Comic Characters Data Mining

Intermediate Report

Huy Tran
u1228479
u1228479@umail.utah.edu

John Duffy
u0783287
u0783287@umail.utah.edu

Ethan Olpin
u1018382
u1018382@umail.utah.edu

Jeff Gay
u1098246
u1098246@umail.utah.edu

1. SIMILARITY

We want to look at how creative the makers were in the last century in terms of visual appearances for characters. Thus we apply the Jaccard similarity on two sets of characters. We split the data into two periods, let's say before 1975 and after 1975. We can see from the similarity report that most of the results range from 0.7-0.9. This is because we are trying to see if DC or Marvel have made characters with new visual styles, i.e. new hair colors, new eye colors. Obviously, there are new styles created, but there should not be too many differences.

Note for characters' identity and align, whether a character has secret/public identity, good/bad/neutral alignment. The similarities are all 1.0 , which is reasonable. The 0.72 similarity we can see from ID and Align in DC characters at the 1975 split is because some of the old characters were not specifically labeled. We can see how characters after 1975 have detailed labels in ID and Align. This is believed to be '**label noise**'. As an optional feature, we can either replace unknown labels with the most common one.

Universe	Period	Attributes	Jaccard Similarity
Marvel	before 75 vs. 75-95	Eye, Hair, Sex	0.825
Marvel	before 75 vs. 75-95	Identity, Align	1.0
Marvel	90-2000 vs after 2000	Eye, Hair, Sex	0.89473
Marvel	90-2000 vs after 2000	Identity, Align	1.0
DC	before 75 vs. 75-95	Eye, Hair, Sex	0.86206
DC	before 75 vs. 75-95	Identity, Align	0.72727
DC	90-2000 vs after 2000	Eye, Hair, Sex	0.89285
DC	90-2000 vs after 2000	Identity, Align	1.0

Table 1: DC & Marvel Characters Appearance Similarity

2. COLLISION COUNTER

We used the backing logic of Coupon Collector and birthday paradox to count the number of collisions. We split the data between four different periods for two data sets, before 1975 vs 1975-1995 and 1990-2000 and after 2000. From here we also created subsets for two different combinations of data parameters, 1. eyes, hair, and sex, 2. identity and alignment.

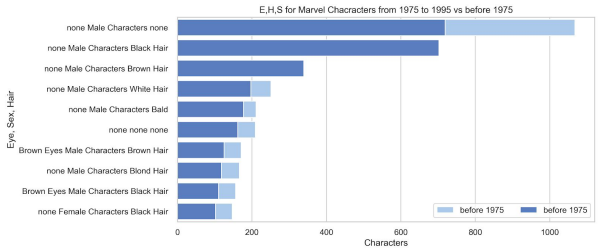

Figure 2.1: Marvel's eye, hair and sex before 1975 vs after 1975

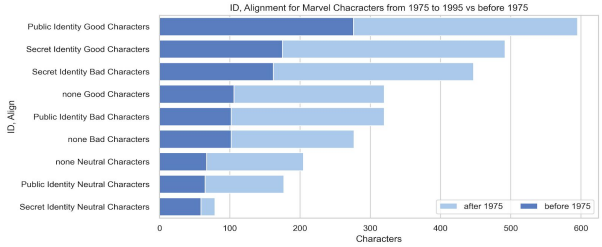

Figure 2.2: Marvel's Identity and Alignment before 1975 vs after 1975

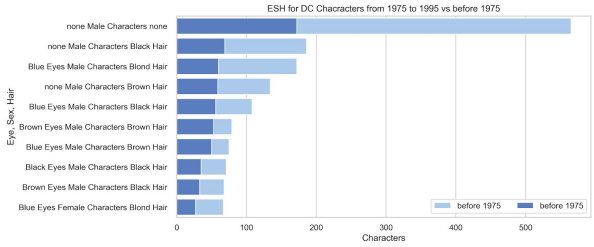

Figure 2.3: DC's Identity and Alignment before 1975 vs after 1975

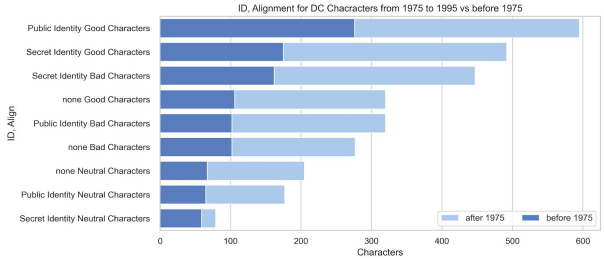

Figure 2.4: DC's eye, hair and sext before 1975 vs after 1975

NOTE: Within the subsets there's a substantial amount of unknown data reported as none. When we checked some data points where it was collected from we found that some unknown features are listed. Due to these none values existing there tends to

be an abnormality of collisions within some labels. As well there are many labels that have one data point, creating outliers.

3. CLUSTERING

We clustered the data using all 15 pairwise combinations of the various ability metrics in the character_stats.csv file. Some data cleaning was necessary, we determined that 71% of the data was usable leaving us with 432 data points. We used k=4 clusters for our data, using Gonzalez to find our initial centers, and then used Lloyd's to refine those centers. We also observed the average distance between all of our clusters reporting the distance between two clusters X, Y as: $D(X, Y) = min_{x \in X, y \in Y} d(x, y)$

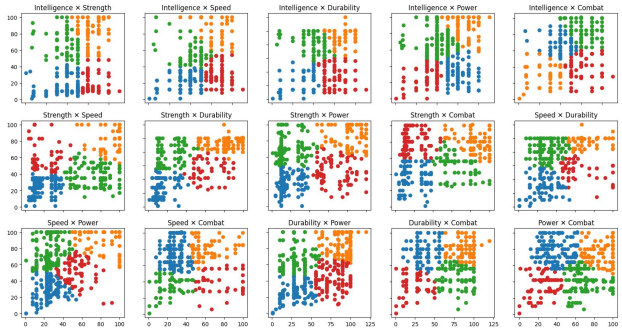

Figure 3.1: Pairwise Metric Combinations Subjected to Clustering

An interesting thing to note, clusters generally had similar centers between power value combinations. This could suggest that characters with higher total values would generally fall in the same cluster, regardless of the combination used. We observe that the maximum distance between two clusters was 12.58, indicating a low degree of separation between clusters, suggesting that our clusters may fail to reveal any meaningful information about the underlying data. We also considered clustering using all six dimensions of the power values, still finding a low degree of separation between our clusters:

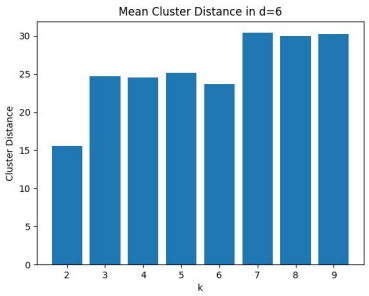

Figure 3.2: Mean Cluster Distance Using 6-Dimensional Points

Metric 1	Metric 2	Cluster Distance
Intelligence	Strength	9.035
Intelligence	Speed	7.455
Intelligence	Durability	8.801
Intelligence	Power	9.41
Intelligence	Combat	9.217
Strength	Speed	12.58
Strength	Durability	9.981
Strength	Combat	8.98
Strength	Power	8.531
Speed	Durability	9.309
Speed	Power	9.084
Speed	Combat	6.567
Durability	Power	8.495
Durability	Combat	8.092
Power	Combat	8.843

Table 2: Mean Cluster Distance for Pairwise Power Combinations

4. NEXT STEPS

From our similarity and collision counter,(similarity in eye hair sex and in identity align) we can pick a random year, let's say 1950. Starting from this year, increment by one year and see how long it took to have the character with the exact same visual appearance (same eye colors, hair, sex etc.) .This is the birthday paradox on real data. Then compare it with the analytical form.
Similarly, we start from say 1980, look at all the eye colors, hair colors with sex available from this period. We want to ask how long did it take Marvel/DC to create characters and match all the visual appearances we have from the 1980. As always, we plan to compare it with the analytical form. This is the coupon collector experiment.

For clustering, we may consider clustering based off of three stat tuples instead of pairs, as this could also show stronger relations between the various power values we have. More clusters could also allow for a more granular look at the data, and could suggest various "classes" of heroes based on where they happen to be clustered. Given the low degree of separation between clusters, we may choose to prioritize techniques other than clustering to reveal more about our dataset.

After we learn more and get a better understanding of the topic of regression, we'll run regression algorithms to produce a predictive model. The regression model will be trained from the average number of appearances of top 10 characters from different time periods before 1990. While the testing data will be average appearances of the top 10 from the 2000s and so on.

Once the topic has been covered in class, we intend to perform network analysis to explore the relationships between different characters' allies, acquaintances and antagonists. We also expect to learn more about the tropes and characteristics of comic crossovers.