

Big Data HW #5

Will Clark & Matthew DeLio
41201-01

May 15, 2015

1 Plotting the Actors' Network

In this section we plot the entire actors' network (see Figure 1 on the following page). For the plot on the left (1a), the node size remained constant with the vertex color varying based on an actor's degree of connectedness (more blue for less connected and more red for well connected). The data set has 7015 actors in it. The degree of connectedness in the set of actors is shown in Figure 2 on the next page. This histogram shows that most actors have worked with between 40 and 60 other actors, although the heavy tail in this distribution indicates that many have worked with far more. For instance, **Vernon Dobtcheff** (the most connected actor in the sample) has worked with 378 others.

When initially inspecting Figure 1a, we got the impression we had done something wrong since there were not many nodes plotted on the graph. To investigate this further, we plotted Figure 1b. In this graph, we used alpha blending (partial transparency) on each vertex node with $1/256^{\text{th}}$ alpha transparency per vertex. Therefore, if 256 or more vertices happened to overlap, it would look as if there were one opaque vertex in its place. This plot clearly shows that first figure groups most actors into clusters and few actors are displayed as independent nodes. This indicates that many groups of actors, possibly from a particular region or sub-genre, tend to work together.

2 Kevin Bacon

Kevin Bacon's network is shown in Figure 3 on page 3. In each of these plots, the node size of each actor is determined according to their degree within the overall actors' network. That is, more connected actors have bigger nodes. There are 96 first-degree connections (shown in red); 2128 second-degree connections (in green); and 5980 third-degree connections (in blue); Kevin Bacon's node is black. The size of his network grows exponentially as we consider more degrees of connection.

In Figure 3a we see his 1st-degree connections and note that while Kevin Bacon is relatively well connected (96 connections), 22.68 percent of his connections have a higher degree than he does. In Figure 3b we visualize 2nd-degree connections and note that of the 2128 connections, 27.99 percent have a higher degree. Because of the way R rendered the visualization, we cannot really see any

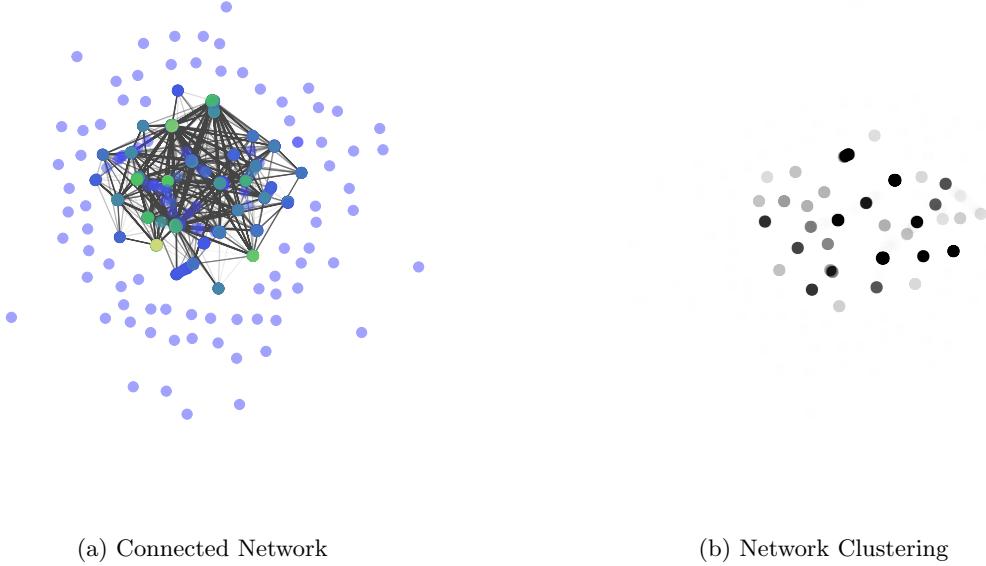


Figure 1: Visualization of Actors' Network

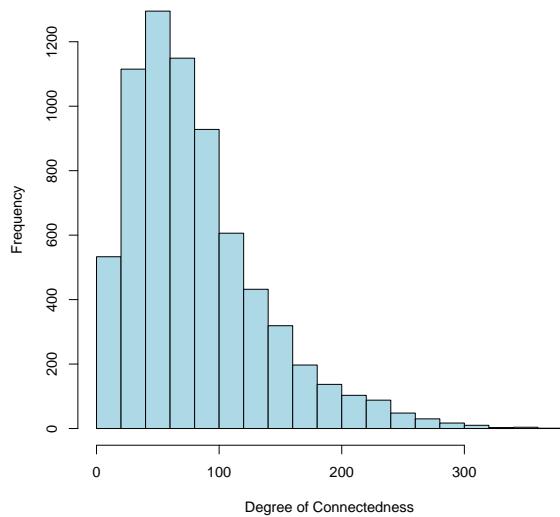


Figure 2: Histogram of Actors' Connectedness

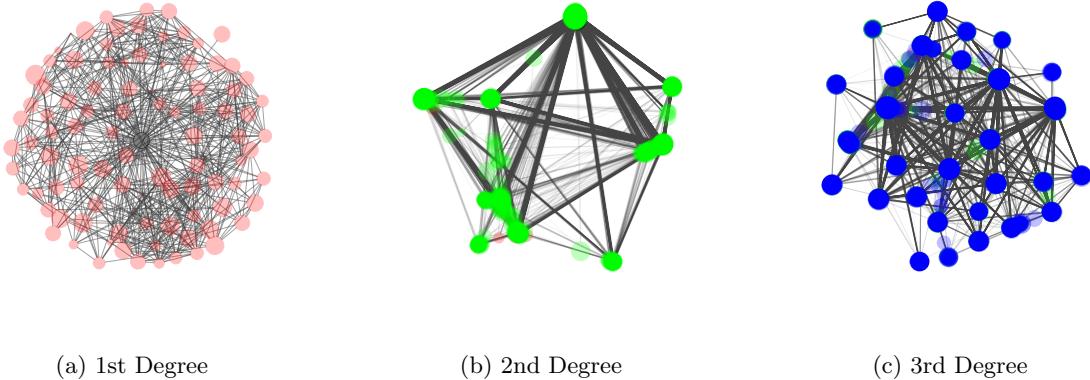


Figure 3: Kevin Bacon’s Network

1st degree connections. In Figure 3c, we visualize 3rd-degree connections and see a much better visual dispersion of the three types of connections. Of the 5980 connections, 31.23 percent have more connections than Kevin Bacon.

3 Most Common and Connected Actors

We can determine the most common actors by summing up the columns of the actors’ matrix and sorting the result. The actor appearing in the most films is **Velimir ‘Bata’ Zivojinovic**, who appeared in 57 films has appeared with 188 other actors. The most connected actor is **Vernon Dobtcheff**, who has been in 47 films and has worked with a total of 378 actors.

The most connected actors and busiest actors (by # of movies completed) during the period are shown in Tables 1 and 2 on the current page and on the following page respectively. Not surprisingly, there is some overlap between the two (**Vernon Dobtcheff** appears in both lists); it is also worth noting that Ron Jeremy is one of the busiest actors in the drama category (leading us to ask—where do these data come from?).

	# of Connections	# of Films
Vernon Dobtcheff	378	47
Jean-François Stévenin	356	36
Jean-Paul Muel	355	40
Roland Blanche	351	40
Victor Garrivier	341	37

Table 1: Most Connected Actors

We randomly select two actors **Jean-Gabriel Nordmann** and **Rosamaria Murtinho** and then determined the shortest path between them. Using R and iGraph, we found that the shortest path contains 3 vertices: Jean-Gabriel Nordmann was in *588 rue paradis* with Claudia Cardinale (I), Claudia Cardinale (I) was in *Fitzcarraldo* with Grande Otelo, Grande Otelo was in *Natal da Portela* with Rosamaria Murtinho. See Figure 4 on page 6 for a visualization of this path.

	# of Connections	# of Films
Velimir 'Bata' Zivojinovic	188	57
Ron Jeremy	234	51
Dora Doll	331	47
Vernon Doubtcheff	378	47
François Berléand	277	42

Table 2: Busiest Actors

4 Association Rules

We first looked for association rules with at least 0.01 percent support and 10 percent confidence; there are around 93 thousand such rules. To make the list of rules more manageable to look at, we changed the minimum level of support from 0.01 percent to 0.1 percent. This change resulted in a set of 18 rules, displayed in Table 3

lhs	rhs	support	confidence	lift
{Royle, David (I)}	=> {Buchanan, Colin (I)}	0.00105	0.8824	743.56
{Buchanan, Colin (I)}	=> {Royle, David (I)}	0.00105	0.8824	743.56
{Royle, David (I)}	=> {Clarke, Warren}	0.00105	0.8824	468.17
{Clarke, Warren}	=> {Royle, David (I)}	0.00105	0.5556	468.17
{Buchanan, Colin (I)}	=> {Clarke, Warren}	0.00105	0.8824	468.17
{Clarke, Warren}	=> {Buchanan, Colin (I)}	0.00105	0.5556	468.17
{George, Götz}	=> {Feik, Eberhard}	0.00119	0.7083	307.50
{Feik, Eberhard}	=> {George, Götz}	0.00119	0.5152	307.50
{Hegstrand, Michael}	=> {Laurinaitis, Joe}	0.00105	1.0000	895.38
{Laurinaitis, Joe}	=> {Hegstrand, Michael}	0.00105	0.9375	895.38
{Foley, Mick}	=> {Austin, Steve (IV)}	0.00105	0.8824	665.29
{Austin, Steve (IV)}	=> {Foley, Mick}	0.00105	0.7895	665.29
{Zivkovic, Vladan}	=> {Cipranic, Ljubomir}	0.00105	0.5556	209.44
{Cipranic, Ljubomir}	=> {Zivkovic, Vladan}	0.00105	0.3947	209.44
{Janicijevic, Dusan}	=> {Zivojinovic, Velimir 'Bata'}	0.00119	0.4595	115.48
{Zivojinovic, Velimir 'Bata'}	=> {Janicijevic, Dusan}	0.00119	0.2982	115.48
{Milinkovic, Predrag}	=> {Cipranic, Ljubomir}	0.00105	0.3659	137.93
{Cipranic, Ljubomir}	=> {Milinkovic, Predrag}	0.00105	0.3947	137.93

Table 3: Association Rules

We can take a close look at the first listed rule to understand what the association rules are telling us. The **lhs** variable is **David Royle** and the **rhs** variable is **Colin Buchanan**. The support of this rule is 0.00105, meaning that in our sample of films, there is a 0.105 percent chance of randomly drawing a film with both of these actors. The confidence of this rule is 0.882—given that **David Royle** is in a movie, there is an 88.2 percent chance that **Colin Buchanan** is in that movie as well. The lift of this rule is 744, which means that the probability of a movie having **Colin Buchanan** is 744 times higher if we know that **David Royle** is in the movie as well.

A more intuitive way of thinking about these results (using the language of more basic statistics)

is that for two actors A and B:

- Support (for the rule) is the joint probability of A and B being in a film;
- Confidence is the probability that B is in a film conditional on A being in the film; and
- Lift is the ratio of confidence (conditional probability) to the unconditional probability (support) of B.

5 Regression Alternative to Association Rules

Here we look at one association rule and replicate its results with a binomial regression. We examine the association between **David Royle** and **Warren Clarke**:

lhs	rhs	support	confidence	lift
{Royle, David (I)} =>	{Clarke, Warren}	0.00105	0.882	468

For notational purposes, we will abbreviate these two as R and C for *Royle* and *Clarke* respectively. First, recognize that the conditional probability (i.e. confidence) can be found by regressing C on R :

$$D_{i,C} = \alpha + \beta_R D_{i,R} + \varepsilon$$

$D_{i,R}$ and $D_{i,C}$ are dummy variables that equal 1 if the actor is in movie i and 0 otherwise. The sum $\alpha + \beta_R$ is the odds multiplier on C given R ; we can convert this to a probability using the logit link:

$$P(C|R) = \frac{1}{1 + \exp(\alpha + \beta_R)} = 0.882$$

This value matches the confidence of our selected rule. The support of our selected rule is the joint probability of R and C being in a film, which we can calculate as the mean of $D_{i,R} \cdot D_{i,C}$; since our variables are dummy variables with value 1 or 0, the mean of this product gives us the probability of both jointly occurring.

$$P(C, R) = \overline{D_{i,C} \cdot D_{i,R}} = 0.00105$$

This matches the support of our given rule. The lift is just the conditional probability (confidence) divided by the unconditional probability of C , which we can similarly calculate as the mean of $D_{i,C}$. The lift is:

$$\frac{P(C|R)}{P(C)} = \frac{1}{1 + \exp(\alpha + \beta_R)} \frac{1}{\overline{D_{i,C}}} = 468$$

This also matches the lift of our given rule.

6 Appendix



Figure 4: Shortest Path Between Actors