



Debunking Hollywood:

A Network Productions Inquiry

By: Jeff Bailey, Hassan Koroma, Harry Choi, and Peter Schnizer

Subtask 1: Data Extraction Methods

Created two scripts,
“dataset_generator” and
“credit_compressor”

- **dataset_generator** loops through the director dataframe and generates a full credits dictionary, which is added to a jsonl.
- **credit_compressor** compressed the inputted jsonl into a gzip.

```
(base) PS C:\Users\hujej\desktop\hw5-project> python dataset_generator.py

Initiating Jeff's IMDB scraping process. For each director inputted in the
initial csv, this script will collect credits with all crew members for e
ach director's feature films. If you would like to filter out specific rol
es or jobs, input a list of strings of roles you would like to include.

Enter path for director information csv: 100_film_directors.csv

Enter name and path of JSONL file: scraped_credits.jsonl

Enter a comma-separated list of valid roles for directors: ['']

we have already collected full credits for the following directors:
{'nm0004716', 'nm0000487', 'nm0160840', 'nm0000876', 'nm0000709', 'nm00275
72', 'nm0668247', 'nm0001068', 'nm0005069', 'nm1950086', 'nm0001331', 'nm0
000500', 'nm0905152', 'nm0001392', 'nm0036349', 'nm0000165', 'nm0476201',
'nm0001054', 'nm1443502', 'nm1490123', 'nm0009190', 'nm0000600', 'nm000039
9', 'nm1119645', 'nm0000631', 'nm0002132', 'nm0000233', 'nm0001081', 'nm00
00217', 'nm1218281', 'nm0501435', 'nm0619762', 'nm0751102', 'nm0037708',
'nm0336620', 'nm0000464', 'nm0169806', 'nm0269463', 'nm0751577', 'nm0893659
', 'nm0000517', 'nm0898288', 'nm0392237', 'nm0946734', 'nm0420941', 'nm000
0229', 'nm0000941', 'nm0000386', 'nm0634240', 'nm0000338', 'nm0000490', 'n
m0853380', 'nm2011696', 'nm0911061', 'nm0327944', 'nm0138927', 'nm1148550',
'nm0000343', 'nm0590122', 'nm1560977', 'nm0001060', 'nm0001814', 'nm0000
116', 'nm0000759', 'nm0000231', 'nm0298807', 'nm0000186', 'nm0000777', 'nm
0570912', 'nm0336695', 'nm0366004', 'nm0000318', 'nm0868219', 'nm0716980',
'nm0426059', 'nm3363032', 'nm0000361', 'nm0905154', 'nm1716636', 'nm03625
66', 'nm1883257', 'nm0001741', 'nm2125482', 'nm0281945', 'nm1503675', 'nm0
200005', 'nm0583600', 'nm0001752', 'nm0000881', 'nm0122344', 'nm0796117',
'nm0510912', 'nm0001005', 'nm0190859', 'nm0000142', 'nm0001631', 'nm000052
0', 'nm0697656', 'nm1347153', 'nm1802161', 'nm0000095'}
100%|████████████████████████████████████████████████████████████████████████████████| 101/101 [00:00<00:00, 2525
0.33it/s]
(base) PS C:\Users\hujej\desktop\hw5-project> python credit_compressor.py

Enter path for collected jsonl dataset: scraped_credits.jsonl

Enter desired path for the zipped dataset (end it with .gz): compressed_cr
edits.gz
(base) PS C:\Users\hujej\desktop\hw5-project> _
```

Subtask 1: Data Extraction Output

Final dataset was a .jsonl list of json-style “director dictionaries”.

This was originally 20 mb, compressed to a 4 mb gzip file.

While filtering crew roles was possible in this code, we collected the entire credits for each movie and then filtered out redundant roles at the network generation step.

```
{
  "dir_id": "nm0009190",
  "name": "J.J. Abrams",
  "gender": "M",
  "ethnicity": "W",
  "otherlabel": "H",
  "movies": [
    {
      "title_id": "tt2527338",
      "title": "Star Wars: Episode IX - The Rise of Skywalker",
      "crew": [
        ["Writing Credits", "Chris Terrio"],
        ["Writing Credits", "Derek Connolly"],
      ]
    },
  ]
}
```


Subtask 2: Network Generation

Pipeline:

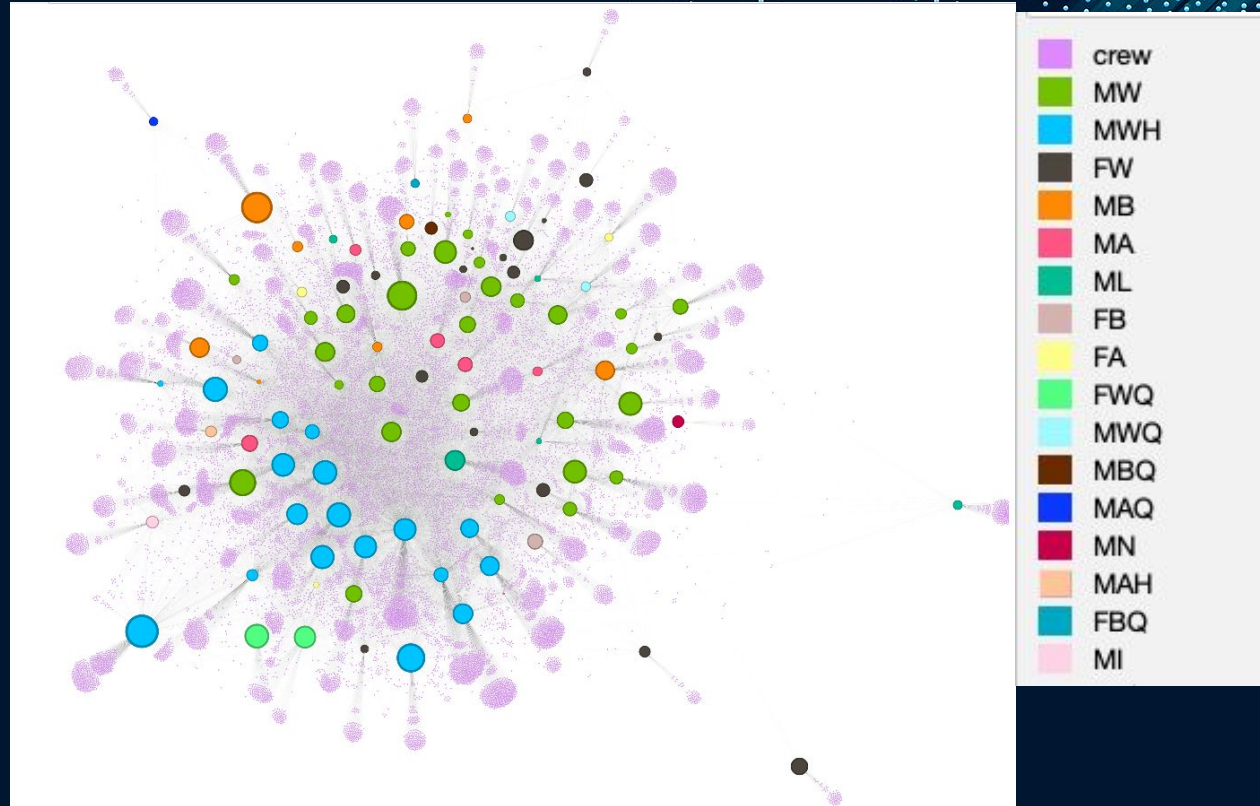
- Preprocessing
 - Load data as a generator
 - Normalize roles
 - Remove extra white spaces
- Node & Edge Attributes
 - Node type, indiv. & grouped labels, AHS
 - Roles, abs. weight, custom crew weight
- Network Models: weighted undirected
 - Full: all director-crew collaborations
 - Main: with role exclusion
 - Bipartite Subnetwork: co-occurrence projections
- Generate Network
 - Create and write network to a .graphml file for visualization in Gephi

```
{
  "dir_id": "nm0009190",
  "name": "J.J. Abrams",
  "gender": "M",
  "ethnicity": "W",
  "otherlabel": "H",
  "movies": [
    {
      "title_id": "tt2527338",
      "title": "Star Wars: Episode IX – The Rise of Skywalker",
      "crew": [
        ["Writing Credits", "Chris Terrio"],
        ["Writing Credits", "Derek Connolly"],
      ]
    },
  ]
}
```

Subtask 3: Network Visualization

Background:

- Layout: ForceAtlas 2
- Size of Nodes: Average Director Homogeneity
- Color of Nodes:
 - Grouped Labels
 - First Letter: Gender
 - Second: Ethnicity
 - Third: Renowned Status



Subtask 4: Analyses

Process:

- Raw Data Analysis
 - Basic summary statistics
- Metric engineering
 - Role homogeneity
 - Weights
- Answering research questions
 - Network characteristics
 - Important nodes
 - Analyzing metric

$$S = \begin{cases} 0 & ; P = 1 \\ \frac{N-1}{P-1} & ; P > 1 \end{cases}$$

N = # times a crew member was re-used in their role

P = # opportunities to be re-used in their role

Relationship Strength (Weights)

$$H = \begin{cases} 1 & ; u = 1 \\ 1 - \frac{u}{n} & ; u > 1 \end{cases}$$

u = unique crew members for role across movies

n = total crew members for role across movies

Role Homogeneity

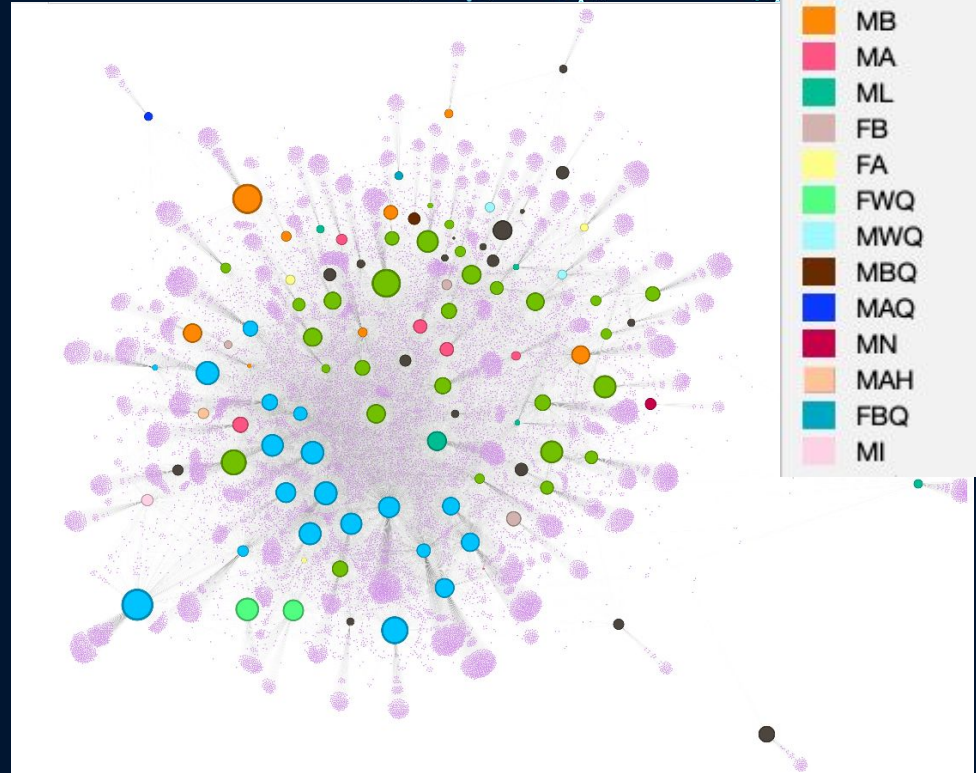
Research Question 2

**Network Characterization &
Properties**



Network Characterization & Properties

- Density: 0.00018
- Triangles: 5339
- Mean Clustering Coeff: 0.00076
- Triangles may not tell the whole story
 - No crew-crew edges



Research Question 3

Interesting Nodes/Links



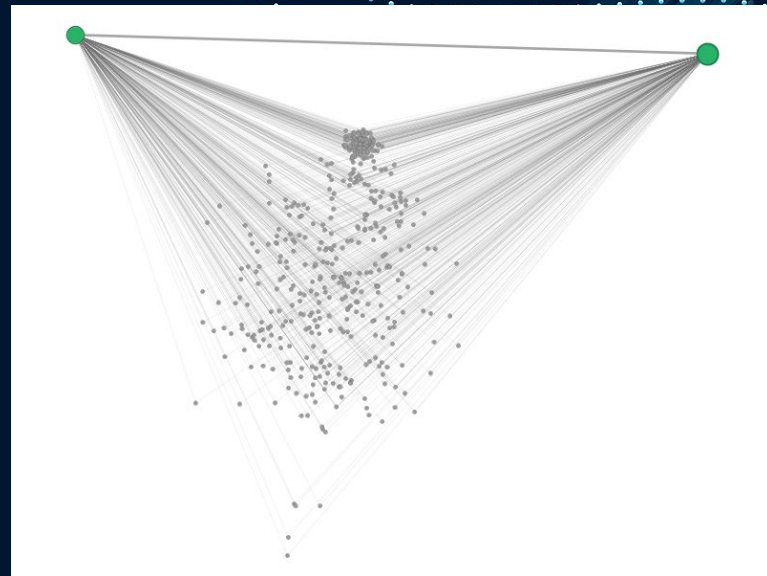
Wachowski Ego

Lilly Wachowski:

- Homogeneity: 0.457 (Rank 8)
- Triangles: 464
- Top 3 Roles: Directed by, Writing Credits, Music by

Lana Wachowski:

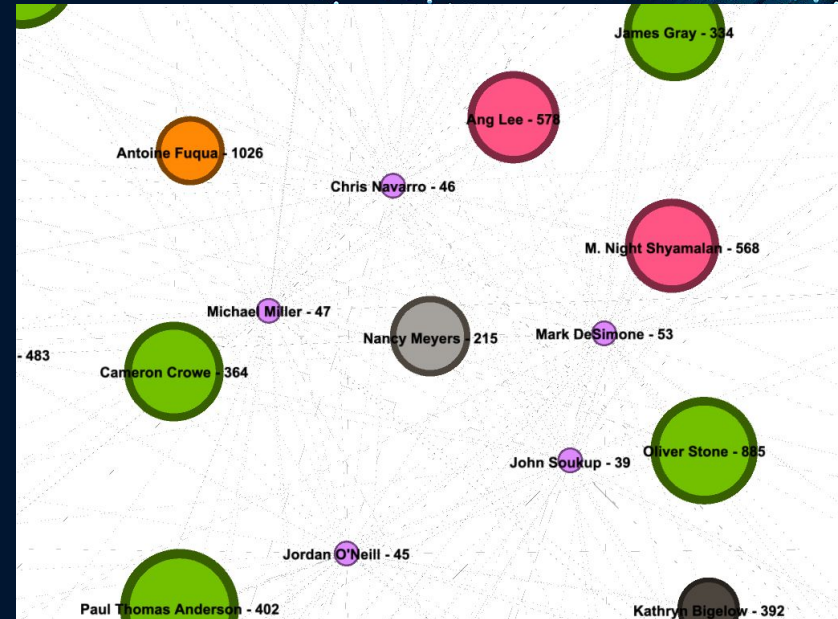
- Homogeneity: 0.415 (Rank 19)
- Triangles: 464
- Top 3 Roles: Directed by, Writing Credits, Costume Design by



Top 5 Highest Degree (Crew)

Rank	Name	Degree	% of director colab
1	Mark DeSimone	53	52%
2	Michael Miller	47	47%
3	Chris Navarro	46	46%
4	Jordan O'Neill	45	45%
5	John Soukup	39	37%

- All sound department guys
- All worked with Tim Burton, Roland Emmerich
- Sound department is not very homogeneous (Ranked 8/10)



Hubs and Interesting Nodes

Top 5 Betweenness Centrality:

1. Steven Spielberg: 53,574
2. Tim Burton: 47,527
3. Ridley Scott: 45,772
4. Ron Howard: 43,388
5. Roland Emmerich: 31,138

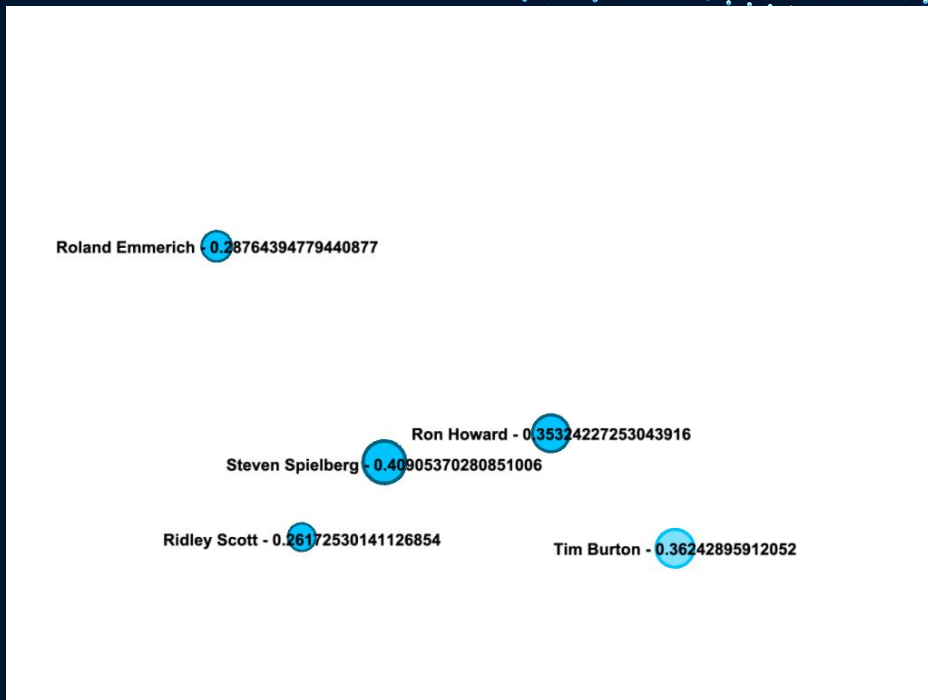
Interesting Scores:

90. Jordan Peele: 3,501
93. Chloé Zhao: 3,249

Mean Betweenness Centrality: 12,475

Top 5 Director Degree:

1. Steven Spielberg: 2,030
2. Tim Burton: 1,702
3. Ridley Scott: 1,643
4. Ron Howard: 1,605
5. Roland Emmerich: 1,195



Research Question 1

**How widespread is the
phenomenon of directors
re-using the same crew?**



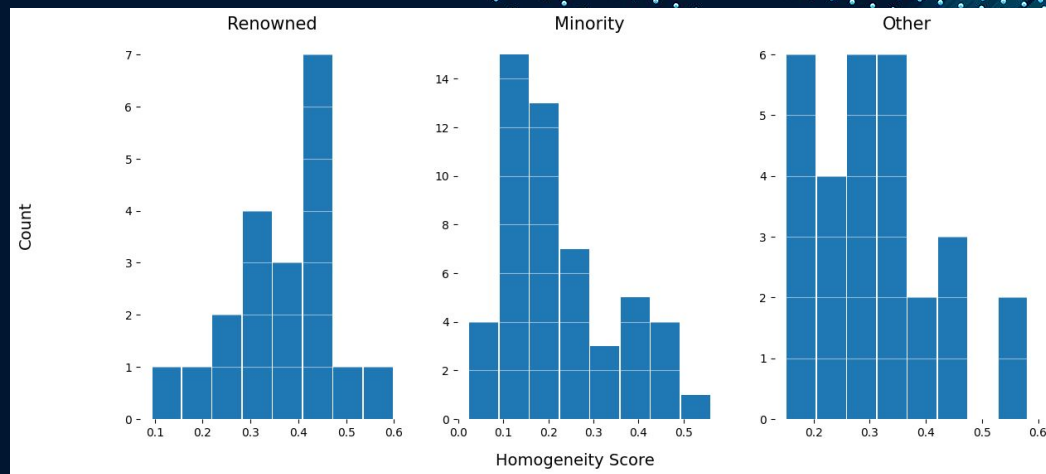
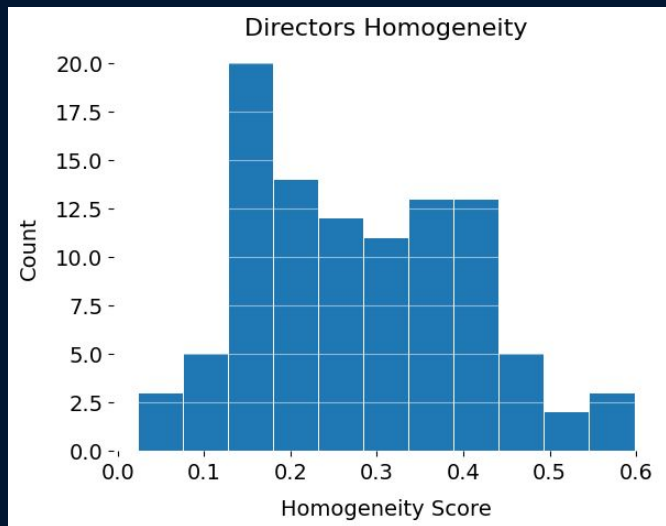
Director Rankings

Ranking of Directors by Homogeneity Score

.
. .
.

Rank	Name	AHS	Top 3 Roles
1.	Peter Jackson	0.598319772586482	Writing Credits - Film Editing by - Cinematography by
2.	Clint Eastwood	0.5817356536996472	Directed by - Film Editing by - Cinematography by
3.	Tyler Perry	0.5595145606597748	Cinematography by - Film Editing by - Production Design by
4.	Steven Soderberg	0.53781649411341	Directed by - Cinematography by - Film Editing by
5.	David Yates	0.5124667078329002	Film Editing by - Writing Credits - Music by

Research Question 1:



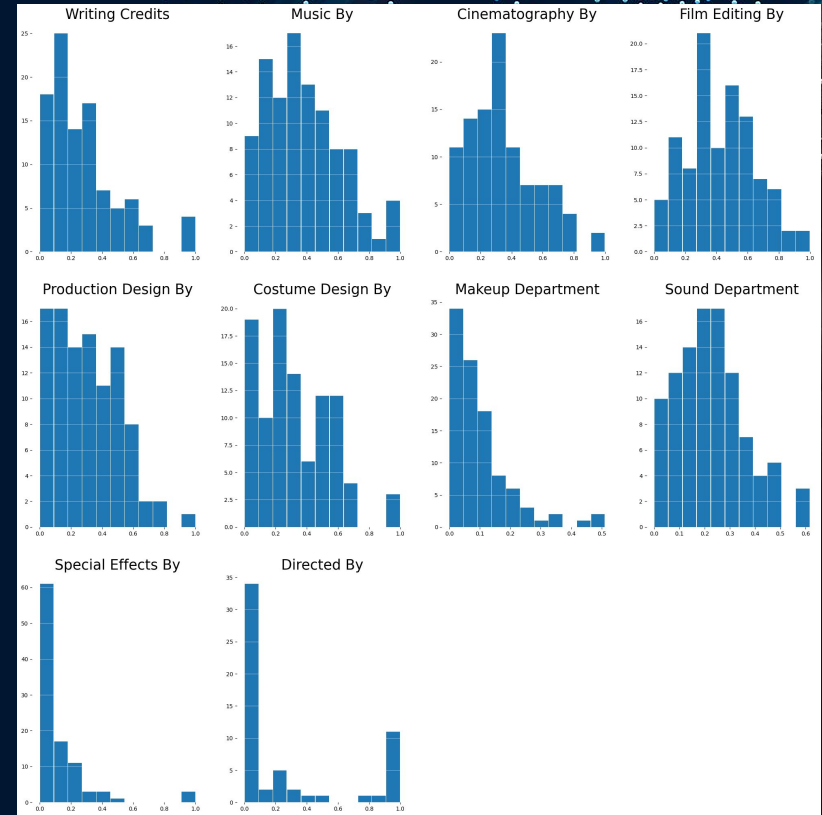
Avg Homogeneity By Role:

- Renowned: 0.36
- Minority: 0.23
- Other: 0.31

Research Question 1:

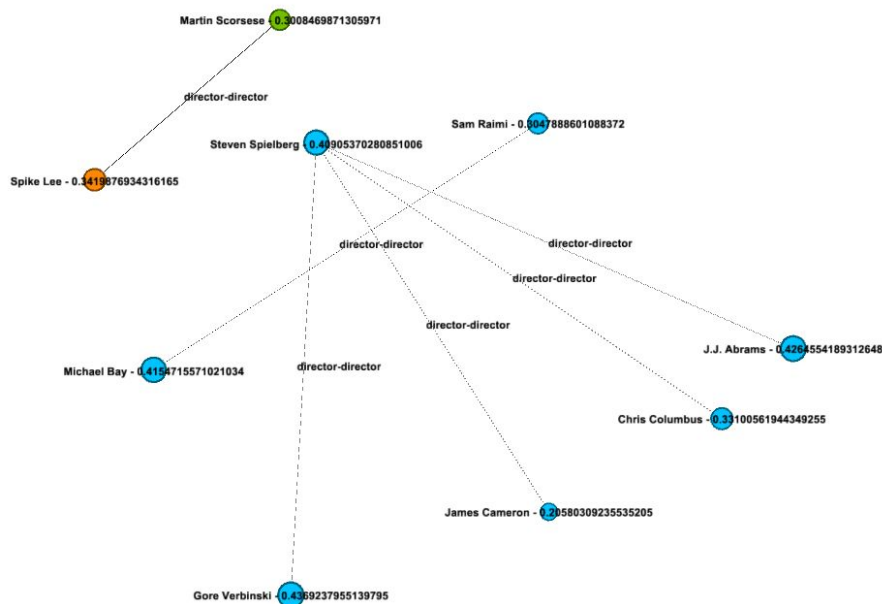
Mean Homogeneity Scores:

1. Film Editing: 0.43
2. Music: 0.38
3. Cinematography: 0.34
4. Costume Design: 0.31
5. Production Design: 0.30
6. Writing Credits: 0.28
7. Directed By: 0.27
8. Sound: 0.23
9. Special Effects: 0.12
10. Makeup: 0.10



Research Question 1:

- Co-Occurrence Network
- Filtered to pairs of directors who have hired 1000 or more of the same crew members
- 6/7 pairings: **Male, White, and Renowned**
- Takeaway: Male, White, and Renowned directors not only re-use the same crew members, but they hire from the same pool of people as each other.



Final Conclusion

- Renowned directors (mostly white men) have the most shared crew and trended with higher crew reuse.
- Most other directors seemed do not have strong networks of crew reuse. (See the many small clouds of crew who only worked on a single movie on our visualization)
- Most minority and less-renowned directors are unable to cultivate strong relationships or crew reuse networks for themselves, and must work with unfamiliar and one-time crew members.
- Lead editors, musicians, and cinematographers are the most commonly reused roles by those directors with strong connections.

THANKS!

Do you have any questions?
networkproducts@ISC..com
+91 620 421 838
NetworkProductions.com



CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.

Please keep this slide for attribution.