STANFORD UNIVERSITY
Stanford Center for Professional Development

HOMEWORK ROUTE FORM

Stanford Center for Professional Development Student Information

Course No. CS224W

Faculty / Instructor Name Jure Leskovec

Date 10-Dec-2013

Student Name Joshua Brian Lindauer

Phone 512.326.2276

Company Carnegie Mellon University

Email brian@shendauer.com

City. Austin

State TX

Country USA

Check One: ○ Homework #: ○ Midterm ● Other Project Milestone

The email address provided on this form will be used to return homework, exams, and other documents and correspondence that require routing.

Total number of pages faxed including cover sheet 10

	For Stanford Use Only	
Date Received by the Stanford Center for Professional Development	Date Instructor returned graded project Score/Grade: (to be completed by instructor or by teaching assistant)	Date the Stanford Center for Professional Development returned graded project:

Please attach this route form to ALL MATERIALS and submit ALL to:

Stanford Center for Professional Development
496 Lomita Mall, Durand Building, Rm 410, Stanford, CA 94305-4036
Office 650.725.3015 | **Fax** 650.736.1266 or 650.725.4138
For homework confirmation, email scpd-distribution@lists.stanford.edu
http://scpd.stanford.edu

Last modified October 27, 2008

CS224w:
Social and Information Network Analysis

Assignment number: Project Final Report

Submission time: 11:00am **and date:** 10-Dec-2013

Fill in and include this cover sheet with each of your assignments. It is an honor code violation to write down the wrong time. Assignments are due at 9:30 am, either handed in at the beginning of class or left in the submission box on the 1st floor of the Gates building, near the east entrance.
Each student will have a total of *two* free late periods. *One late period expires at the start of each class.* (Homeworks are usually due on Thursdays, which means the first late periods expires on the following Tuesday at 9:30am.) Once these late periods are exhausted, any assignments turned in late will be penalized 50% per late period. However, no assignment will be accepted more than *one* late period after its due date.

Your name: Joshua Brian Lindauer, Raj Bandyopadhyay, Rafael Guerrero (Group 8)
Email: brian@shendauer.com **SUID:** jbl, rajb2, rplatero

Collaborators:

I acknowledge and accept the Honor Code.

(Signed) J. Brian Lindauer

(For CS224w staff only)

Late periods: 1 2

Section	Score
1	
2	
3	
4	
5	
6	
Total	

Comments:

Egonetwork Properties and Successful Filmmaking*

Raj Bandyopadhyay†
Atlanta, GA
plusbzz@gmail.com

Rafael Guerrero-Platero‡
Atlanta, GA
fourais@gmail.com

Brian Lindauer§
Austin, TX
brian@shendauer.com

ABSTRACT

Small-world networks are widely studied in many contexts due to their prevalence. For example, the collaboration network of movie actors is a small-world network, with short diameter and high clustering[6]. However, the availability of large-scale, crowd-sourced metrics of movie popularity and success allow us to gain deeper insight into how global and local network properties actually influence the careers of individual artists (both actors and directors) and the critical or popular success of the movies they act in. In this project, we analyze critic and audience ratings data from the website Rotten Tomatoes to provide insights into these questions. After verifying the global small-world and power-law properties of the networks, we use Personalized PageRank to estimate the quality of individual artists to the movies they work on. Then, we use both Spearman's rank correlation coefficient and linear regression for various egonet graph properties to determine whether they are correlated with our artist quality metric. In performing this analysis both for the entire dataset, and as time series data, we conclude that small world properties of film artist's egonets correlate negatively with the quality of their artistic output. Instead, the most valuable predictors of their success seem to be having ties to multiple disparate communities. Finally, we find that the artist ratings time series lags the egonet property timeseries by three years.

1. INTRODUCTION

There's an old saying about success that "it's not what you know, but who you know". Scientists studying social networks have taken this a step further in some cases to show how the specific structure of the social network around you predicts success[5]. We set out in this research to determine whether network properties correlate with successful moviemaking. In particular, we would like to discover whether any local network properties of actors and directors (which we refer to collectively as "artists") are predictive of the success of their movies, and whether those properties and the success of an artist's ventures evolve together over time. What do the properties of this collaboration network look like and do they differ when partitioned by genre? This extension of prior work in [5] to study the evolution of the network over time could provide valuable insights into the role of network structure in the success of creative ventures.

*Stanford CS224W Project Final Report

†SUID: rajb2

‡SUID: rplatero

§SUID: jbl

2. RELATED WORK

2.1 Barabasi et al (2002): Evolution of the social network of scientific collaborations

This paper[1] models the dynamics that govern the topology and evolution of social networks, and in particular the collaboration network of scientists. While key network properties like average path length, clustering coefficient and average degree are usually interpreted as static values, the authors show that these values change over time. This paper gives us a good reference methodology to do temporal analysis on our movie-artist bipartite graph, and it will be of interest to see if the same temporal tendencies observed by the authors (i.e. decreasing average path length, decay in clustering coefficient, increasing of average degree) are also observable in our dataset.

2.2 Uzzi and Spiro (2005): Collaboration and creativity: The small world problem

In this insightful paper[5], the authors investigate the bipartite network of creative artists working on Broadway musicals from 1945 to 1989. In addition to the global small-world coefficient Q of a network in a given year, they investigate the impact of local network properties, such as local density, structural holes and closeness centrality on the financial and artistic success of the musicals. Their most interesting finding is that the relationship between success and Q is parabolic; as Q increases, the probability of a musical succeeding increases upto a point, after which it falls. Their explanation is that a Q that is too small inhibits creativity due to lack of new ideas, while a Q that is too big inhibits creativity by inducing homogeneity.

2.3 Joshi et al (2010): Movies and revenues: an experiment in text regression

This paper[3] is representative of a somewhat popular topic area in machine learning – predicting box office revenues of movies. The authors' major contribution is the incorporation of the text of pre-release critic reviews into their predictive linear model (with elastic net regularization). The paper reports results using a number of combinations of metadata features and text features. Our interest pertains to the use of actor and director metadata to predict movie outcomes. The regression weights offer one possible way to quantify the contribution of each individual to success or failure of movies. However, after studying this possibility, we found that it did not provide accurate results. This is explained further below.

3. ROTTEN TOMATOES DATA SET

3.1 Description

The data set we use in this project has been collected from the Rotten Tomatoes website by Chau et. al. [2]. It contains movies, ratings by critics and users, actors, directors, links to more information and links to similar movies. More specifically, it contains 197789 movies ranging from 1891 to 2016 and metadata for each movie, including studio name, up to 10 genres, 5 cast members, and 2 directors, the MPAA rating, runtime in minutes, critic rating (from 1-100), audience rating (from 1-100), and links to similar movies as suggested by human contributors. A "rating" here represents the percentage of reviews that were positive.

3.2 Preprocessing

As Figure 1 shows, the majority of movies in the database were released after 1990. In addition, both critics and audience ratings show a nostalgia effect. That is, older movies get higher ratings, presumably because only the best older movies are actually watched and reviewed today. To avoid confounds related to this earlier data, we work only with the data from 1990 onward.

Furthermore, the "Music" and "Sports" genres do not employ the type of collaborative filmmaking we seek to study here because they consist almost entirely of concert videos and highlights of a particular sports team's season. Because of this, we filter out all movies from these two genres before performing our analysis.

From the remaining data, we form two graphs. The artist-artist graph contains a node for each actor or director and an edge if the two artists contributed to the same film. The movie-artist bipartite graph contains two node types (movies and artists) with an edge between them if an artist contributed to a movie.

4. GLOBAL NETWORK PROPERTIES OF ARTIST AND MOVIE NETWORKS

We initially investigate the global properties of two networks: the artist-artist collaboration network, and the movie-artist bipartite network.

4.1 Unipartite artist-artist collaboration network

The artist collaboration graph has 227139 nodes and 1244329 edges, with a diameter of 15 and average clustering coefficient (CC) of 0.75. This suggests a small-world network.

As shown in Figure 2, the degree distribution suggests a power law with estimated =2.42. The high fraction of nodes with CC=1 is an artifact of collapsing a bipartite network, due to a high number of cliques (one per movie)[4]. In addition, there is no obvious way to incorporate movie ratings in this network. For that, we must turn to the bipartite movie-artist graph.

4.2 Bipartite artist-movie network

The artist collaboration graph has 293672 nodes and 407502 edges, with a diameter of 36 and average CC of 0.39. Note that the CC is for the bipartite graph, as defined in [4]. The degree distribution of artists i.e. the number of movies an artist has contributed to follows (as expected) a power law distribution with an estimated =2.6. The bipartite CC distribution is peaks at very low and high CCs (Figure 3). Both networks have a giant connected component containing over 75% of the nodes.

5. ESTIMATING ARTIST CONTRIBUTIONS TO MOVIE RATINGS

A naive approach to quantifying artist quality might simply use the average critic or audience score over all of the movies that the artist contributed to. However, this gives equal credit to a master actor starring in the film and a lesser contributor who had less impact on the films success. We look for superior solutions to tease apart the contributions of the various contributors to each film.

5.1 Failed attempt to use Linear Regression to estimate artist quality

Our first attempt uses Linear Regression to estimate individual artist quality. A linear model would explain which present artists had the largest positive and negative impacts on the movie. We set up the regression as L2-regularized least squares after initial attempts without regularization resulted in a cascade of very large positive and negative coefficients:

$$\theta = \arg\min_{\theta} \|\theta^T X - Y\|_2^2 + \lambda \cdot \|\theta\|_2$$

Our input matrix X is the adjacency matrix of the movie-artist bipartite graph, where each row of the matrix represents a movie. Y is a vector of either audience or critic scores.

This method produces a puzzling scoring where middling actors rise to the top of the list. Results for this and following artist scoring methods are show in Tables 2 and 1. Analysis shows two probable reasons. First, an artist is penalized when they work with a very highly-rated artist, and great artists are likely to work together often. That is, if you have a movie with a critic score of 90%, 3 incredible artists collaborating in it might each be given a regression weight of only 0.3. The second, flip side, of this is that an okay artist working with a very bad artist, even on a poorly-rated movie, sees their coefficient artificially boosted to offset the very bad artist's highly negative coefficient. These failures lead us to look in a different direction for artist scoring.

5.2 PageRank based only on link structure (PR)

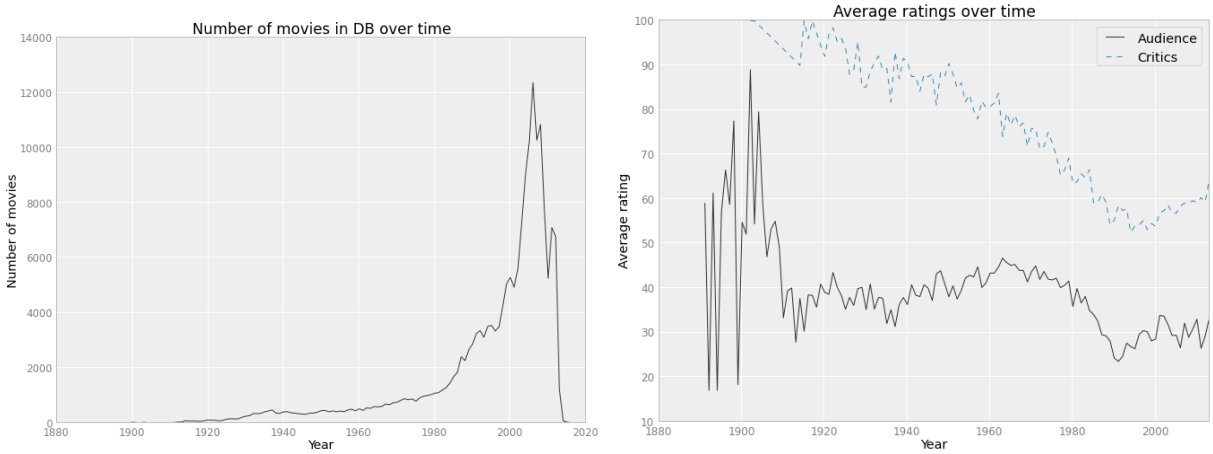

Figure 1: Basic information about movies and their ratings.

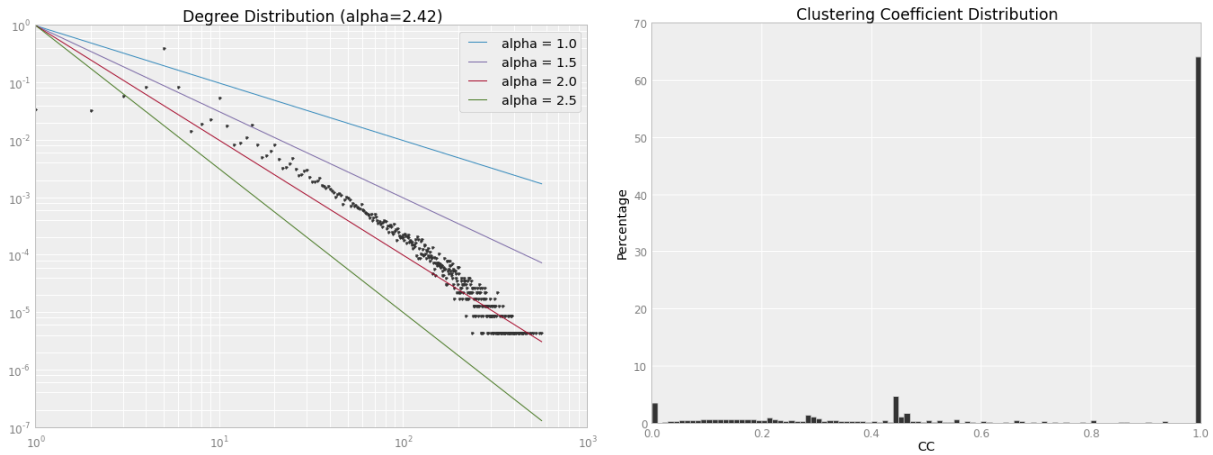

Figure 2: Distribution of degree and clustering coefficient for unipartite actor-actor graph.

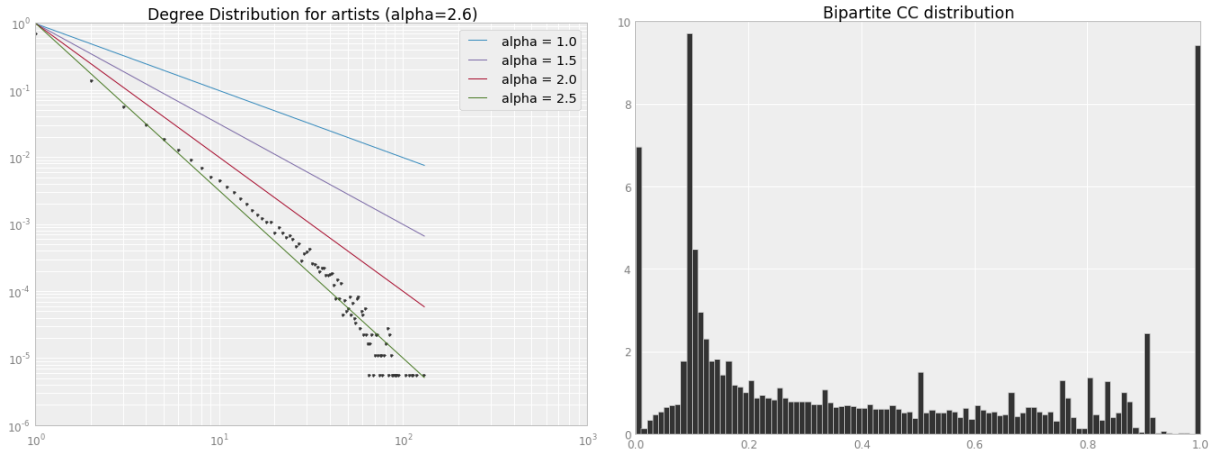

Figure 3: Distribution of degree and clustering coefficient for bipartite movie-actor graph.

Linear Regression	PR	PPR	PPR^2
1. Robin Harris	1. Samuel L. Jackson	1. Samuel L. Jackson	1. Samuel L. Jackson
2. Peter Sallis	2. Bruce Willis	2. Robert De Niro	2. Robert De Niro
3. D.L. Hughley	3. Robert De Niro	3. Anthony Hopkins	3. Anthony Hopkins
4. Jennifer Saunders	4. Nicolas Cage	4. Bruce Willis	4. Matt Damon
5. Arsinee Khanjian	5. Matt Damon	5. Nicolas Cage	5. Johnny Depp
6. Nathan Gamble	6. Woody Harrelson	6. Matt Damon	6. Julianne Moore
7. Nathan Fillion	7. Willem Dafoe	7. Johnny Depp	7. Ben Kingsley
8. John Waters	8. John Leguizamo	8. Julianne Moore	8. Steve Buscemi
9. Rodolfo Coria	9. Julianne Moore	9. Woody Harrelson	9. Bruce Willis
10. Mark Chandler Bailey	10. Anthony Hopkins	10. Ben Kingsley	10. Nicolas Cage

Table 1: Top 10 Actors by Critic Rating

Linear Regression	PR	PPR	PPR^2
1. Lucia Mendez	1. Jeff Smith	1. Barney	1. Barney
2. Alejandro Sanz	2. Barney	2. Muppets	2. Muppets
3. Garth Brooks	3. Anthony Wong Chau-Sang	3. Anthony Wong Chau-Sang	3. David Attenborough
4. Linda Cheung	4. Eric Roberts	4. Martin Sheen	4. Martin Sheen
5. Lou Albano	5. Joe Estevez	5. Mary-Kate Olsen	5. Anthony Wong Chau-Sang
6. Doug Stanhope	6. Simon Yam	6. Simon Yam	6. Mary-Kate Olsen
7. Bettany Hughes	7. Michael Madsen	7. David Attenborough	7. George Carlin
8. Rossana SanJuan	8. Martin Sheen	8. Andy Lau	8. Simon Yam
9. Charles Bukowski	9. Debbie Rochon	9. Peter Coyote	9. Noam Chomsky
10. Graham Hancock	10. Muppets	10. Noam Chomsky	10. Danny Glover

Table 2: Top 10 Actors by Audience Rating

For our next method, we calculate the standard Google PageRank for each node to investigate the baseline impact of the link structure of the bipartite graph. We perform this calculation on two graphs: the original graph and a graph only containing movie nodes with critics ratings, along with the associated artists.

As expected, the top ranking actors are simply the ones that are most popular. The audience list is more diverse internationally, while the critics list is focused on Hollywood actors. However, this method does not use any information about the ratings, which is our next step.

5.3 Personalized PageRank with ratings as teleport vector (PPR)

In order to incorporate ratings information, we use PPR. On teleportation, rather than selecting uniformly at random from a personalization set, we select from the set of all movies weighted by their audience or critics rating from 1-100. For an artist node, the teleport weight is set to 1.

These results are more sensible than the general PR. Why does it work? Since a random walker teleports more often to a highly rated movie, this in turn leads to a higher probability of reaching an artist in that movie. The more an artist is in highly rated movies, the higher the probability mass received by that artist.

5.4 Personalized PageRank with squared ratings as teleport vector (PPR^2)

Our results in the previous section are promising, but high ranking of Bruce Willis and Nicolas Cage as top critically-acclaimed artists cause us some concern about the fate of humanity. Our hypothesis is that the signal from ratings is not strong enough to overcome the information in the link structure. In order to emphasize the ratings, we re-run PPR, this time using the square of the ratings as teleport weight for movie nodes.

Bruce Willis and Nicolas Cage move down the rankings and Woody Harrelson is replaced by Steve Buscemi in Table 1, making for a list which we are satisfied passes a plausibility test.

One final enhancement, setting edge weights based on the ordering of the actors to give most credit to the movie's biggest contributors, does not appear to improve the results.

We would like to acknowledge two possible criticisims of our methodology. We used a fairly subjective critera for determining which of these artist rating methods was best. When we couldn't justify surprising results with the data, we continued to try new and enhanced methods. However, our purpose was only to find a method that produced a plausible and defensible scoring of artists with which to we could compare egonet properties.

Additionally, an astute reader may note that, taken to its limit, the PPR^2 method almost reduces to each artist's average movie rating. But it provides a couple of advantages over taking the average rating. First, it gives more credit to artists who have high ratings in multiple films. Second, PPR^2 is mostly driven by that average rating, but is also driven in smaller part by the quality of the artist's surrounding network. This is an improvement over simply using the average rating because working with other artists who are highly rated is a social signal that an artist is of high quality.

Henceforth, when we refer to "artist quality", we are measuring that in terms on this PPR^2 metric.

6. EGONET PROPERTIES OF ARTIST NETWORKS

Uzzi and Spiro find a strong correlation between critic success of theater artists and the small world properties of their collaboration network[5]. We set out to investigate whether a similar correlation exists for film artists when looking at their local ego-networks. To that end, we compute the following egonet properties. Each of these properties is computed over each artist's egonet, which consists of the artist plus any other artist with which they have collaborated.

- **nn**: Number of nodes.
- **ne**: Number of edges.
- **di**: Diameter. The longest shortest path length.
- **sp**: Average shortest path length.
- **dc**: Degree centrality. The percentage of nodes that the ego node is connected to. This is always 1.0 and should be excluded from future versions of this paper. Because it appears in some plots, it remains here for explanation.
- **bc**: Betweenness centrality. Fraction of shortest paths that pass through the ego node.
- **cc**: Clustering coefficient. The fraction of possible triangles that exist.
- **de**: Density. The fraction of possible edges that exist.
- **cstr**: Constraint. Equivalent to density. This refers to how constrained an artist is by being deeply embedded.
- **bo**: Brokerage. The fracion of possible edges that do **not** exist.
- **effsz**: Effective size. The number of neighbors of an ego node minus the average number of edges between its neighbors.
- **efficie**: Efficiency. Effective Size normalized by the actual size (i.e., the percentage of non-redundant neighbors).
- **wcc**: The number of weakly-connected components that remain after deleting the ego node.

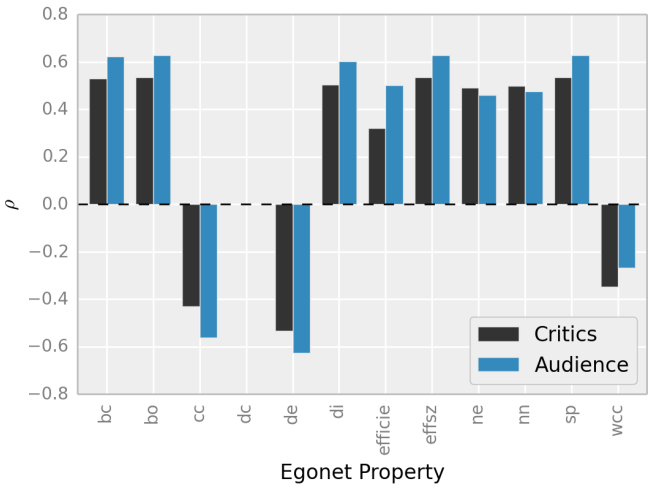

Figure 4: Spearman correlation between egonet properties and artist quality. Contrary to previous results, clustering coefficient (cc) is negatively correlated with quality.

These properties are not independent and some are nearly or literally equal. Specifically, we would expect brokerage, betweenness centrality, efficiency, effective size, average shortest path length, and diameter to all vary together. Likewise, we expect density, constraint, and clustering coefficient to vary together. Although it might seem that the **wcc** property should vary with brokerage/betweenness, we see below that this is not always the case.

7. RELATIONSHIP BETWEEN EGONET PROPERTIES AND ARTIST QUALITY

We measure the correlation between network properties and artist quality using Spearman's rank correlation coefficient (ρ). This measures how well the relationship between the two variables can be described as a monotonic function, without making any assumptions about linearity or even magnitude.

7.1 Using Data as a Whole

We do find strong correlations between egonet properties and artist quality, though ρ does not come close to 1.0 or -1.0, so it is clear these are not the only factors at play. The Spearman's ρ test yields a surprising result, since we might have expected, based on the global effects measured in [5], to find a positive relationship between clustering coefficient and artist quality. For film artists, the opposite appears to be true. Egnonet properties related to high embeddedness correlate *negatively* with artist quality, while properties related to high betweenness correlate positively with quality (Figure 4). It appears that in the film industry, at least, it is more advantageous to be on the periphery of multiple communities than to be deeply embedded in a single one.

7.2 Analysis by Genre

For the most part, egonet property correlations are consistent across genres, though the magnitudes may vary. Two exceptions are worth noting. First, for films labeled as "cult

movies", smaller egonets are positively correlated with quality, unlike most other genres, where larger egonets correlate positively with quality (Figure 5).

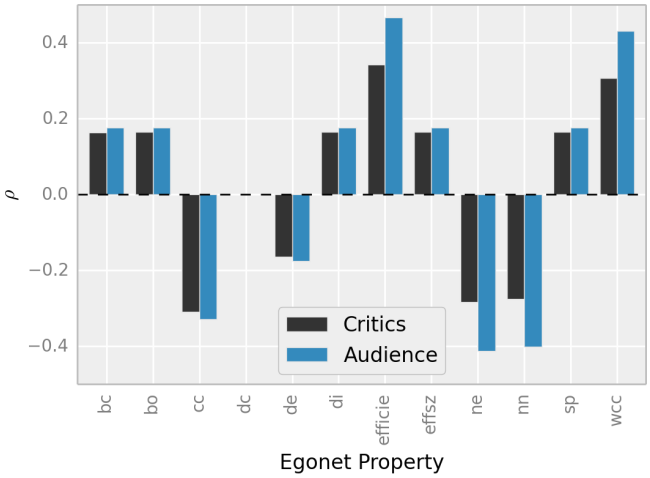

Figure 5: Spearman correlation between egonet properties and artist quality for cult movies. Unlike other genres, smaller egonets (small nn) have a positive correlation with artist quality.

Second, the correlation of the **wcc** metric with quality varies between genres. For some genres, like cult movies, being connected to many disparate components is an indicator of quality, while for others genres and for artists overall, there is small or sometimes even negative correlation between **wcc** and quality (Figure 6).

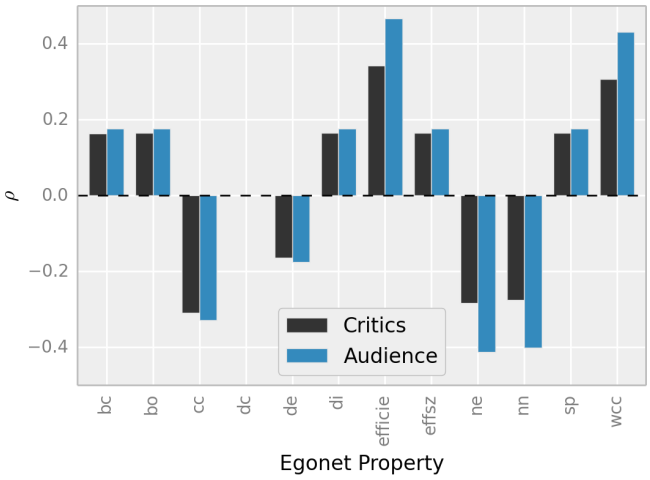

Figure 6: Note the negative correlation between wcc and quality. Comedic actors do not gain by working with disconnected communities.

7.3 Artist network examples

Two illustrative examples of this brokerage/quality principal are shown below. Michael Apted (Figure 7), a british director, writer, producer and actor has high brokerage/betweenness and a high artist quality score. Andy Samberg (Figure 8), a purely comedic actor has low brokerage and a correspondingly lower artist quality rating. It is bit difficult to see in the graph plots, but Apted's neighbors (less Apted) make up 4 connected components, while Samberg's neighbors (less Samberg) make up one giant connected component.

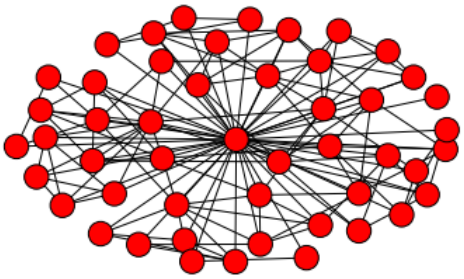

Figure 7: Michael Apted has high brokerage and a correspondingly high artist quality rating.

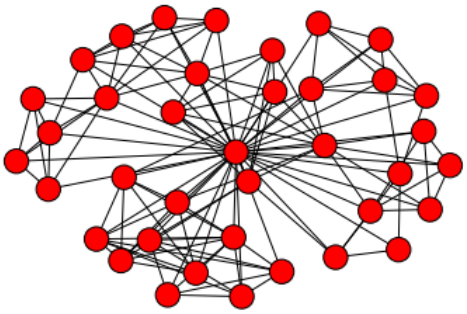

Figure 8: Andy Samberg has low brokerage (high embeddedness) and a correspondingly lower artist quality rating.

8. TEMPORAL ANALYSIS

8.1 Finding optimal lag

Next, we examine the correlation between egonet properties and artist quality as they vary over time. To do this, we compute each property, including our PPR^2 score for each year to capture the evolution of the properties over time. The annual datapoints are cumulative, so the properties for year Y are computed using data for movies released from 1990 (the beginning of our data) through year Y. The Spearman ρ for data broken out in this format is approximately the same as for the data when examined as a whole, though, not surprisingly, the predictive power is slightly less than when the data is analyzed as a whole.

Although we can't prove causality, we would like to know whether a change in egonet properties precedes a change in artist quality or vice versa, and the size of that lag, if any. To measure this, we run the Spearman test for lags ranging from 0 to 8 years. At each lag value, we compute the mean absolute correlation (MAC) for the properties at that lag. If we let ρ_{yi} be the quality correlation of property i when computed with a lag of y years, then our optimal lag L is computed as:

$$L = \arg\max_{y} \frac{1}{n} \sum_{i=0}^{n} |\rho_{yi}|$$

This results in an optimal lag of 3-4 years (3 years for audience ratings, 4 for critics ratings). That is, we observe egonet property changes 3-4 years before we observe the full impact of the corresponding artist quality change. The difference between the 0 year lag and 3 year lag is not enormous – it is on the order of 0.1 in the unitless Spearman $[-1, 1]$ scale – but it is measurable.

8.2 Verification with regression

Given this time lag, we would like to know whether we can predict future artist quality from present-day egonet properties. In some sense, this is an alternate verification of the lag result using Spearman's correlation coefficient, but it is also a predictive model for future artistic success.

To set up the test, for each lag value, we take the egonet properties from year Y and quality scores from year $Y + lag$. We train a model using linear regression with Lasso regularization using a randomly selected 75% of the data. Then we test our error using the remaining 25% of the data. We repeat this test 3 times for each lag value. The results confirm an optimal 3 year lag of quality changes after egonet changes, as seen in Figure 9.

8.3 Precise relationship of egonet properties to artist quality

Though we have established a correlation between egonet properties, we still do not know whether the relationship is linear, exponential, quadratic, or something else. Uzzi and Spiro find a parabolic effect between small world properties and artist quality in the theater community[5]. For their dataset, there is a "sweet spot" where a world is small, but not too small. We would like to know whether this holds true for the film community as well. As discussed in Section 6, many of our properties are closely related and will have similar or identical behaviors. In plotting the individual properties against artist quality, we find that there are only 3 families of plot shapes, which are either exponential or linear. None are parabolic (Figure 10). Again, our film community results diverge from previous results in the theater community.

8.4 Discussion of Results

The 3 year lag between the observed change in an artist's egonet and the change in their movie ratings is perhaps not surprising. One possible explanation is that their network improves, which gives them entrée into an improved project. Alternately, perhaps there is a latent change in the artist's abilities which effects both the egonet change and the later ratings bump. Other confounding factors could be present. For instance, perhaps the network structure somehow affects ratings in ways we cannot observe in our data, and which have nothing to do with an artist's film-making talents (e.g., improved public relations). As usual, we must point out that we have only proven correlation and cannot make any strong statements about true causation. However, all of our results point to Granger causality of of egonet properties on artist quality.

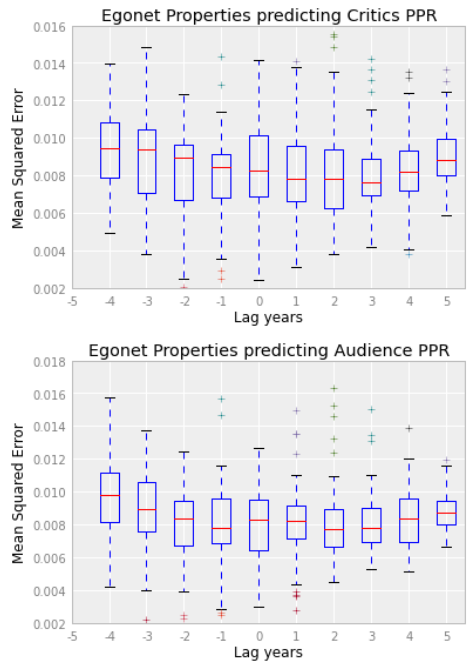

Figure 9: Mean-squared error for audience and critic regression using various lag values.

9. FUTURE WORK

We showed correlation and time series sequencing to arrive at a conclusion of Granger causality, but we did not run formal Granger causality tests against the data to get a p-value for that conclusion. This would be a logical next step in ensuring the validity of the result.

Additionally, the use of PageRank at all as an artist scoring mechanism presents a possible threat to validity in that the graph properties we are measuring might be inherently tied to our artist scoring measure. A round of tests using a completely graph-independent measure, such the average rating of movies to which the artist contributed, would be an effective way to allay that concern.

10. CONCLUSION

Our discovery of the strictly positive correlation between brokerage and high artist ratings came as a bit of a surprise, since we expected our findings to be closer to previous work on global network data from the musical theater community. These findings are, however, consistent with the theory of the "strength of weak ties". Artists in the film industry who straddle multiple communities seem to gain benefit from those associations, allowing them to create work that garners

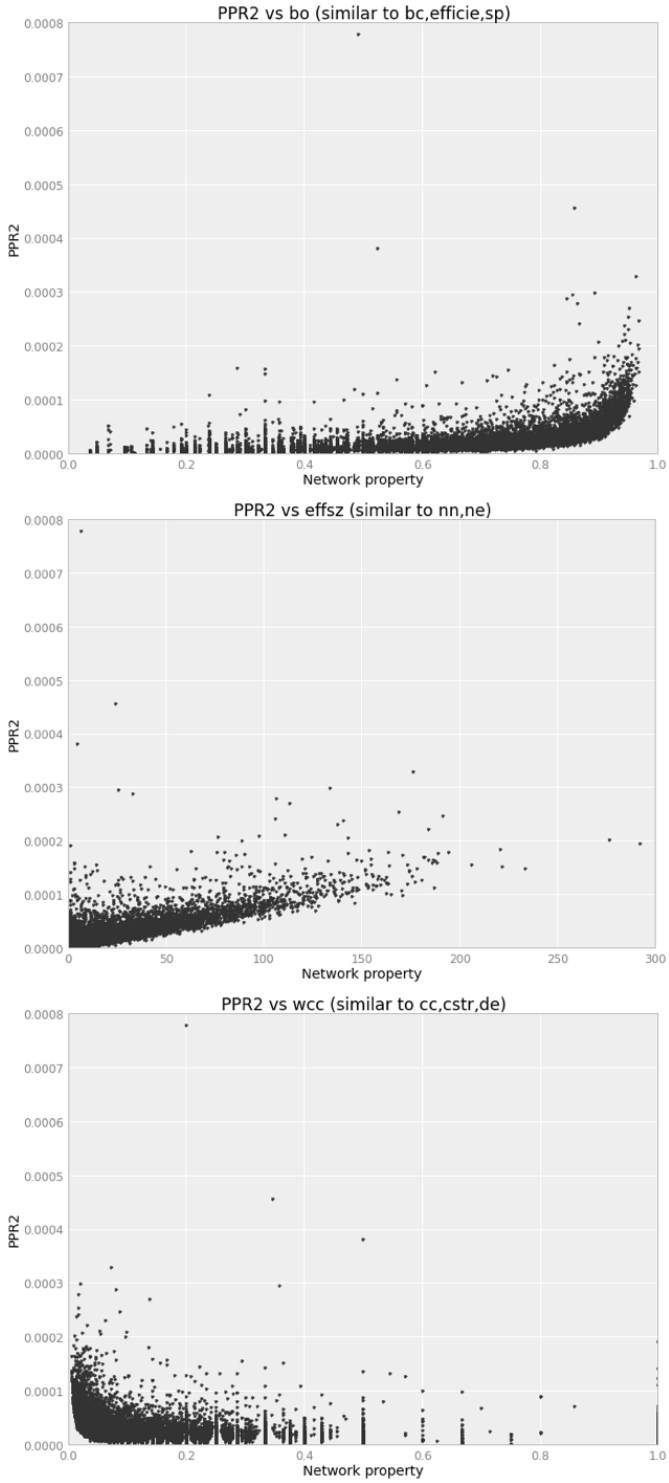

Figure 10: Individual egonet properties plotted against artist quality result in only 3 families of plots, none of which is parabolic.

more critical and popular acclaim than their more insulated counterparts.

11. REFERENCES

[1] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3):590–614, 2002.

[2] P. Chau and A. Goel. Rotten Tomatoes data set. Personal communication, Nov. 2013.

[3] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296. Association for Computational Linguistics, 2010.

[4] M. E. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001.

[5] B. Uzzi and J. Spiro. Collaboration and creativity: The small world problem. *American journal of sociology*, 111(2):447–504, 2005.

[6] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *nature*, 393(6684):440–442, 1998.