Network Analysis

Nirmal Kanagasabai
School of Computer Science
McGill University
Email: nirmal.kanagasabai@mail.mcgill.ca

***Resumo*—The purpose of this exercise is to analyze relationships that are found in the Whos-Dated-Who dataset and make significant inferences about the community. Data collection from WhosDatedWho.com website and the collected data is cleansed. Later, NetworkX, a Python package is used for creation, manipulation, and study of structure and dynamics of the network under consideration. An algorithm is defined to compute overlapping daters and inferences are made by comparing the count of overlaps to the total dataset that was used. From the results, it could be inferred that the percentage of overlapped daters in the dataset of celebrities considered for network analysis (celebrities are actively involved in dating), was found to be 50.86%.**

I. Introduction

To analyze the relationships in the Whos-Dated-Who website, the following steps were carried out:

1) Data Collection and Cleansing
2) Network Analysis using NetworkX and Gephi
3) Inference based on the count of overlapping daters

II. Data Collection and Cleansing

To begin with, the dataset is extracted from Whos-Dated-Who website using Dynamic Scraping (for scraping the URLs of all celebrities in the infinite scrolling website and Static Scraping (for scraping the relationships of one celebrity).

The dataset that was extracted previously was made use of [1]. However, through further analysis, it could be inferred that there were few irregularities in the dataset that I originally had. So, the steps were repeated again (this time, carefully avoiding the irregularities).

As the actual analysis revolves around evaluating the abundance of overlapping daters, all those celebrities who didn't have any relationships in the past or those celebrities for whom there is no record on the WhosDatedWho.com website is removed from the actual dataset.

III. Network Analysis using NetworkX and Gephi

To carry out the Network Analysis on the dataset that was considered, NetworkX [2], a Python package that is used for creation, manipulation and study the structure and dynamics of the networks and Gephi [3], an open source Graph Visualization and Exploration tool were made use of.

The dataset which was prepared had two columns: 'Celebrities' and their corresponding 'Partners'. This was nothing but an edge-list with two nodes (one - Celebrity, other - Partner). The dataset from the .csv file was read as a Pandas dataframe. A directed graph was created and the edges were added from the edge-list. NetworkX, by default, creates the two nodes on either side of the edge (if it is not already created). In total, the graph had 54,315 nodes and 72,365 edges.

A. Overlapping Daters - Definition

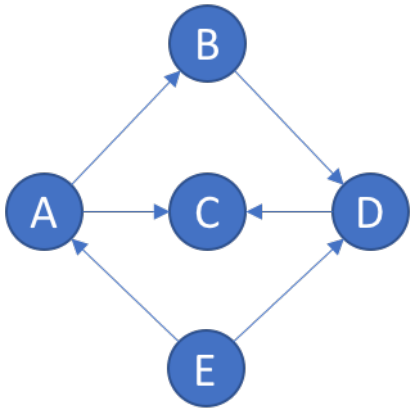
An algorithm was defined to find the overlapping daters. As per definition, two daters, X and Y, are overlapping if they share more than one dating partner in common. For example, if X has dated individuals A, B and C, and Y has dated A, B, D, E, there are two common partners (A and B). Hence, they the celebrities 'X' and 'Y' are overlapping daters.

B. Common Out Neighbors and Overlapping Daters

The graph under consideration was 'Directed'. The focus was on how many 'celebrities' are overlapping daters. Hence, I wanted to find the 'out degree', i.e., the number of celebrities dating their partners. The key consideration here is that 'the partners may or may not be celebrities themselves'. In the example graph below [Fig. 1], the common out neighbors are:

Celebrity 1	Celebrity 2	Common Partner
B	E	set([D])
E	B	set([D])
A	D	set([C])
D	A	set([C])

Figura 1: Example - Directed Graph



To achieve this, the NetworkX's successors() and intersection() functions were used. That is, for every Graph with two nodes 'i' and 'j', the function common_out_neighbors() return a set which is an intersection of the successors of 'i' and the successors of 'j'.

The returned set consists of the partners that are common for both the celebrities. If the size of this set is greater than or equal to '2', a set of those two celebrities is prepared. The overlapping celebrities along with their common dating partners are written to a .csv file. The pseudocode can be found in Figure 2.

Figura 2: Pseudocode - Overlapping Daters

```
common_out_neighbors(Graph, Celebrity1, Celebrity2):
        celeb1 = Graph.successors(Celebrity1)
        celeb2 = Graph.successors(Celebrity2)
        return set(celeb1.intersection(celeb2))

overlapping_daters(G)
        For n in G.nodes():
                For m in G.nodes():
                        if (n != m):
                                overLaps = common_out_neighbors(G, n, m)
                                if len(overLaps) >= 2:
                                overLappingCelebrities = set([n,m])
```

IV. Results and Discussion

The following parameters (Degree Distribution and Modularity) of the Whos-Dated-Who network were computed using Gephi.

A. Degree Distribution

The average degree (degree distribution) of the network was calculated. It depicts as to how interconnected is each node in the given network. It can be observed from Figure 3 that the average degree is 1.332.

B. Modularity

The modularity of the network was also calculated. It is one of the best ways to visualize communities and is widely used as a measure of how good clustering is. It clearly illustrates as to how many more connections a section of a graph has compared to a random graph. I made use of 'Randomizing' property which produces a better decomposition but increases the computation time. The modularity of the Whos-Dated-Who network is 0.892 and the number of communities is found to be 10,752 which is plotted as the y-axis and the size (number of nodes) was plotted in the y-axis. This can be observed in Figure 4.

Figura 3: Degree-Distribution

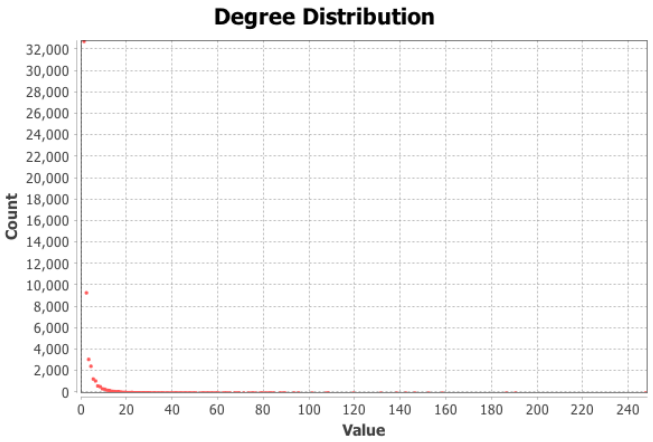
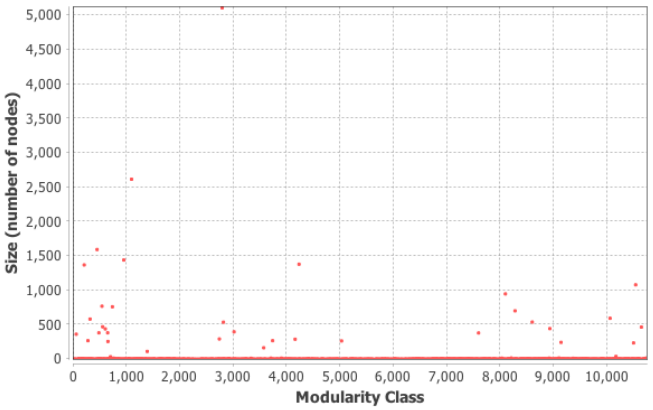


Figura 4: Communities Size Distribution



C. Overlapping Daters

Out of 53,302 celebrities that were extracted from Whos-Dated-Who.com, 27,281 celebrities didn't date anyone or their records of past relationships weren't found in the website. Hence, they were neglected and the count of unique celebrities that were included in the Network Analysis study is 26,021.

Based on our algorithm, we retrieved a .csv file containing 23,685 entries of over-lapping daters. This list was surprisingly high as only 26,021 celebrities were used for the study. Upon close analysis, it could be found that the partners of the celebrities were celebrities themselves thereby leading to duplicate values in the overlapped daters list. For example, 'A' dated 'X' and 'Y'. Likewise, 'B' also dated 'X' and 'Y'. In this case, as per the rule, 'A' and 'B' are overlapped daters. However, 'X' and 'Y' are also celebrities themselves. This way, we will have 4 entries (2 each for the pairs (A,B) and (X,Y)). The duplicates were removed and we had 13,235 overlapped daters.

Entity	Count
Total count of Celebrities	53302
Count of Celebrities who doesn't date anyone	27281
Count of unique celebrities in NxDataset	26021
Count of unique Overlapping Daters	13235

If we are to calculate the percentage of over-lapped daters in the dataset we considered for Network analysis, we obtained 13,235 daters out of 26,021 unique celebrities. This accounts for 50.86% of the celebrities that were studied.

Likewise, if we compare the 13,235 overlapped daters against the total count of celebrities (before removing the celebrities with '0' date relationships - 53302 celebrities), we obtain 24.83% which is still a larger percentage.

From the overlapped-daters list, one very interesting observation is the number of partners that were commonly dated by two celebrities. The maximum number of partners dated by two celebrities (maximum number of nodes that were commonly shared by two other nodes) is 43. Celebrities, 'Julian','Rocco Siffredi', who are typically porn stars were involved in multiple encounters with similar partners which made this list stand out from the rest. The next significant number of common partners dated by two celebrities who are American actresses (Ava Gardner and Lana Turner) is 25.

V. Conclusion

With the dataset we obtained, it can be inferred that overlapping daters have been a constant phenomenon among celebrities who are actively involved in dating. The number of quadrilateral relationships (celebrities having 2 or more partners in common) present in the graph is higher (which clearly depicts the number of overlapping daters).

Referências

[1] Nirmal Kanagasabai, Derek Ruths, "Scraping Static and Dynamic Websites", Social Media Analytics, McGill University.

[2] NetworkX: Python Package(https://networkx.github.io/)

[3] Gephi: Open Graph Visualization Platform (https://gephi.org/)