2013 IEEE Military Communications Conference

Red Black Network: Temporal and Topological Analysis of Two Intertwined Social Networks

Saurav Pandit*, Jonathan Koch*, Yang Yang*, Brian Uzzi† and Nitesh V. Chawla*
*Dept. of Computer Sc. & Engg.
Univ. of Notre Dame
†Kellogg School of Management
Northwestern Univ.

***Abstract*—In this paper we introduce and study the properties of certain kind of interdependent networks that we collectively call a *Red Black Network* – two intertwined social networks that work together towards a series of events (missions or performances). More specifically, members of one of the two networks is responsible for planning and organizing the events. They will generally be referred to as *artists*. Members of the other network, henceforth called *actors*, are responsible for the execution of the events. Using temporal data from the performing arts industry, we study the co-evolution of two such co-dependent social networks. We find that the statistical properties of two such networks are highly correlated, and use that finding to devise a prediction mechanism for such properties in a scenario when one of the two networks is invisible or only partially visible. This also sets up a framework for our ultimate goal of *temporal, semi-blind, multi-relational* link prediction.**

I. Introduction

Two related networks, made up of two different entities but heavily relying on each other, are abundant in nature. However the precise dynamics of how two such networks interact with and depend on each other are largely unknown to us. Often, they can rely on each other to such an extent that any perturbation in one network can spread and disrupt the other network. A perfect example is a case study of cascading power failure in Southern Italy. The network of power stations in Italy were controlled by a network of computer servers and the network of computers were powered by the power stations. Due to the nature of this inter-dependence, a small power failure had propagated to both networks such in an extent that caused blackout in almost half of Italy. We intend to study the inner workings of such *intertwined* networks. In this presentation we describe our initial findings, ongoing experiments and future goals.

To avoid further confusion, one subtle difference is to be made between *intertwined* networks and *multi-relational* networks (another hot research area these days). An example of a multi-relational network can be a network of students in a high school with different sets of edges among them. One set of edges could mean their friendship, another set of edges could mean phone conversations between them, physical contacts or a number of other things. Note that here the set of nodes remain the same. Whereas an intertwined network consists of two completely different networks somehow functioning together towards a common goal. We can see them as the union of two non-overlapping graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ connected by a set of cross-edges $E = \{(v_1, v_2) | v_1 \in V_1, v_2 \in V_2\}$. Note that intertwined networks have (at least) three different types of edges present and hence can be seen as multi-relational networks. At times we shall even treat them so, but largely it is important to keep in mind that we are dealing with networks that are more than simply multi-relational. Let us try to shed more light on exactly what type of interdependencies are of interest to us.

A. Motivation

Given that two networks can interact with each other in many different ways, it is not necessary that it will be possible to precisely quantify the interdependency across all intertwined networks. But nevertheless, we believe that the insight we gain from one pair of intertwined networks can be used and extrapolated to other *similar* intertwined networks. In particular, we intend to study two *social networks* that work together – one network of entities who do the planning and decision making, the other network of entities who execute the job. Due to lack of existing terminology (to the best of our knowledge), we call this combination of two intertwined networks a *Red Black Network*[1].

We can think of military mission design as an example. Military missions depend broadly on two groups of people – let us call them Intelligence team and Action team. An Intelligence team may include people involved in reconnaissance, collecting and analyzing data and strategizing. An Action team may include soldiers, specialists and support teams. Given enough data on past success or failure of these groups, it may be possible to gain valuable insight on what subsets of these groups work better together. Also note that the adversaries of the military or law-enforcement agencies also work in similar ways. If we look at the mafia, cartels or terrorist networks, we'll find that there are some people (usually in the background and hard to find) planning and masterminding operations, and others (relatively more visible and more prone to capture) who carry out the operations. On that context, we may be able to predict whether two individuals are likely to work together in the future and even predict the outcome of a future collaboration. Being able to manipulate this network structure to one's advantage and being able to predict future events in a timely manner are at the core of Counter-insurgency (COIN) strategies, where one needs to deal with such layered,

[1] The name “Red Black Network” is not related to the data structure “Red Black Trees”, but rather it reminds us of the intertwined red (Live) and black (Neutral) wires used in household electrical circuits.

978-0-7695-5124-1/13 $31.00 © 2013 IEEE
DOI 10.1109/MILCOM.2013.128

IEEE computer society

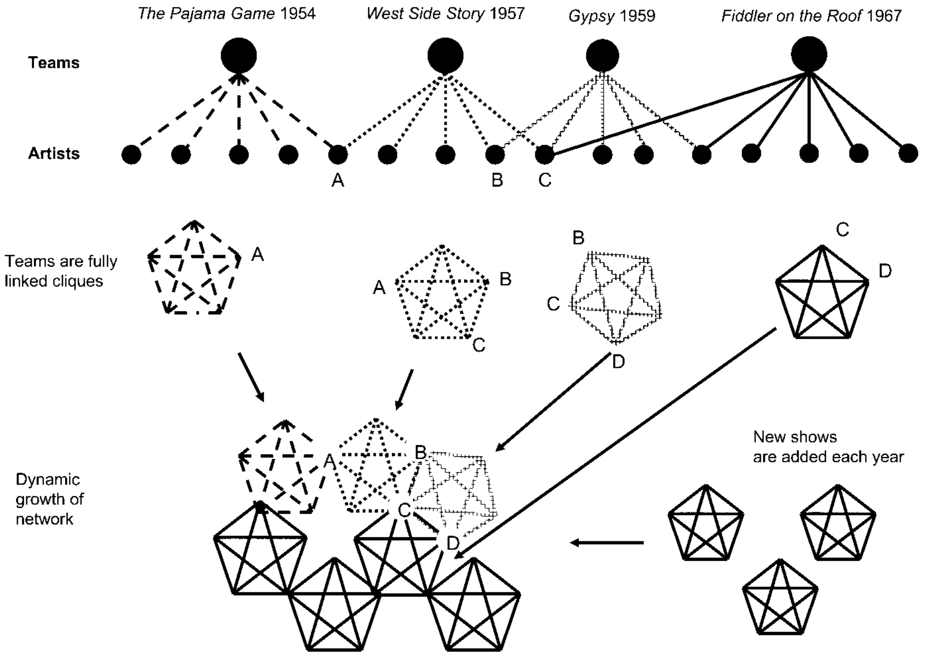

The Pajama Game 1954
West Side Story 1957
Gypsy 1959
Fiddler on the Roof 1967
Teams
Artists
A
B
C
Teams are fully linked cliques
A
A
B
C
B
C
D
C
D
Dynamic growth of network
A
B
C
D
New shows are added each year

Fig. 1: Broadway creative artist network. Figure is illustrative but based on actual data; A = Harold Prince (producer), B = Stephen Sondheim (composer/lyricist in *Gypsy* and lyricist in *West Side Story*), C = Arthur Laurents (librettist), and D = Jerome Robbins (director). As the fully linked cliques are connected to each other through artists who are part of multiple teams, the frequency of between-clique connections is disproportionately made up of repeated ties and third-party ties. This pattern is illustrated by the high connectivity among the artists who separately worked on *West Side Story*, *Gypsy*, and *Fiddler*, and the frequency of the repeated and third-party ties among B and C, and C and D, Sondheim and Laurents, and Laurents and Robbins.

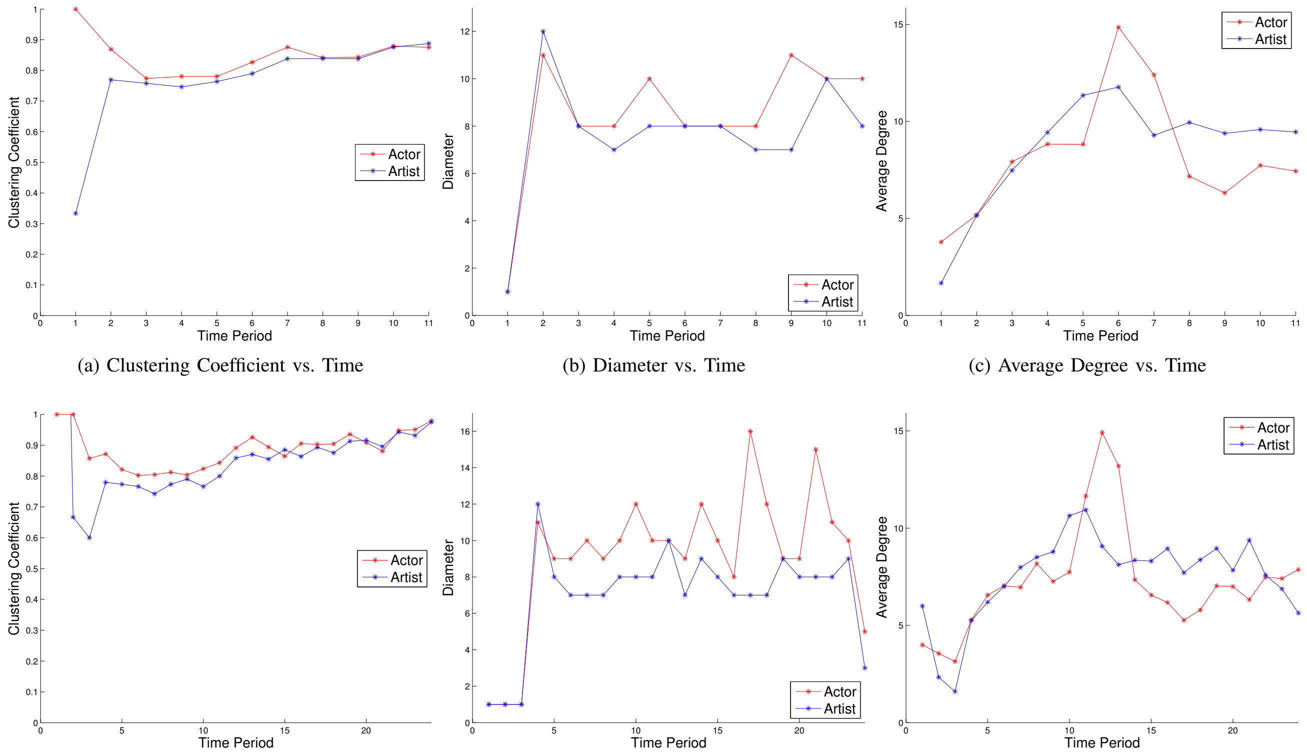
intertwined networks of leaders, insurgents and even non-combatants. We should also mention that such intertwined networks are not just limited to the military scope and, in fact, abundant in civilian societies as well, e.g. sports teams, performing arts groups etc.

From a network science perspective, we ask the following questions: How do the properties of the two networks evolve over time? Are statistical properties and/or events (e.g. link formation) in one network driven or influenced by the other network, and if so does it work both ways? If only one of the two networks is visible to us and the other one is hidden or partially hidden, what kinds of predictions can we make about the hidden network? In a Red Black Network, each edge or relationship represents a collaboration between two entities. We can use the history of outcomes of these collaborations, if available, and thus increase the types of relationships in the multi-relational network (by labeling the edges). Then we can further investigate if it is possible to use this additional information to make these predictions more precise.

II. The Broadway Dataset

The dataset used for our experiments relates to information collected for about 120 years (from 1880 to 2000), regarding musicals on or around the famous theater district of New York City – The Broadway. Courtesy: Brian Uzzi [1]. There are three main entities in this dataset: **musicals**, **actors** and **artists**. Naturally, artists can have various roles: writer, producer, director, choreographer etc. The social network for artists and the one for the actors constitute this particular Red Black Network. Each link in these social networks represents that the two involved nodes (artists or actors) worked together during some period of time for a musical. Hence the graph representation contains three different types of edges: **artist–artist**, **actor–actor** and **artist–actor**. The graph representing the complete Red Black Network basically consists of a set of overlapping cliques, each clique representing the artist and actor set of a musical production over a time period. Please see Figure 1 for an illustration.

One important feature of this dataset is that performance history of musicals are indeed available. Three performance metrics for the musicals are recorded: critics review (scale of 1 to 5), runtime (in days) and financial success (a binary indicator). Often critics notes are also available. For the purpose of data cleaning, these metrics for the performance of a musical is reduced to a *hit* or *fail* classifier. Due to missing data, mixed reviews and other reasons some performances cannot be classified as hit or fail. We put them in a third class between hit and fail. For our experiments we will ignore the original set of metrics and use the reduced performance metric



(a) Clustering Coefficient vs. Time

(b) Diameter vs. Time

(c) Average Degree vs. Time

(d) Clustering Coefficient vs. Time

(e) Diameter vs. Time

(f) Average Degree vs. Time

Fig. 2: Comparing topological statistics of the two networks over time. Periods of 10 Year intervals **(top)** and 5 Year intervals **(bottom)**.

with values *hit*, *average* or *fail*. As each edge in this network corresponds to a single musical production, the edges can also be attributed these performance metrics.

III. Statistical Similarities

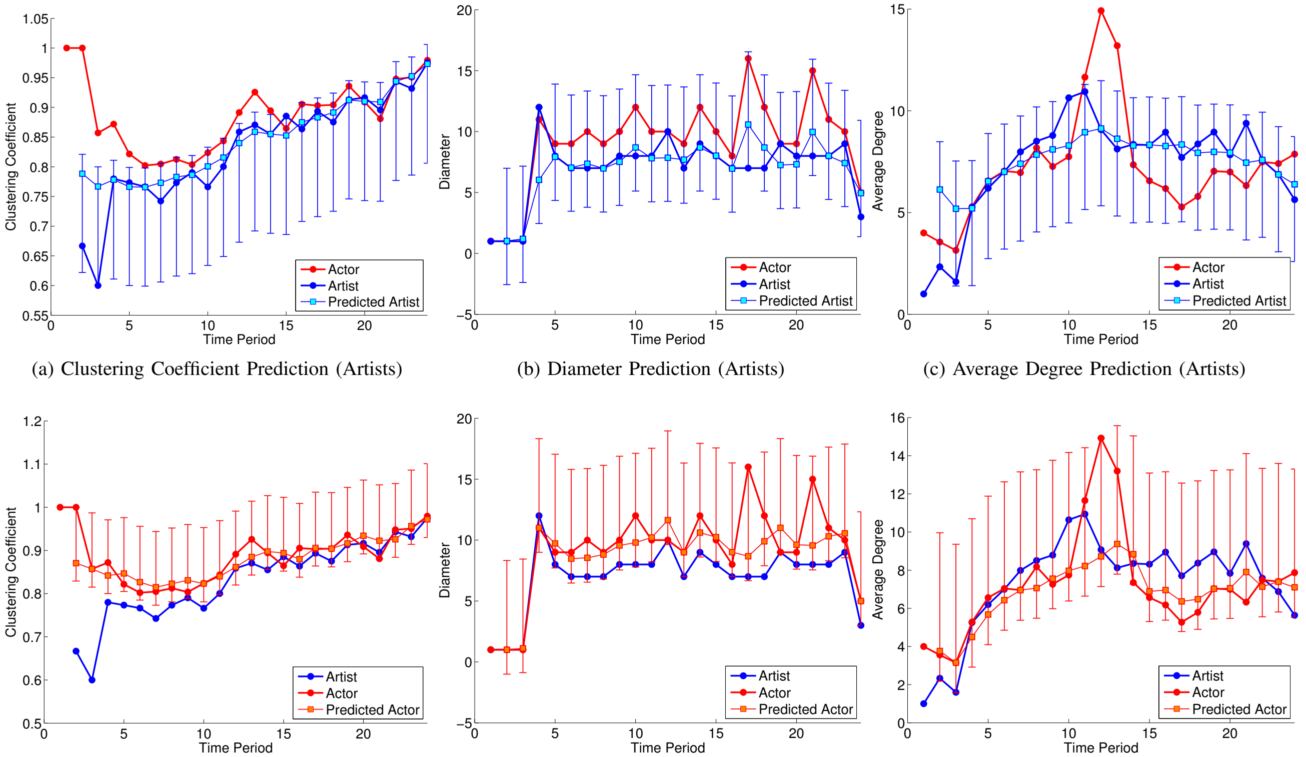
Our first set of experiments were to study how the two networks evolve over time and whether there is a significance correlation in changes between the two networks [11] [13]. We start this set of experiments by splitting the dataset temporally. Remember that the data spans over approximately 120 years. We run different instances of our experiments with the data split into 10 year, 5 year and 2 year periods.

We look at several statistical properties over time, such as Average Degree, Diameter, Clustering Coefficient, Eccentricity etc. Note that these properties only give a global picture of that graph. Hence, we have also looked at the node-property distributions (such as Degree Distributions, Clustering Coefficient Distributions) of the network snapshots and computed the Pearson Correlations over time. Furthermore, we have computed the Graphlet Degree Distribution (GDD) Agreement over time, that gives a much more accurate picture of local similarities and topological alignments.

For the sake of clarity, instead of presenting the results of the entire set of experiments (i.e. with 10 year, 5 year, 2 year periods, and with all the properties mentioned above), we present some representative results. All our conclusions were made from the full set of results and they will be published in a detailed version of this paper. However, the representative results should be enough to get the essence of our findings to the reader.

Let us turn our attention to Figure 2. The average clustering coefficient, diameter and degree are plotted over time respectively for 10 year periods and 5 year periods. The key observation from these plots (and other plots not shown here) is a positive temporal correlation. One would notice that even though the values for two networks may differ, they peak and bottom out, at least locally, in and about the same time periods. We see this positive correlation in all the statistical properties mentioned above but not shown here, except for one (eccentricity).

Note that this positive correlation in statistical properties, even though a strong evidence that changes in any of these networks influence each other, but not a conclusive proof. For example we cannot say if it is the case that one of the two networks is changing autonomously or independently and the changes in the second network is a consequence of the changes in the first network. Nor can we conclusively rule out the case that changes in both networks are results of external stress or stimuli, e.g. social and cultural changes, economic depressions, wars etc. That is an interesting question for social



(a) Clustering Coefficient Prediction (Artists)

(b) Diameter Prediction (Artists)

(c) Average Degree Prediction (Artists)

(d) Clustering Coefficient Prediction (Actors)

(e) Diameter Prediction (Actors)

(f) Average Degree Prediction (Actors)

Fig. 3: Predicting topological properties. Periods of 5 Year intervals. Successful 1-step-ahead predictions when the artist network is hidden **(top)** and when the actor network is hidden **(bottom)**. For example, when the artist network is hidden (**top**), at each point in the artist time-series (blue) we only have knowledge of the past values. We also have past and current knowledge of the actor time-series (red). Based on those information, predictions are made. We see that actual values (which are *unknown* at the time of prediction) fall within the error range of the prediction at 95% confidence level.

scientists and historians. Nevertheless, this positive correlation is essentially good news for us. It helps us build prediction frameworks, knowing that the changes in two networks, for whatever reason, follow each other. This framework and the results are described in the next section.

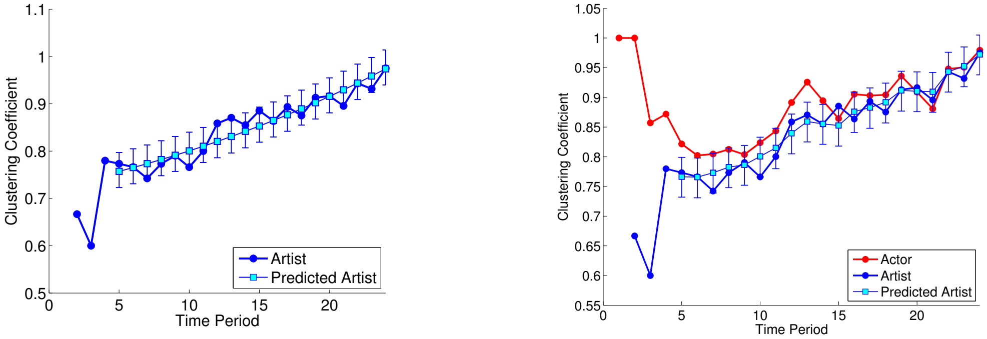
IV. Step-Ahead Prediction

Given the statistical similarities in the two networks over time, the next question we ask is: If we have complete information of one network and *some* history of the other network, can we predict the future statistical properties of the second network? More precisely, can we do a "step-ahead prediction" on the properties of the second network based on its history and the knowledge of the first network?

In order to perform step ahead predictions on the Broadway data, we used the forecasting plugin available in Weka 3.7.3. The plugin allowed us to predict one time series' behavior in the future, based on its own past behaviors and the characteristics of another time series. In our case, this meant it was used to predict one statistical property's behavior for one network (either **actor-actor** or **artist-artist**), given a starting point and the same data from the other network.

To achieve this type of analysis in Weka [2], the following settings were used: Basic configuration in the forecasting plugin would be set to target one property of a given network and output confidence intervals of 95%. Then, in the advanced configuration, the custom lag length was set to 1 (i.e. a "1-step-ahead" prediction), in order to supply the forecasting algorithm with only the starting point of the time series to be predicted. Overlay data was then selected in a way such that the same statistical property of the opposing network was used. Lastly, output predictions for both text and graphical output were set to the target property with no future predictions beyond the end of the series made. Figure 3 details the outcome of the experiments.

The specifications we presented, as well as the results in Figure 3 is for 1-step-ahead prediction. Predictions can also be made further into the future, e.g. we have performed experiment with 4-step-ahead predictions. With a 5 year period a 4-step is equivalent of 20 years. Note that further into the future a prediction is made, more difficult the process becomes without the help of the overlay series. However, we were able to improve the predictions significantly with the help of the overlay series. Please see Figure 4 for more details.



(a) Clustering Coefficient Prediction (Artists)

(b) Clustering Coefficient Prediction (Artists)

Fig. 4: Predicting Clustering Coefficient 4 steps ahead. Periods of 5 Year intervals. Without overlay **(left)** and with overlay **(right)**. Consider a time point in the series when the only data available are 20 years old CC information for the artists network. Based on that we try to predict the CC value of that network (**left**). We observe that, due to the lack of enough information, the predictions grow monotonously. On the **right**, we do the same predictions, but now with up-to-date data available about the actors network. In the later case, we are able to make more meaningful predictions. Similar results are seen on the other statistical properties as well.

V. Link Prediction

With very encouraging results on the similarity of two networks over time, to an extent where we are able to predict the future statistics of one network with information from the other network, we move on to build some link prediction schemes.

Most of our approaches of link prediction are based on measures for analyzing the "proximity" of pairs of nodes in the network. Higher proximity scores between nodes suggest larger probabilities that they have connection. Feature-based link prediction methods can be categorized as: (1) methods based on node neighbors; (2) methods based on ensemble of all paths. In category (1) there are several base line predictors, such as *Common Neighbors* (CN) [6], *Jaccard Coefficient* (JC), *Adamic Adar* (AA) [3] and *Preferential Attachment* (PA) [4], [5], which employ the nodes neighboring information to compute the proximity between them. In category (2) a number of methods refine the notion of shortest-path distance by implicitly considering the ensemble of all paths between two nodes, e.g. Katz [7] heuristic. However, the *PropFlow* (PF) [8] method has been shown to outperform the baseline predictors mentioned above and, till date, is one of the highest performing predictor known. In this paper we use PropFlow as a benchmark and build further on it.

A. Multi-relational Link Prediction

First, we reiterate that not only we have past data available about pairwise collaborations, but each of those pairwise collaborations are part of a musical production. And we have performance metrics available for the productions. As mentioned before, we have reduced the performance metrics into a *trinary* metric: $\{hit, average, fail\}$. Hence, not only that we know whether two individuals have worked together in the past, we also know the outcome of those collaborations. Hence we attempt to go beyond predicting whether two individuals will collaborate or not, and predict the outcome of such collaboration. In other words we attempt to predict not only the *formation* of a link, but also the *quality* of a link that is predicted to form.

We know that a Red Black Network such as the Broadway network has three types of edges: **actor-actor**, **artist-artist** and **artist-actor**. Now by labeling each edge as *hit*, *average* or *fail*, we can classify the edges into nine categories. For the multi-relational link prediction, we use an unsupervised scheme known as *Multi-relational PropFlow* (MRPF) [9]. Table I shows the AUROC results under PF and the improvement under MRPF.

TABLE I: **AUROC on Broadway Dataset**

Edge-type	PF	MRPF	Improvement
Artist Artist Hit	0.739	**0.753**	1.89%
Artist Artist Average	0.634	**0.690**	8.83%
Artist Artist Fail	0.789	**0.801**	1.52%
Actor Actor Hit	0.584	**0.803**	37.50%
Actor Actor Average	0.551	**0.635**	15.25%
Actor Actor Fail	0.532	**0.669**	25.75%
Actor Artist Hit	0.573	**0.825**	43.97%
Actor Artist Average	0.565	**0.747**	32.21%
Actor Artist Fail	0.727	**0.765**	5.23%

B. Temporal Link Prediction

The next piece of framework we put in place has to do with predicting the *timing* of the formation of a link. For this we use a supervised link prediction scheme (*Bagging with logistic regression*) [14] that takes into account timing issues like *activeness* of nodes or *recency* [14] [10] [12] of formed links. The AUROC results on **actor-actor** and **artist-artist** networks are shown in Tables II and III respectively.

TABLE II: **AUROC Comparison on Broadway Actors Collaboration Network**

Temporal Predictor	T_PA	T_CN	T_JC	T_AA	
AUROC	**0.709**	**0.685**	**0.712**	**0.712**	
Static Predictor	PA	CN	JC	AA	PF
AUROC	0.63	0.517	0.517	0.518	0.536

TABLE III: **AUROC Comparison on Broadway Artists Collaboration Network**

Temporal Predictor	T_PA	T_CN	T_JC	T_AA	
AUROC	**0.721**	**0.726**	**0.753**	**0.715**	
Static Predictor	PA	CN	JC	AA	PF
AUROC	0.67	0.561	0.559	0.535	0.62

VI. Conclusion

In this paper we present the Red Black Network - a network consisting of two intertwined social networks. In particular, we chose to look at the Broadway network in which the two entwined networks consisted of one of actors and one of artists; respectively, the players and the leaders if you will. We have been able to show via statistical similarities and step-ahead predictions on the Broadway network, that even when looking at the two social networks in isolation, they still tend to follow one another closely. For what reason that is, be it social stimuli, network properties, etc. we are not completely sure of yet but let us take a step back and think of what this means. Regardless of why one network tends to influence the other, this phenomenon is occurring, and because of that, we can try to build prediction frameworks to better predict these types of Red Black Networks. We knew that the network we looked at is temporal and multi-relational in nature, and thus were able to able to use link prediction models that took advantage of these network features to garner better results than normal.

Therefore, moving ahead one question is can we create a prediction scheme that utilizes all of these features of a Red Black Network listed above, in tandem, to produce optimal results? This may seem like a very pigeon-holed question considering it only relates to Red Black Networks, but another driving force behind this analysis is the abundance of which these networks exist in human society: sports teams, militaries, company management, product design, performing arts, the list goes on and on. Furthermore, in most cases, all of the examples listed here aim to be the best or most efficient in whatever industry they are apart of. This reason even further puts an onus on the importance of our continued research within this area, but also allows us to use the success rate data that goes along with Red Black Networks, to even better our Red Black Network link prediction approach.

Lastly, our findings and investigations have also led us to pursue one final area of research moving ahead, that being temporal, semi-blind, multi-relational link prediction. Imagine you are studying a Red Black Network and wish to predict how it evolves as time advances. More often than not, we are able to see the actors within the Red Black Network, because they are the ones who are carrying out all of the jobs/missions. At the same time, artists/leaders are usually hidden to the public eye. Therefore, we ask ourselves the question: with artist data hidden as we move forward in time, can we use future interactions between actors and past artist/actor interactions to predict how artist links will form in the future? Then to extend from that, is it possible to analyze artist cliques, artists influence over the network, and artist clique generation and modification into the future?

Acknowledgement

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] B. Uzzi, and J. Spiro, Collaboration and Creativity: The Small World Problem in American Journal of Sociology, Vol. 111, No. 2. (1 September 2005), pp. 447-504.

[2] I. H. Witten, and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San Francisco, California, USA, second edition, 2005.

[3] L. Adamic and E. Adar, Friends and neighbors on the web. Social Networks, 25:211-230, 2001.

[4] A.-L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaboration, Physica A, 311(3-4):590-614, 2002.

[5] M. E. J. Newman, Clustering and preferential attachment in growing networks, Physical Review Letters E, 64, 2001.

[6] M. Mitzenmacher, A brief history of lognormal and power law distributions, In Proceedings of the Allerton Conference on Communication, Control, and Computing, 2001.

[7] L. Katz, A new status index derived from sociometric analysis, Psychometrika, 18(1):39-43, 1953.

[8] R. Lichtenwalter, J. Lussier, and N. Chawla, New perspectives and methods in link prediction, in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010, pp. 243-252.

[9] R. Johnson,Y. Yang, E. Aguiar, A. Rider, and N. V. Chawla. ALIVE: A Multi-Relational Link Prediction Environment for the Healthcare Domain, Third Workshop on Data Mining for Healthcare Management, PAKDD 2012.

[10] A. Potgieter, K. April, R. Cooke and I. Osunmakinde, *Temporality in link prediction: Understanding social complexity*, Sprouts: Working Papers on Info. Sys. 2007.

[11] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert and T. Vicsek, *Evolution of the Social Network of Scientific Collaboration*, Physica A, 2002.

[12] Z. Huang and D. Lin, *The Time-Series Link Prediction Problem with Applications in Communication Surveillance*, INFORMS Journal on Computing, 2009.

[13] T. C. Mills, *Time Series Techniques for Economists*, Cambridge University Press, 1990.

[14] Yang Yang, Nitesh V. Chawla, Yizhou Sun, and Jiawei Han, Link Prediction in Heterogeneous Networks: Influence and Time Matters Proc. of the 12th IEEE International Conference on Data Mining (ICDM'12), Brussels, Belgium, Dec. 2012.