing capability on a 12Mb gen 2 core i5 processor.

Changing Costumes in a Phone Booth

Analysis of the Marvel Comics network as a unipartite is core to this project. Since functions in Gephi and R/iGraph weren't available, a different method to produce the unipartite network data was needed.

Having the source data was available in .csv format, readable by Excel, permitted an Excel Visual Basic macro to be used to convert the bipartite edges in the original file (Figure 1) to

edges reflecting a unipartite structure (Figure 5). Although Visual Basic usage is relatively rare compared to tools such as Python, it is familiar and capable of delivering the required functionality.

The conversion was accomplished by programmatically writing a new dataset containing an edge between every pair of characters appearing in each issue (i.e. linked to a single issue node in the original file). In the resulting dataset, source and target nodes are all 'character' nodes; no 'issue' nodes exist in the new file. Edge records also includes a type field indicating Undirected, as shown in Figure 4.

Source	Target	Type
IRON MAN/TONY STARK	JARVIS, EDWIN	Undirected
IRON MAN/TONY STARK	QUICKSILVER/PIETRO M	Undirected
IRON MAN/TONY STARK	ROSS, GEN. THADDEUS	Undirected
IRON MAN/TONY STARK	SCARLET WITCH/WANDA	Undirected
IRON MAN/TONY STARK	SENTINELS	Undirected
IRON MAN/TONY STARK	THOR/DR. DONALD BLAK	Undirected
IRON MAN/TONY STARK	VISION	Undirected
IRON MAN/TONY STARK	WONDER MAN/SIMON WIL	Undirected
JARVIS, EDWIN	QUICKSILVER/PIETRO M	Undirected
JARVIS, EDWIN	ROSS, GEN. THADDEUS	Undirected
JARVIS, EDWIN	SCARLET WITCH/WANDA	Undirected

Figure 5- Excerpt of Marvel unipartite network dataset, containing one edge for every pair of characters appearing together in each issue. This repetitive structure is reflected by Edwin Jarvis' appearing as the first Target node in the first (Source: Iron Man) group of edges, then appearing as the Source in the second (Source: Jarvis) group of edges, with characters below Jarvis in the first edge group appearing as Targets of the edges in the second group.

Following final execution of the Visual Basic conversion macro, the resulting Source-Target edge dataset defining the Marvel Comics unipartite network was loaded into Gephi for visualization and analysis. Character-name edge definitions were also converted to numeric edge IDs and loaded into R/iGraph for additional analysis.

Superhero Network Transformation

Unipartite and bipartite network dataset metrics for the Marvel Comics network are listed in Figure 6 for comparison.

As expected, conversion to unipartite mode made a noticeable difference in most network metrics. The number of network nodes was significantly reduced but the number of edges increased by 70%.

The average degree in the network increased tenfold and the maximum degree increased. The node having the highest degree also changed. The average path length and diameter both decreased, and the clustering coefficient increased from zero. The proportion of nodes in the giant component stayed the same.

The average degree of a nodes' neighbors across the network, which was not measured for the bipartite mode, would have been expected to increase also, based on the changes in network diameter, average degree and clustering.

	This Study	
Network mode	Bipartite	Unipartite
Network type	Directed	Undirected
Nodes (N)	19,291	6,426
- Character nodes	6,467 (33.5%)	6,426 (100%)
- Issue nodes	12,824 (66.5%)	--
Edges (m)	96,519	167,207
Issues/character	15	--
Characters/issue	7.5	--
Average degree (z)	5.003	52
Maximum degree	1,625	1,906
Avg path length (APL)	4.47	2.638
Giant Component	19,230 (99.7%)	6,408 (99.72%)
Diameter	11	5
Clustering coefficient (C)	0	0.78
Neighbors' Avg Degree <knn>	--	334.1

Figure 6 - Comparison of Marvel Comics bipartite and unipartite network metrics. Bipartite data replicated from Figure 1. Unipartite metrics were calculated using Gephi and R/iGraph.

Gephi's Superpower

The repetitive edge structure of the unipartite dataset has the effect of almost doubling the number of edges in the dataset. This is due to the fact that in the unipartite data structure produced, an edge is created each time a pair of characters appears together (e.g. collaborates) in an issue. This means that a great number of edges are created for:

(a) characters that Marvel writers "cast" in many issues and titles to increase circulation due to their popularity (e.g. Captain America, Spiderman, etc.)

(b) characters who frequently appear together in teams (e.g. Avengers, Fantastic Four, X-Men, etc.)

This multiplicity of edges in the input dataset provides a serendipitous result: when the dataset is read into Gephi, rather than record multiple edges between the same pair of characters, Gephi apparently uses the existence of multiple edges between characters to accumulate a 'weight' attribute for the edge between that pair of characters. The weight value can be seen in the Data Lab view and is apparently used in calculation of other metrics where appropriate.

More Kryptonite

During the programming and debugging process, additional 'data glitch' items were noted

in the input (bipartite) dataset. There were a number of duplicated character-issue lines buried within the 96,529 edges defined in the bipartite input dataset, as if "placeholder lines" (using real character-issue pairs, not dummy data) were pasted into the bipartite edges list during its creation from the Marvel Chronology Project database and never overwritten or removed.

These extra lines, when processed by the Visual Basic program, produced additional character-character edges in proportion to the number of placeholder lines encountered. When read into Gephi, these spurious edges result in very high weights (generated by Gephi's superpower) for the associated character-character edges.

Although several groups of these placeholder lines were caught and eliminated while testing the Visual Basic program, at least one group escaped notice. It can be seen as the highest weighted edge in the unipartite dataset in Gephi Data Lab view when sorting on the Weight attribute. Luckily, it can be easily removed using Gephi's built-in tools.

When working with such large datasets, it continues to be good practice to be alert for concealed data artifacts that may impact results.

Marvel Superheroes vs. Mortal Men

The Marvel Comics unipartite network was compared to an Erdos-Renyi G(n,m) random graph, generated with the number of nodes and edges of the Marvel Comics network. Results are shown in Figure 7.

	E-R Random Graph	Unipartite
Network type	Undirected	Undirected
Nodes (N)	6,426	6,426
Edges (m)	167,207	167,207
Average degree (z)	52	52
Avg path length (APL)	2.643	2.638
Giant Component	6,426 (100%)	6,408 (99.72%)
Diameter	4	5
Clustering Coeff	.0081	.1945
Avg Clustering Coeff (C)	.0081	.7811
Maximum degree	79	1,906
Neighbors' Avg Degree <knn>	53	334.1

Figure 7 - Comparison of Marvel Comics unipartite network metrics to those of an Erdos-Renyi random graph. Unipartite data replicated from Figure 4. Random graph metrics calculated using Gephi and R/iGraph. Items highlighted in yellow show significant variance between the two graph types.

As anticipated, the values of metrics relating to the graphs' gross construction parameters are very similar. However, in other areas (highlighted in yellow) the metrics reflect significant differences in internal structure between the E-R random graph and the Marvel Comics network.

Power Law Distribution

The degree distribution in the network was examined using R/iGraph, resulting in the plot shown in Figure 8. The distribution portrayed appears to indicate a straight-line power law signature for degree larger than approximately ten.

The PLFIT program, which attempts to fit the dataset to the power law equation below, was used to confirm this.

$$p(x) \sim x^{-alpha} \quad \text{for } x >= xmin$$

The PLFIT program produced an alpha value of 2.39 for values larger than 110, with goodness-of-fit well below tolerance.

This calculated value of alpha is within the 2-3 range generally associated with social network power law degree distributions. It is within the range for scientific collaboration networks (see Figure 11) observed by Newman in the SIAM paper, and is very close to the alpha of 2.3 for the film actors collaboration network also reported by Newman in the PNAS paper listed above.

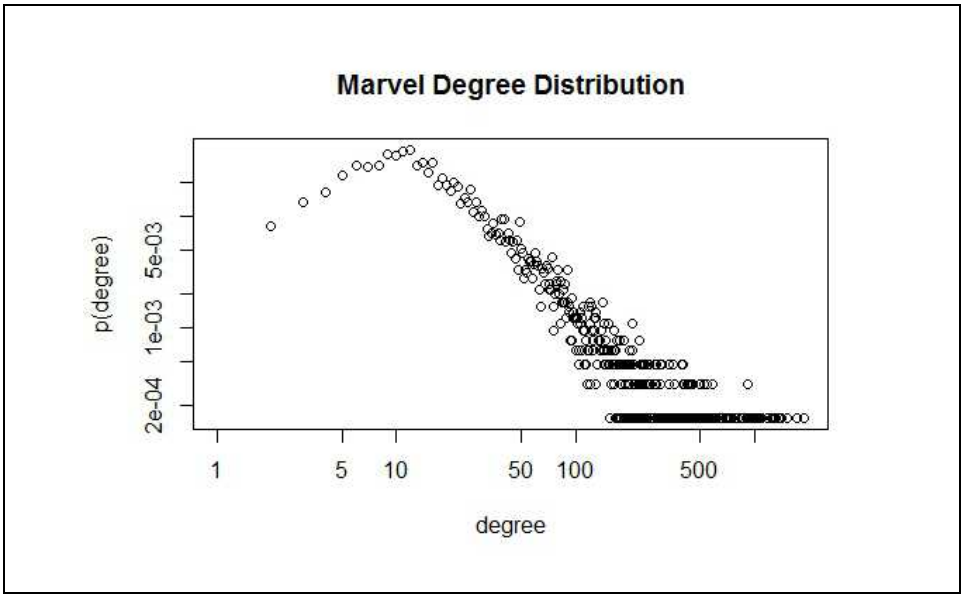

Figure 8 – Degree distribution of Marvel Comics unipartite network.

It was noted above, when discussing edge proliferation during mode conversion, that some characters appear repeatedly in many different issues. Gleiser uses this in his analysis of the Marvel Comics network degree distrbution by plotting Newman's proposed measure for the strength of ties between characters w against p(w), resulting in a very clean power law distribution plot with alpha of 2.26.

Neighbor Degree Correlation

In many real-world social networks, nodes of high degree tend to be linked to other nodes of high degree and those of low degree with other low degree nodes (assortative mixing). To determine whether this is the case in the Marvel Comics network, the degrees of all network nodes' neighbors knn were calculated using R/iGraph and plotted in Figure 9.

Although knn values vary greatly for nodes of degree less than approximately 200, beyond this point the distribution appears to converge into a line with negative slope. This indicates that in the Marvel Comics network, nodes of increasing degree tend to be linked to nodes of decreasing degree, indicating the opposite of the expected assortivity (disassortivity).

Gleiser calculated that the slope for that part of the curve has a power law exponent of 0.52, noting that a well-known technological network (the Internet) has a very similar value of 0.5 for this same characteristic.

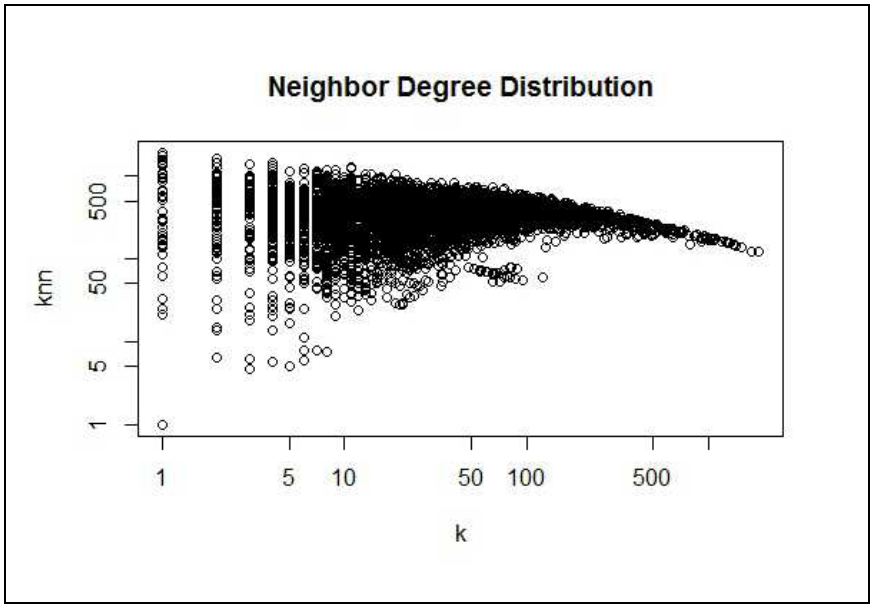

Figure 9 – Degree of a node's neighbors plotted against the degree of the node in the Marvel Comics network. The negative slope visible at k>200 indicates disassortivity.

Collaboration Networks Compared

Newman's study of scientific collaboration networks reviewed six archives. He noted that one of those archives, the SPIRES database, was atypical due to its subject area. Figure 10 below presents a comparison of the Marvel Comics network metrics and the ranges reported by Newman for other scientific collaboration networks excluding SPIRES.

	Marvel Comic Network	5 Scientific Collaboration Networks - Newman
Character/author nodes	6,467	8,360 <-> 1,520,200
Issue/paper nodes	12,824	13,170 <-> 2,163,900
Node ratio (% chars/issues)	50%	63% <-> 70%
Issues/character	15	2.55 <-> 6.4
Characters/issue	7.5	1.99 <-> 3.75
Collaborators (z)	5.0	3.59 <-> 18.1
Mean distance (APL)	2.64	4.6 <-> 9.7
Max distance (diameter)	5	14 <-> 31
Giant Component	6408	5,835 <-> 1,395,700
Giant Component %	99.7%	57.2% <-> 92.6%
Clustering coefficient (C)	0.78	0.066 <-> 0.496
Power law distrib exponent	2.39	0.91 <-> 2.5

Figure 10 - Comparison of Marvel Comics network metrics to ranges of values reported by Newman for scientific collaboration networks studied. Metrics highlighted in yellow reflect the unitpartite graph structure.

In terms of size, the Marvel Comics network lies on the low end of the node count range of those studied by Newman (19,000 nodes vs. 21,000 nodes), as well as the relative proportion of author nodes to paper nodes (50% vs. 63%).

The character/issue and issues/character ratios in the Marvel Comics network are almost double the high end of the ranges for these values observed by Newman. However, the average number of collaborators (i.e. characters/issue) for the SPIRES database (which was identified as atypical of collaboration networks) was reported to be nine authors per paper. Newman comments on this:

> *"The reason for this last impressive figure is that the SPIRES database contains data on experimental as well as theoretical work. High-energy experimental collaborations can run to hundreds or thousands of people, the largest author list in the SPIRES database giving the names of a remarkable 1,681 authors on a single paper."*

The SPIRES number is more in line with the 7.5 average observed in the Marvel Comics network. However, since the largest number of characters in a single Marvel issue is 111 (vs. the SPIRES 1,681), further similarities between the Marvel Comics network and SPIRES seem unlikely.

Although the average number of collaborators (degree) and giant component size measured in the Marvel Comics network are within Newman's reported ranges, the mean distance (APL) and maximum distance (diameter) are well below Newman's minimum values. In addition, the clustering coefficient of the Marvel Comics network is almost double the upper end of Newman's range. Deeper investigation of the similarities and differences between the Marvel Comics network and other real-world collaboration networks could be a direction for further study.

Both Alberich and Gleiser discuss ways in which the Marvel Comics network differs from a real-world social network. It was conjectured that, rather than growing out of real collaboration events and other social situations which connect individuals, Marvel

character relationships are created and reused by the writers based on the commercial popularity of certain characters with their target audience. Another interesting direction for further study might be to investigate this idea, comparing the circulation figures for Marvel titles (which are available online) to the areas of divergence from ‘real-world’ social networks.

Superhero Network Visualization

In contrast to the Marvel Comics bipartite network view, in which character-character interactions are defined by the collaboration events (the ‘issues’ nodes) present in the network itself, the unipartite network links characters direcly to other characters. The breadth of a character’s interactions is reflected by the degree of the character’s node, and the extent to which those interactions recur within a consistent community or team (e.g. a subset of the network) is reflected by the character’s node’s modularity.

In Figure 11, which provides an overview of the entire range of character-character interactions (e.g. the unipartite network), nodes are colored to indicate membership in one of the character groups that consistently appear together (e.g. teams of characters, reflected by Gephi’s modularity class metric) and sized by their degree. A higher degree indicates interaction with a larger number of other characters.

Relative node positioning reflects the Force Atlas layout algorithm with repulsion strength adjusted to 500 (almost double the default value), driving more separation between recurring teams. This results in placing those characters (i.e. nodes) which consistently appear together but seldom interact outside their community, in a peripheral or “orbital” position. On the other hand, characters and character groups (e.g. teams) which interact frequently with other characters appear much closer to the center of the sphere. This layout structure results in “orbital” groups of unique colors, reflecting that they form a modular community having few interactions with other groups (disconnected components). One completely isolated group can be observed at the 3:00 o’clock position in Figure 11.

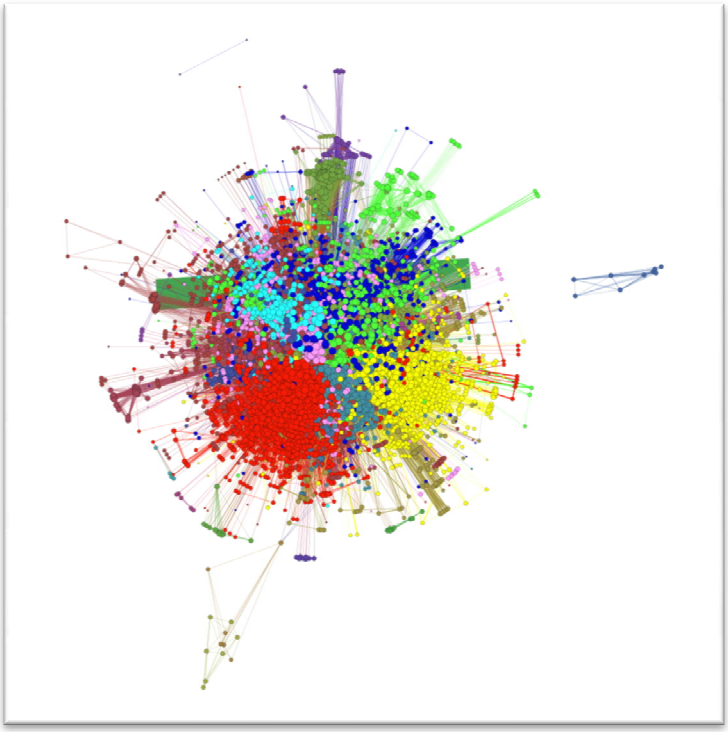

Figure 11 – Overview of the Marvel Comics unipartite network. Colors indicate in which modular sub-group (i.e. community or team) the node is a member. Node size reflects degree, the number of other characters interacted with. Position on the radial axis indicates the level of interaction with other characters or groups of characters.

A close-up view of a subset of the Marvel Comics network is shown in Figure 12, in which filtering is used to focus on highly active characters in the network. Display parameters are unchanged from Figure 11: node color indicates membership in a group (modularity class) and node size is scaled by level of interactions (degree). Node position was minimally adjusted manually to improve caption readability.

Almost the entire network seen in Figure 11 has been eliminated from view by filtering out nodes with degrees of less than 1000 (maximum in the network is 1906). Only those characters that interact maximally frequently with other characters -- in this case, each other – remain visible. Character teams are segregated by color, with high frequency of team interaction indicated by the heavier edges between team members.

The popular Marvel characters Dr. Strange, Hulk, Thor and Spiderman are also shown, their node colors indicating membership in other communities of characters not represented in the figure, although high levels of interaction with the red, purple and yellow teams is indicated by the edges between them.

Gradually lowering the display filter's degree cutoff value would reveal broader but less frequent, character associations. Alternatively, by resetting the filter to single out a

specific community (i.e. a single modularity class value) the entire "web" of relationships around Spiderman could be examined. In the paper referenced above, Gleiser discusses performing a similar analysis using the Watts & Strogatz clustering coefficient.

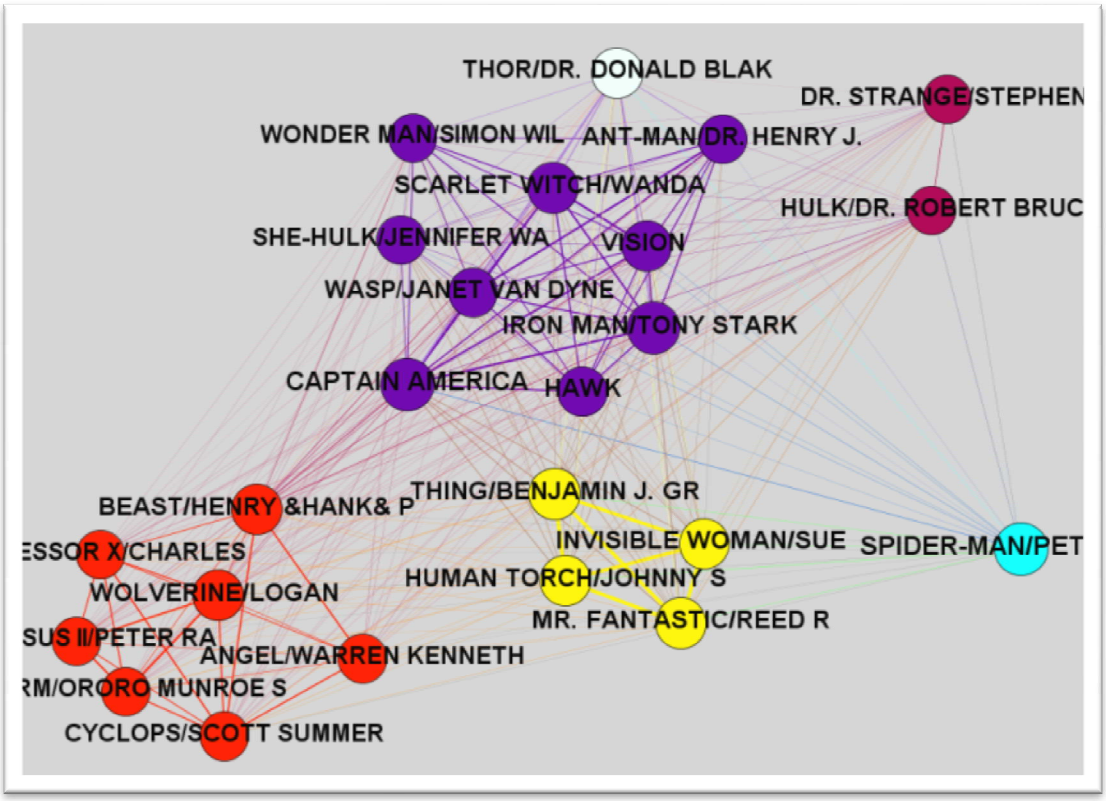

Figure 12 – "Narrow focus" view of high degree nodes (characters) in the Marvel Comics unipartite network. Colors represent community affiliations (red = X-Men, purple = Avengers, yellow = Fantastic Four) reflecting Gephi's modularity class.

Conclusion

With much guidance from the prior work of Alberich, Gleiser and Newman, the work performed in this study confirms that the Marvel Comics network displays some attributes of real-world collaboration networks, but is at the same time different in important ways.

Much work has already been done on the Marvel Comics network, and it was interesting and instructive to explore and replicate it, and to apply newly-learned tools while doing so. It was also rewarding to discover the depth and extent to which M.E.J. Newman has contributed to documenting and advancing our knowledge of networks.

Ideas not covered in SMA class (prior to the timing of this project) were uncovered and/or explored, including degree correlation and assortivity, details of projecting a bipartite network, and the breadth of available information documenting a broad range of network classes.

Some areas within this project’s scope for possible follow-up were also identified.

========== Supporting Info ==========

Marvel Unipartite Network R Script

```
>
#########################################################################
#########
>
> library(igraph)
> setwd("C:\\Users\\Rick\\Desktop\\R Analysis")
> edgelist <- read.table("Unipartite Numeric Edge List Weighted.txt",
header=TRUE)
> graph <- graph.data.frame(edgelist, directed=FALSE)
>
> summary(graph)
IGRAPH UN-- 6426 167207 --
attr: name (v/c), Weight (e/n)
>
> nodes = vcount(graph)
> edges = ecount(graph)
> deg = degree(graph)
> maxdeg = max(degree(graph))
> avgdeg = mean(degree(graph))
> diam = diameter(graph)
> apl=average.path.length(graph)
> gblcc=transitivity(graph, type="global")
> avgcc=transitivity(graph, type="average")
> lc <- largest.cliques(graph)
> cln <- clique.number(graph)
> knnres <- graph.knn(graph)
> knnv <- knnres[[1]]
>
> nodes
[1] 6426
> edges
[1] 167207
> maxdeg
[1] 1906
> avgdeg
[1] 52.04077
> diam
[1] 5
> apl
[1] 2.638427
> gblcc
[1] 0.1945397
> avgcc
[1] 0.7810964
> cln
[1] 111
> mean(knnv)
[1] 334.1283
> table(clusters(graph)$csize)

   2    7    9 6408
   1    1    1    1
>
> data <- degree.distribution(graph)
> plot(data, log="xy", xlab="degree", ylab="p(degree)",main = "Degree
Distribution")
Warning message:
In xy.coords(x, y, xlabel, ylabel, log) :
```

```
  1496 y values <= 0 omitted from logarithmic plot
>
> data <- degree.distribution(graph, cumulative=TRUE)
> plot(data, log="xy", xlab="degree", ylab="p(degree)",main =
"Cumulative Degree Distribution")
>
> data <- data.frame(x=knnv, y=degree(graph))
> plot(x~y, data=data, log = "xy", xlab="k", ylab="knn",main =
"Neighbor Degree Distribution")
>
> E(graph)$weight <- seq(ecount(graph))
> data <- data.frame(x=graph.strength(graph,
mode="all"),y=degree(graph))
> plot(x~y, data=data, log = "xy", xlab="k", ylab="Weight",main =
"Weighted Degree Distribution")
>
> a = plfit(degree(graph))
> a
$xmin
[1] 110

$alpha
[1] 2.39

$D
[1] 0.03688578

>
> a = plfit(knnv)
> a
$xmin
[1] 686.6667

$alpha
[1] 5.868632

$D
[1] 0.0249429

>
> a=plfit(graph.strength(graph, mode="all"))
Error: cannot allocate vector of size 1.1 Gb
> a
$xmin
[1] 686.6667

$alpha
[1] 5.868632

$D
[1] 0.0249429
```

E- Random Graph Network R Script

```
> ######################################################################
> ######################################################################
>
> library(igraph)
> setwd("C:\\Users\\Rick\\Desktop")
> graph <- erdos.renyi.game(6426, 167207, type="gnm", directed = FALSE,
loops = FALSE)
```

```
>
>
> summary(graph)
IGRAPH U--- 6426 167207 -- Erdos renyi (gnm) graph
attr: name (g/c), type (g/c), loops (g/x), m (g/n)
>
> nodes = vcount(graph)
> edges = ecount(graph)
> deg = degree(graph)
> maxdeg = max(degree(graph))
> avgdeg = mean(degree(graph))
> diam = diameter(graph)
> apl=average.path.length(graph)
> gblcc=transitivity(graph, type="global")
> avgcc=transitivity(graph, type="average")
> lc <- largest.cliques(graph)
> cln <- clique.number(graph)
> knnres <- graph.knn(graph)
> knnv <- knnres[[1]]
>
> nodes
[1] 6426
> edges
[1] 167207
> maxdeg
[1] 82
> avgdeg
[1] 52.04077
> diam
[1] 4
> apl
[1] 2.642751
> gblcc
[1] 0.008075722
> avgcc
[1] 0.008065105
> cln
[1] 4
> mean(knnv)
[1] 53.02101
> table(clusters(graph)$csize)

6426
   1
>
> data <- degree.distribution(graph)
> plot(data, log="xy", xlab="degree", ylab="p(degree)",main = "MU
Degree Distribution")
Warning message:
In xy.coords(x, y, xlabel, ylabel, log) :
  32 y values <= 0 omitted from logarithmic plot
>
> data <- data.frame(y=knnv, x=degree(graph))
> plot(x~y, data=data, log = "xy", xlab="k", ylab="knn",main =
"Neighbor Degree Distribution")
>
> a = plfit(degree(graph))
> a
$xmin
[1] 38

$alpha
[1] 3.5
```

```
$D
[1] 0.2163889

>
>
> data <- degree.distribution(graph)
> plot(data, log="xy", xlab="degree", ylab="p(degree)",main = "MU
Degree Distribution")
Warning message:
In xy.coords(x, y, xlabel, ylabel, log) :
  32 y values <= 0 omitted from logarithmic plot
>
> data <- data.frame(y=knnv, x=degree(graph))
> plot(x~y, data=data, log = "xy", xlab="k", ylab="knn",main =
"Neighbor Degree Distribution")
>
> a = plfit(degree(graph))
> a
$xmin
[1] 38

$alpha
[1] 3.5

$D
[1] 0.2163889
```

Bipartite to Unipartite Visual Basic Script

```
Option Base 1
Dim In_Index, Out_Index, In_Issue As Double
Dim This_Issue, This_Char As Variant
'

Sub Build_Links_No_Circs()

    Application.Calculation = xlCalculationManual

    Call Clear_Edges

    In_Index = 1
    Out_Index = 1
    This_Char = Worksheets("Marvel Bipartite").Range("A1").Offset(In_Index, 0).Value

Do Until This_Char = ""
    In_Issue = In_Index
    This_Issue = Worksheets("Marvel Bipartite").Range("A1").Offset(In_Issue, 1).Value
    This_Char = Worksheets("Marvel Bipartite").Range("A1").Offset(In_Issue, 0).Value

    Do Until (Worksheets("Marvel Bipartite").Range("A1").Offset(In_Issue + 1, 1).Value <> This_Issue _
            Or Worksheets("Marvel Bipartite").Range("A1").Offset(In_Issue + 1, 1).Value = "")
        If This_Char = Worksheets("Marvel Bipartite").Range("A1").Offset(In_Issue + 1, 0).Value Then
GoTo Skip_Circ
'         -- Write Source Node
        Worksheets("Marvel Unipartite").Range("A1").Offset(Out_Index, 0).Value = This_Char
'         -- Write Target Node
        Worksheets("Marvel Unipartite").Range("A1").Offset(Out_Index, 1).Value = _
            Worksheets("Marvel Bipartite").Range("A1").Offset(In_Issue + 1, 0).Value
'         -- Write Link Type
        Worksheets("Marvel Unipartite").Range("A1").Offset(Out_Index, 2).Value = "Undirected"
'         -- Write Issue Generating Link as Reference
```

```
        Worksheets("Marvel Unipartite").Range("A1").Offset(Out_Index, 3).Value = This_Issue

        Out_Index = Out_Index + 1
Skip_Circ:
        In_Issue = In_Issue + 1
        Loop

    In_Index = In_Index + 1
    This_Issue = Worksheets("Marvel Bipartite").Range("A1").Offset(In_Issue, 1).Value
    This_Char = Worksheets("Marvel Bipartite").Range("A1").Offset(In_Issue, 0).Value

    Loop

    Application.Calculation = xlCalculationAutomatic

End Sub


Sub Clear_Edges()

'   Ensure Autofilters Set to "All"

'    Worksheets("Marvel Unipartite").Range("A1:C1").AutoFilter field:=1
'    Worksheets("Marvel Unipartite").Range("A1:C1").AutoFilter field:=2
'    Worksheets("Marvel Unipartite").Range("A1:C1").AutoFilter field:=3

    Worksheets("Marvel Unipartite").Range("A2:z680000").ClearContents

End Sub
```