Lowering Six Degrees of Kevin Bacon by Genetic Algorithm

Brendan J. Babb  
Department of Computer Science  
University of Alaska Anchorage  
Anchorage, Alaska, USA  
brendanbabb@gmail.com

Bilal Gonen  
Department of Computer Science  
University of West Florida  
Pensacola, Florida, USA  
bgonen@uwf.edu

Abstract—Six Degrees of Kevin Bacon refers to a game where you try to link any actor to Kevin Bacon by the fewest movies. Every actor has a Bacon number. If an actor appears in the same movie as Kevin Bacon his Bacon number is 1. This paper creates a new movie with a cast that lowers the weighted average for Kevin Bacon, which could be useful for increasing the connectivity of a network.

Keywords—complex networks; genetic algorithm, six degrees of separation

I. INTRODUCTION

Six degrees of separation is the idea there are on average, six links between any two people on the planet. Six degrees of Kevin Bacon [1] is a popular game that plays along with this concept; it is based on the idea that every actor is connected to Kevin Bacon by six links or less. The Bacon number of an actor or actress refers to the number of links in the shortest path to Kevin Bacon from the actor or actress. Kevin Bacon has a Bacon number of 0. If an actor has been in a movie with Kevin Bacon, his or her Bacon number is 1. An actor who has not co-starred with Kevin Bacon who is in a movie with a co-star of Kevin Bacon will have a Bacon number of 2. There might be multiple paths to arrive at Kevin Bacon, but the Bacon number refers to the total number of links in the shortest path. If an actor cannot be connected to Kevin Bacon, their Bacon number is infinity.

Six Degrees of Kevin Bacon became popular in the late 1990s and was turned into a board game. The Oracle of Bacon [2], is a web page that can compute the Bacon number of any actor as well as the shortest path between any two actors. Of the 4.5 million actors currently listed in the Internet Movie Database (IMDB), roughly 10% have an infinite Bacon number. In addition to these actors, there are 200+ others who have a Bacon number ranging from 7 to 10, which some say, disproves the six degrees of Kevin Bacon. The Oracle of Bacon also shows the weighted path average for several other actors, and shows that Kevin Bacon is not the center of the movie world. The weighted average is calculated by taking the number of actors one degree away, call this A, and then the number of actors 2 degrees away, call this B. A simple weighted average would be (A*1+B*2)/(A+B). You would extend this formula for 3 degrees away, etc. Kevin Bacon's weighted average number is currently 2.984. Many other actors have a lower weighted average of links, like Donald Sutherland (2.822), Dennis Hopper (2.827), and Harvey Keitel (2.811).

The movie actor network is a small world network which is denoted by a short average path link and a higher degree of clustering than would be found in an Erdos-Renyi random graph. In this paper, we are interested in computing the cast that would reduce the weighted average the most.

While the actual cast of a new Kevin Bacon movie is not of significant importance, there are potential applications for being able to decrease the size of a small world network with regard to a given node. These applications include peer to peer packet networks, power grids and advertising on social networks. The methods are used to calculate and find the needed new links to decrease the size of the Kevin Bacon small world could prove useful as well.

Challenges in working with the IMDB network are that the database includes documentaries, award shows, TV movies, video games, adult and animated films. This blurs the definition of co-starring with Kevin Bacon. Should people who voice characters in an animation be considered connected? Should adult movie actors be included? An award show frequently has hundreds of actors appearing in the show. If Kevin Bacon appears in the show then all the other actors have a Bacon number of 1 because of the show. The actors aren't acting together in the sense of a movie, so the award shows were removed from the data set. In the movie JFK by Oliver Stone, actual archival footage of John F. Kennedy is used, so in some networks he would have a Bacon number of 1. It is important to be consistent on what definitions for vertices and edges were used for the actor movie network in this paper and in previous work, so that you can compare apples to apples.

The Oracle of Bacon has focused on the question of who makes a better center than Kevin Bacon, but not on how to make Kevin Bacon a better center. Previous work has shown the various components and cliques within the movie actor network, but not how to connect them. In this paper, we will answer those questions using different techniques to categorize the clusters and components of the network and finding ways to join disparate clusters. It might be possible to create a new movie that lowers Bacon numbers but doesn't have Kevin Bacon in the movie.

The next section will contain background on previous research on this problem, and will be followed by a
methodology section, a results section and finally the conclusion.

II. BACKGROUND

In [3] the author shows how Six Degrees of Kevin Bacon can be used to help teach graph theory by engaging his undergraduate students. Written in 2004, the author laments that there isn’t much large network data, and that no database lists all of his friends, indicating his research was before the advent of Facebook. The paper is slight and doesn’t break much ground on techniques. That said, the paper made a few interesting observations. It mentioned that Olga Holts is part of a 29 vertex component of 1920s Estonian silent film actors, and has an infinite Bacon number. These isolated components could be useful in decreasing the size of the small world. It goes on to point out that Actor Wesley Snipes has 180 fewer 1st level co-stars than Kevin Bacon, but has an overall weighted average number less than Bacon. The paper also asks about what happens to an actor’s Bacon number once they have died or retired. Overall the paper is very introductory and focused on using the actor movie database as a viable teaching resource. The paper probably only appeared in the literature search because of Kevin Bacon in the title.

Duncan Watts was one of the first people to study the small world phenomenon with larger data sets with Steven Strogatz and in [4] he explores the small world of Kevin Bacon. The author observes that there is a major connected component in the network that contains most of the actors. In 2000 when the research was conducted, it was comprised of 225,000 actors and the average path length between any two actors was less than 4. The network is classified as a small world, as previously mentioned because compared to a random network of the same size, the average path for the actors is 3.65 and for a random network it is 2.99. The clustering coefficient for a random network is 0.00027, but for the actor network it is 0.79, several degrees larger. Watts compares this to a power grid network and a worm cellular network and observes the same similar average path link to a random network's average path link, but a much higher clustering coefficient. The book describes a famous experiment Stanley Milgram conducted where he asked people in the Midwest to send a letter to someone they didn’t know in Boston with the rules that they had to mail the letter to someone they knew on a first name basis that might be more likely to get the letter to the final destination. Many of the letters did make it to the destination and Milgram estimated in the US that everyone was connected on average by 6 connections. If this was attempted by randomly choosing the person to send the letter to, it would take many more connections, but people were able to make educated choices about different groups and which group might be more closely connected to the final destination, either by region or profession. This is useful in terms of thinking of movies by genre and decade released. The author makes reasonable assumptions in the book and discusses many other small worlds that could prove useful to this problem.

Malcolm Gladwell in a New Yorker article [5] discusses the Six Degrees of Lois Weisberg, and summarizes the amazing number of connections that Lois has accumulated over her lifetime. He suspects it is because of several different circles of friends that she is part of. Gladwell also briefly touches on Kevin Bacon and how the actor has worked in many genres which he suspects gives him more connections. He delineates the movies of Burgess Meredith and how he starred in movies, and also played bit parts, and was in Oscar films and also really bad films. The key take away is that in any of the small worlds, people that are close to the center have interacted with many tight sub-components. Although his article is written for a general audience there are useful references and thought experiments.

In [6] there is a focus not only on analysis of the IMDB but on the visualization of the network. The authors focus on Kevin Bacon in part of the paper and create a temporal animation of the Kevin Bacon graph. They identify a technique of using (p,q)-cores and 4-rings that help them isolate islands and meaningful sub-graphs. They look at actors in the IMDB connected to Kevin Bacon by decade and are surprised to find a large number actors with a Bacon number of 1 in the early part of the 20th century. The anomaly turns out to be caused by incorrect dates for movie years and for actors in movies with the same title being attributed to the earlier movie. JFK is also shown to have a Bacon number of 1 as previously mentioned, and a methodology for dealing with this, must be determined. The groupings of actors throughout the decades shows how certain actors would work consistently together as in the case of Dan Ackroyd, Bill Murry and Chevy Chase co-starring in several movies in the 1980s. Their analysis also shows adult films as a small sub-component where actors frequently work together but are far from the center of the overall network. Arnold Schwarzenegger starts to form political connections, when he runs for governor and that lowers Bill Clinton's Bacon number due to documentaries about politics. This reaffirms the need to decide on what movies should constitute a link. Their work also shows that different groups that align with genre fall out naturally when you look at actors that have worked frequently together within a decade. The (p,q)-core and 4-ring need to be studied more extensively but appear to be very useful tools for solving the thesis of this current paper. Overall the paper is excellent and links to animations of Kevin Bacon's network over time, that are not currently working, but I think would provide very useful insight into how to shrink the Bacon world.

III. METHODOLOGY

In order to test if a new movie lowers the weighted average we must be able to compute the new average. We set out to use Pajek [7] as the tool to analyze the data set. There were two data sets available. The Notre Dame Actors database for Pajek with roughly 200,000 movies. There is a data set from the University of Florida sparse matrix collection [8] that has 428,440 movies from 1865 to 2005, with 896,308 actors and is in a Matlab sparse matrix format. The data set contains over 3,782,463 entries of 1, indicating an actor has appeared in a movie.

Finding communities is NP complete [9] so instead of using an exhaustive search to find the separate communities, we decided to use a Genetic Algorithms (GA) [10] to generate potential new movies with Kevin Bacon and 14 actors and evaluate the new weighted average (WA) score. The search
space is 896,308 choose 14 which is a very large number but the choice of actors can be shrunk down based on node count and other filters.

GAs are based on Darwinian evolution and are valuable for searching large search spaces. A population of candidate solutions is created, in this case it is a vector with Kevin Bacon's actor number and 14 other actors chosen at random. An example candidate solution would be (1,2,3,...,14,15), that is 15 numbers long. As an example, assume an initial population of 100. The fitness of each individual is evaluated by adding it as a new movie containing the 15 actors to the data set and calculating the new weighted average (WA) for Kevin Bacon. Each member of the population has a fitness value and two parents are chosen based on their fitness, with higher fitness individuals being chosen to reproduce more often than lower fitness individuals. A more fit individual will have a lower WA and will have a higher probability of being chosen. The two parents will mate and have two children that have part of their actors from one parent and part from the other parent based on a crossover point chosen at random. If more than one of the same actors are in a child, the duplicates will be removed and additional random actors will be chosen to make a list of 15 actors. Mutation also occurs at random and replaces an existing actor with a new random actor. This process goes forward replacing the parents with their children. If the crossover rate is 80% out of 100 individuals, 80 are replaced with child individuals and 20 members of the original population move forward unchanged in each generation. There is a mutation probability that is usually 5% that would choose 5 random individuals in this case and randomly change an actor in the individual. This makes up a new generation. Each member of the new generation has their fitness evaluated and then the process takes place all over again. Slowly from generation to generation the best overall individual's fitness improves. Individuals who have a higher fitness have a higher chance of being chosen for crossover and their grouping of actors tend to propagate through the population, resulting in individuals with 15 actors that lower the WA for Kevin Bacon.

The weighted average score takes almost two seconds to compute in Matlab.

The data set contains a binary sparse matrix, which has the number of movies rows by number of actors columns. If actor Y is in movie X then the matrix A(X,Y)=1 and if the actor is not in that movie than the value is 0. There are auxiliary tables that contain the genre, year, and actor names. The movie names are not in the data set. In first looking at the data set, I tried to see how much the weighted average went up by removing a given Kevin Bacon movie. There was an increase of 0.02 when I removed one movie which I traced back to the Independent Film Awards of 2005. Since award shows contain many different actors, but not acting together in a movie, I would like to remove this genre from the data set. We also discovered a link from Fiona Apple to Kevin Bacon in the data set. When we traced down the movie, we discovered that it was a documentary called Music with interviews of musicians including the Bacon Brothers (a band including Kevin Bacon and his brother). I was interested in what the average number of movies an actor has appeared in the data set. Analyzing this data led me to an actor that had appeared in over 2000 movies, who turned out to be Peter North, an adult film actor. The intent of the game of Six Degrees of Kevin Bacon is to name Hollywood movies that two actors appeared in together, not award shows or documentaries or music videos. I removed genres such as Documentaries, Short Films, Award Shows, Animation, and Adult. This might remove key Bacon movies and other key movies, but otherwise I don't have a good way of removing the adult movies.

We contacted the Oracle of Bacon and to see if I could use their data set. It is updated monthly and has better genre descriptions and since it includes movies up to 2012, it includes an extra 7 years of data. They responded but the data format was too difficult to parse in a short time frame.

Using the GA allowed me to start discovering combinations of actors that lower the WA for Kevin Bacon. After attempting to use the Oracle of Bacon data with little success, the University of Florida data was selected again. The sparse matrix was pruned by removing the following genres; adult, animation, musicals, music, award shows, game shows, documentaries, short films and TV shows. After the movies are removed, all actors who are now in 0 movies are removed.

Experiments were run with the GA for casts that did not include Kevin Bacon, but the improvement was very slow and that approach could be examined in further research.

IV. RESULTS

The pruned data set contained 143,325 movies and 640,039 actors and actresses. The data set contained movies from the 1900s up to 2004. The data set was in a sparse matrix D, with D(M,A)=1 indicating actor A was in movie M. If actor A did not appear in movie M, then D(M,A)=0. A simple GA was used to create a vector of 14 integer values from 1 to 640,039 (the number of actors) and the actor value for Kevin Bacon was added to the vector to create a 15 actor individual in the GA population. The fitness was evaluated by appending the new vector to the dataset, creating a new movie containing the actors. The weighted average was then calculated for this new data set and the WA was the fitness for the individual. The GA used a population of 40 individuals and a 30% crossover rate was used with single point crossover. The mutation operator mutates an actor by replacing the actor with another actor chosen at random. The GA ran for 600 generations and lowered the WA from 2.6868 to 2.6650 in the best run.

<table>
<thead>
<tr>
<th>Bacon #</th>
<th>Current</th>
<th>New Bacon Movie</th>
<th>Weighted Cur.</th>
<th>Weighted New</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3149</td>
<td>3163</td>
<td>3149</td>
<td>3163</td>
</tr>
<tr>
<td>3</td>
<td>230346</td>
<td>235863</td>
<td>460692</td>
<td>471726</td>
</tr>
<tr>
<td>4</td>
<td>363212</td>
<td>365783</td>
<td>1089636</td>
<td>1097349</td>
</tr>
<tr>
<td>5</td>
<td>34371</td>
<td>6369</td>
<td>137484</td>
<td>25476</td>
</tr>
<tr>
<td>6</td>
<td>1705</td>
<td>1611</td>
<td>8525</td>
<td>8655</td>
</tr>
<tr>
<td>7</td>
<td>184</td>
<td>178</td>
<td>1104</td>
<td>1068</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>11</td>
<td>77</td>
<td>77</td>
</tr>
</tbody>
</table>

Fig. 1 Change in number of actors at each Bacon Number and the weighted change.
Fig. 2 Weighted Actors vs. Bacon Number – Blue is current, and red is after the new movie.

You can see the number of actors of Bacon Number x stays relatively the same, except with Bacon Number 5, where the new cast drops the number from 34,371 to 6,369.

The cast of this new movie consisted of Jose Julio Spiewak, Pierre Juvenet, Hector Bidonde, Oya Palay, Boris Novikov, Elza Gomes, Gordon Kennedy (III), Veikko Linna, Gustavo Rojo (I), Jonas Gardell, Sanna Majuri, Durga Khote, Birgitta Andersson (I), Marisa Zotti and Kevin Bacon. None of these actors are well known to the average American. Many actors are predominately in foreign films. Jose Julio Spiewak, for example was born in Poland in 1931 and has been in 38 films. His last film was a Brazilian film in 1984 and he had 719 co-stars total. Whereas Marisa Zotti was in 2 films and has only 5 co-stars total. Durga Khote is an Indian actress who appeared in 143 films from 1931-1983, and had 843 co-stars total. Visualization would help gain insight into the different cliques in the film world, but it appears that the actors chosen are from particular world regions of film. This makes sense, as starring with a well-connected Bollywood star would shorten the distance between Kevin Bacon and several Bollywood actors.

V. Conclusion

Using a GA, you can search a fairly large search space of movies and actors and find combinations of actors that lower the WA for Kevin Bacon. But without proper visualization you don't learn anything intuitive about why the actors were chosen. Visualization could help see what cliques and components were possibly joined by linking these actors to Kevin Bacon. The GA worked quite well and didn't require tweaking to get good results. Our work shows that using a GA would be a promising method for working with complex networks when finding communities would be very time intensive. This provides a technique when you are interested in making a smaller world for a particular node in a network and not necessarily the entire network. If it could be shown that the actors chosen were from different communities, you might be able to use the quick GA method to identify actors from different communities and then map out those actors’ networks. Reversing the problem of community detection.

VI. Future Work

Visualization of the network would allow insight into how to lower the WA. There are many possible methods for reducing the size of the data set to see what connections could be made. The GA could also be enhanced with a more advanced mutation and crossover and possibly just select candidate actors from actors who had a Bacon number larger than one. We are looking into methods for discovering the communities within the data set based on Newman’s article [11] on finding communities by eigenvectors and modularity. Creating weak links from Kevin Bacon to the separate communities will hopefully lower the WA, although the computational complexity needs to be researched. Additional methods for analyzing the network and finding good weak links can help lower the WA.

REFERENCES