Visualizing a Network of Film Actors With Box Office Mojo

Christopher Redino

January 26, 2016

Outline

- Goals
- Data
- Code
- Results
- Improvements

Goals

- Goals
- Data
- Code
- Results
- Improvements

- See how connected/segregated the community of film actors are (in terms of who co-stars with whom) by considering different ways of grouping actors.
- For now this is just a qualitative first attempt, but quantifiable results can come later.
- May be of utility to actors, see which studios/genres are more “clicky”, inform career decisions. See who may be linked to whom more quickly with an interactive map.

Data

- Goals
- Data
- Code
- Results
- Improvements

- Data set was scrapped from boxofficemojo.com.
- The scrapper and a sample csv can be found at https://github.com/csredino/Data-Science.

Data

- Goals
- Data
- Code
- Results
- Improvements

- As scrapped data has one row for each movie page on Box Office mojo.

movies

Filter

	title	director1	director2	domestic	distributor	release	genre
1	$9.99	N/A	N/A	$56,900	Regent Releasing	December 12, 2008	Animation
2	$upercapitalist	N/A	N/A	$16,600	Truly Indie	August 10, 2012	Thriller
3	(500) Days of Summer	Marc Webb	N/A	$35,233,100	Fox Searchlight	July 17, 2009	Romance
4	(Untitled)	N/A	N/A	$246,100	Samuel Goldwyn	October 23, 2009	Comedy
5	...And Justice for All	N/A	N/A	$107,727,500	Columbia	October 19, 1979	Drama
6	1+1=11	N/A	N/A	$178,400	Eros	March 28, 2003	Unknown
7	1,000 Times Good Night	N/A	N/A	$54,200	Film Movement	October 24, 2014	Foreign
8	10	N/A	N/A	$242,194,400	Warner Bros.	October 5, 1979	Romantic Come
9	10 Items or Less	Brad Silberling	N/A	$103,300	ThinkFilm	December 1, 2006	Comedy
10	10 Questions for the Dalai Lama	N/A	N/A	$265,000	Monterey Media, Inc.	April 27, 2007	Documentary
11	10 Things I Hate About You	N/A	N/A	$61,024,900	Buena Vista	March 31, 1999	Romantic Come
12	10 to Midnight	N/A	N/A	$18,497,100	MGM	March 11, 1983	Unknown
13	10 Years	Jamie Linden	N/A	$210,000	Anchor Bay Films	September 14, 2012	Comedy / Dram
14	10,000 B.C.	Roland Emmerich	N/A	$107,193,300	Warner Bros.	March 7, 2008	Period Adventur
15	100 Bloody Acres	N/A	N/A	$6,200	Doppelganger Releasing	June 28, 2013	Horror Comedy
16	The 100-Year Old Man Who Climbed Out the Window...	N/A	N/A	$815,800	Music Box Films	May 1, 2015	Foreign
17	101 Dalmatians	N/A	N/A	$456,744,300	Disney	January 25, 1961	Animation
18	101 Dalmatians (1996)	N/A	N/A	$247,236,700	Buena Vista	November 27, 1996	Family Comedy
19	101 Dalmatians (Re-issue) (1969)	N/A	N/A	$102,929,600	Disney	January 1, 1969	Animation
20	101 Dalmatians (Re-issue) (1979)	N/A	N/A	$61,466,100	Disney	January 1, 1979	Animation
21	101 Dalmatians (Re-issue) (1985)	N/A	N/A	$72,862,200	Buena Vista	December 20, 1985	Animation
22	101 Dalmatians (Re-issue) (1991)	N/A	N/A	$117,325,900	Buena Vista	July 12, 1991	Animation
23	101 Reykjavik	N/A	N/A	$176,700	Menemsha	July 27, 2001	Unknown
24	102 Dalmatians	N/A	N/A	$100,001,200	Buena Vista	November 22, 2000	Family Comedy
25	10th & amp; Wolf	N/A	N/A	$67,800	ThinkFilm	August 18, 2006	Crime Drama

Code

- Goals
- Data
- Code
- Results
- Improvements

- Want to generate network plots showing how film makers are connected through their collaborations (think Kevin Bacon).
- Several packages can do this but they all expect the data in a certain form.
- Most of the work is transforming/cleaning the data
- To make a network map requires two data frames, one for “nodes” and one for “links”

Code

- Goals
- Data
- Code
- Results
- Improvements

- Node frame is at its minimum just a list of all the actors.
- All other columns are their properties that can be used for grouping/slicing later, to categorize subsets of the community and see which subsets actually divide them in terms of collaborations.

Filter

	name	group	studio	genre	budget_avg	box_office_avg	id
1	Aaron Eckhart	1	Warner Bros.	Drama	47775000	58373804	0
2	Aaron Johnson	1	Universal	Drama	47928571	47053800	1
3	Abigail Breslin	1	Buena Vista	Drama	43586364	65781463	2
4	Adam Brody	1	Warner Bros.	Comedy	35062500	42980963	3
5	Adam DeVine	1	Universal	Comedy	17000000	124381850	4
6	Adam Sandler	1	Sony / Columbia	Comedy	60411111	93136417	5
7	Adam Scott	1	Paramount	Comedy	35863636	30985117	6
8	Alan Alda	1	Universal	Comedy	55000000	47734832	7
9	Albert Brooks	1	None	Comedy	31925000	83142240	8
10	Alfred Molina	1	Miramax	Drama	60880952	57568535	9
11	Alicia Silverstone	1	Warner Bros.	Comedy	45000000	43891254	10

Code

- Goals
- Data
- Code
- Results
- Improvements

- Link frame has one row for each connection.
- This is a bit harder to produce, especially if you want to do it quickly.
- Scrapper can save up to 18 contributors for a movie, for ~13,000 movies this is over 2 million possible connections.
- Vectorizing the loop over movies is the most important, but the rest can be put in nested loops without much cost.

Filter

	src	trgt	weight
1	0	2	1
2	0	3	1
3	0	12	1
4	0	17	1
5	0	26	1
6	0	30	1
7	0	31	1
8	0	58	1
9	0	74	1
10	0	92	1
11	0	93	1
12	0	112	1
13	0	117	1
14	0	119	1
15	0	120	2
16	0	133	1
17	0	148	1
18	0	164	1
19	0	180	1
20	0	191	1

Results

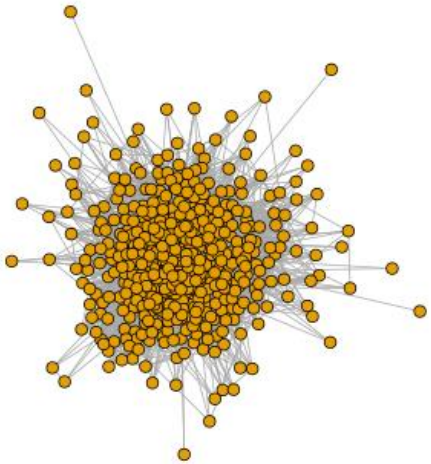
- Goals
- Data
- Code
- Results
- Improvements

- R script can transform the data as output by the scrapper.
- Can look at actors, directors, writers, producers, and composers.
- Can categorize by genre preference, studio preference, average budget of works, and average domestic box office of works.
- For the sake of time, I'll only show results for actors and only categorized by studio and genre preference.

Results

- Goals
- Data
- Code
- Results
- Improvements

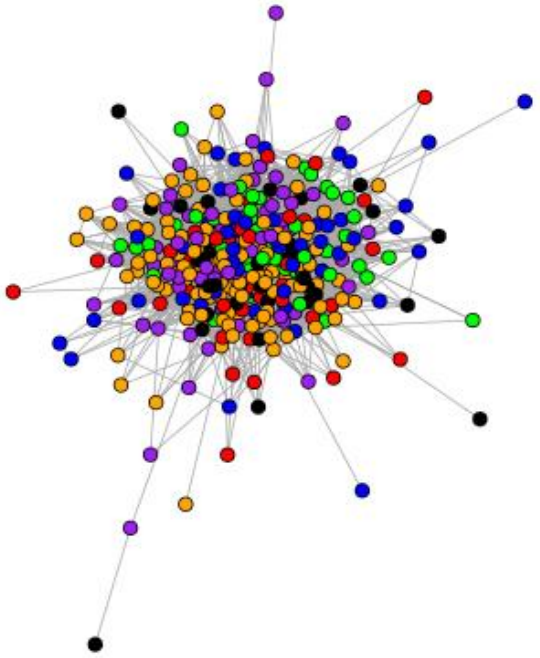
- Looking at all actors at once with no categories just gives a “hairball”.
- Even with incomplete records removed, its still hundreds of actors with thousands of connections.



Results

- Goals
- Data
- Code
- Results
- Improvements

- If we color by category “studio”, we start to see some bunching, but its hard to discern patterns with so many nodes

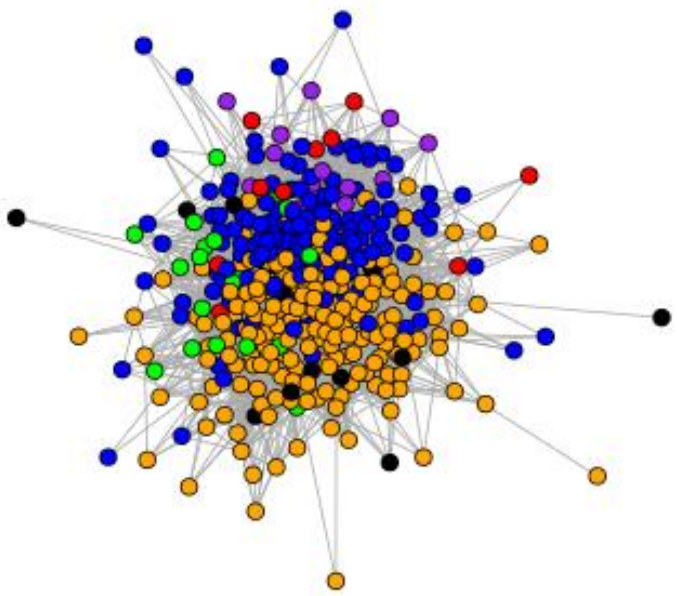


```
"Warner Bros."] = "orange"
"Fox"] = "blue"
"Paramount"] = "black"
"Sony/Colombia"] = "green"
"Beuna Vista"] = "red"
"Universal"] = "purple"
```

Results

- Goals
- Data
- Code
- Results
- Improvements

- Coloring by genre has a much more noticeable effect.
- Comedy and drama actors are apparently pretty segregated.
- Other genres seem preferential in their connections.

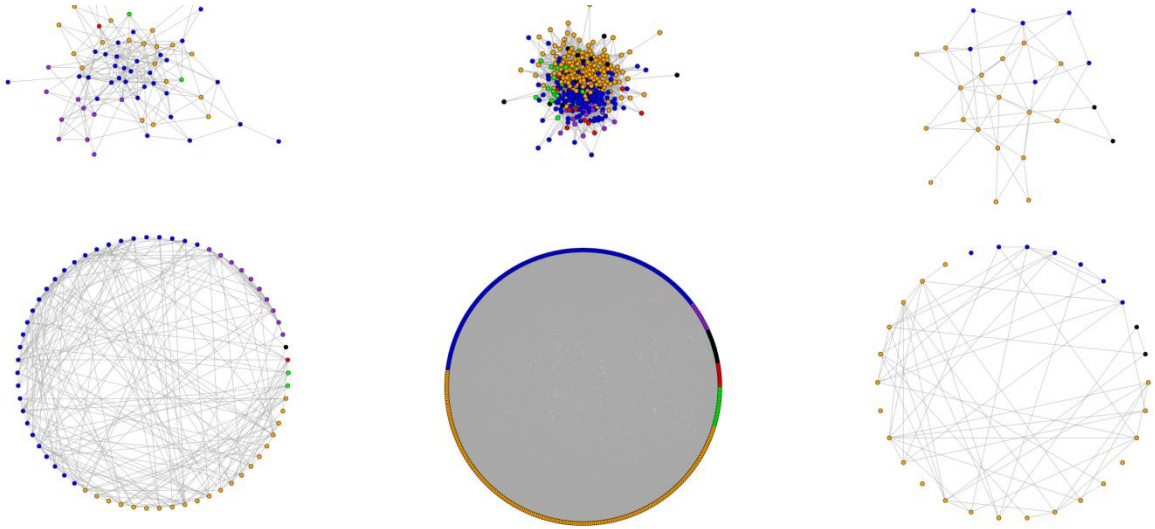


```
"Comedy"] = "orange"
"Drama"] = "blue"
"Horror"] = "black"
"Action"] = "green"
"Romance"] = "red"
"Fantasy"] = "purple"
```

Results

- Goals
- Data
- Code
- Results
- Improvements

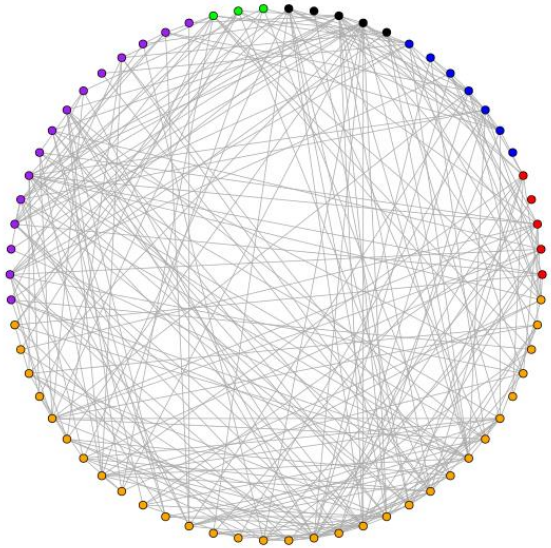
- Even with the distinct regions of color, it's a bit messy.
- Can try looking at smaller subsets.
- Can try different ways of organizing the map visually.



Results

- Goals
- Data
- Code
- Results
- Improvements

- Maybe network structure for studios would be easier to see with a similar treatment?

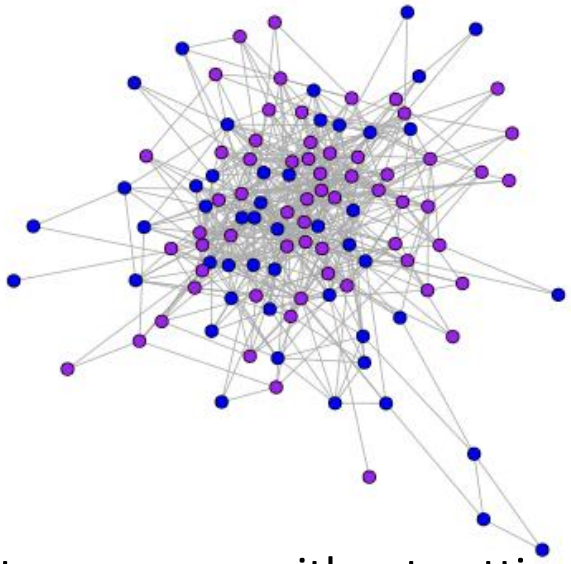


- Hard to say . . .

Results

- Goals
- Data
- Code
- Results
- Improvements

- Maybe network structure for genre was easy because there were two dominate categories?
- Can try to look at two major studios only, Universal and Fox.



- Most we can say without getting quantitative is that the segregation by genre is more distinct.

Improvements

- Goals
- Data
- Code
- Results
- Improvements

- More complete records by merging with other data sets.
- Account for overlapping genres, such as romantic comedies.
- Color the links themselves based on what two nodes they are connecting.
- Try visualization methods meant specifically for larger networks, such as hive plots.
- Get quantitative: easy ways to parameterize “more segregated” would be to count the colored links for different groups, or determine the average path length to node of another color.