CS224W Project Proposal
Outcome Prediction in a Chess Tournament

Nopphon Siranart, Karen Truong, Vincent Pierre Berges

October 20, 2016

1 Introduction

In a tournament where individuals or teams are randomly drawn to compete with others, predicting the outcomes of a game between contestants who are yet to play against each other based on historical results is a challenging problem. Typical methods currently used include ranking players by the total points or percentage of game wins. This way of predicting is generally inaccurate as it is possible that top players in the ranking might have only faced weak opponents. In this project, we we formulate the problem of predicting the outcomes as a graph network prediction problem. We first examine and discuss a number of papers on the topic that might be useful before discussing our proposal in greater detail.

2 Literature Review

In our search for existing literature on predicting competition outcomes, we first reviewed a paper covering PageRank, a common ranking algorithm, where we want to predict outcomes based on the ranking of players computed from previous results. To brainstorm a more probabilistic approach to anticipating the result of pairwise comparisons, we explore possible variations of the Bradley-Terry model, before going onto think more about the theory behind signed, directed-edge networks, which look at existing connections to common neighbors to perform link prediction. More details about how we can apply each of these techniques to our project will be discussed in the next section.

2.1 PageRank Algorithm

- **Generalizing Googles PageRank to Rank National Football League Teams (Govan, Meyer, Albright, 2008) [2]**

  This article goes over different ranking algorithms for nodes into a graph including the well known PageRank. Using multiple ranking algorithms and previous game results, we can estimate the strength of a team. The article uses the example of football teams. Based on the history of settled scores between teams, we can determine which team is stronger than the other. Once the score is assigned comparing two teams is easy, the team with the highest rank is predicted to beat the team with lowest rank. There might be improvements to this approach. For example, we could relate the ranks directly to a probability or combine multiple ranks into one.

One thing missing from this paper is that, although they present multiple ways of ranking football teams and talks briefly about the strength and the weakness of each model, they do not evaluate the performance of the models and do not make comparison between each of them explicitly. We have no idea after reading this paper which method we should use and in what situation. This paper also assumes that there is no tie between the teams. Hence, it can be applied only to a game with no draw.

- **Other Papers**

  We get motivations from these two papers: *PageRank Beyond the Web* by David F. Gleich [1] and *Who Is the Best Player Ever? A Complex Network Analysis of the History of Professional Tennis* by Filippo Radicchi [4]. The first one discusses how PageRank can be applied into other domains ranging from bibliometrics, social and information network analysis, link prediction and recommendation to systems analysis of road networks, as well as biology, chemistry, neuroscience, and physics. This gives us a general idea of how applicable PageRank is. In the second paper, it discusses how we can use PageRank to rank tennis players. This paper uses a method similar to [2], but it also compares its ranking system with the real-world ranking with nice visualization.

2.2 Signed Network Prediction

- **Signed Networks in Social Media (Leskovec, Huttenlocher, Kleinberg, 2010) [3]**

  This article uses triads, configurations of 3 nodes, to make predictions on the signs and direction of edges. Signed networks can be analyzed through two perspectives: the structural balance theory, which models edges as likes and dislikes, depending on the sign; and status theory, which interprets edges as whether the source node outranks the destination node. For the interest of our project brainstorm, we choose to restrain our review of this paper to the analysis of status of the different nodes of the graph. By analysing the triads in the graph, one can determine the fraction of each of the four types of triads. In a graph where we have two nodes A and B who both have some directed, signed edge to a set of nodes X, we have context about A and B. When given a new unknown edge between A and B, i.e. the "closing edge", one can make a prediction of what direction this edge will have - or equivalently what kind of triad A, B, and X will form - by using the prior probability distribution of triads over the rest of the graph.

  One shortcoming of the signed network construction in this paper is the absence of triad configurations that include neutral edges. In social media, a neutral edge might be interpreted as saying the two persons know each other yet do not consider themselves either friends or foes. In the status example of voting used throughout the article, this could be considered as an abstention. If we consider that a "neutral edge" between two nodes means that the two nodes have similar status, there are no longer only four types of triads, there are new possible configurations and we can assign a probability of direction to an edge based on the number of triads corresponding to each configurations the edge is connected to. By introducing a new edge type, we can improve on the existing research since we can now represent a wider range of relationships between the nodes, which might bring to light interesting dynamics in the network.

Another improvement we could try is an extension from triads to systems involving four, maybe even 5 nodes to make predictions on the direction of an edge. In the paper, context between A and B was defined to be nodes X with whom and A and B have links. In this augmented case where we look beyond the triad, the contextualized links that would be involved in the prediction extends to include additional nodes where the node may not have a link to both A and B, but perhaps node X has a link to A and Y and node Y has a link to B, so there is more context for the prediction.

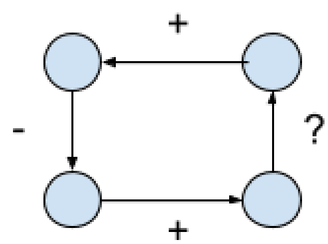

Figure 1: Prediction in a 4-node graph

2.3 Bradley-Terry Model

- **Ties in Paired-Comparison Experiments: A Generalization of the Bradley-Terry Model (Rao and Kupper, 1967) [5]**

  Instead of establishing a discrete prediction for whether or not node i outranks node j, we can formulate the problem as calculating the likelihood of i beating j. Probabilistic prediction of superiority in rankings has been previously studied by statisticians. One model for pairwise comparison is the Bradley-Terry model. Based on known comparisons of different pairs, the Bradley-Terry model uses maximum likelihood estimation to determine the probability that item i "beats out" item j. The model defines $P(i\ beats\ j)$, the probability that i outranks j, as $\frac{\pi_i}{\pi_i+\pi_j}$, where each π_i, π_j are variables representing the "merit" of i and j, or an index of preference, respectively. One variation of this takes the logit of the fraction, where π is an exponential score function, resulting in the simplified expression for $\Pr(ibeatsj) = V_i - V_j$, a log-linear interpretation. (The difference $V_i - V_j$ between responses should have a logistic distribution).

  In this particular paper, Rao and Kupper propose a modification to the Bradley-Terry model to allow for ties, if the magnitude of the observed difference between i's merit and j's merit is below a certain threshold $\eta = \ln\theta$. In their development of this threshold, the formula assumes that each pairwise comparison has r observations. The modified probabilities that account for ties become:

$$P(i|i,j) = P(Z_ij > \eta) = \frac{i}{i+j}$$

$$P(j|i,j) = P(Z_{ij} < -\eta) = \frac{\pi_i}{\theta\pi_i + \pi_j}$$

$$P(0|i,j) = P(|Z_{ij}| < \eta) = \frac{(\theta^2-1)\pi_i\pi_j}{(\theta\pi_i+\pi_j)(\pi_i+\theta\pi_j)}$$

where $Z_{ij} = V_i - V_j\ \theta$ is determined iteratively using Eq. 3.5 in the paper while also adjusting the maximum likelihood estimators in Eq. 3.4 [5]

One disadvantage of this approach is that the parameters are dependent on a logistic (squared hyperbolic secant) density distribution of the items and the past judgements on them, but it is unrealistic to assume that data can obey such a curve. It may be interesting to apply this method when the item is represented as a node in a network and "merit" parameter π_i of each item is derived from the node's network properties. There have been other papers that have applied the Bradley-Terry model to predicting the outcome of sports tournaments, but little work has been done towards applying Bradley-Terry at a large scale, for instance, networks with thousands of edges and nodes.

3 Project Proposal

Based on our previous literature reviews and discussions, we would like to apply network analysis models and algorithm to tackle the problem of predicting the outcomes of a matchup between players in a tournament. There are two things that we would like to accomplish in this project. First, we would like to have a model to predict the result of any matchup between players that achieve higher accuracy than the baseline model. Second, we would like to have a probability model that will output the likelihood of the game results. This can be used in various applications. For instance, we can use the model to predict ties or a chance that each player will win and build a recommendation system to match two players based on their preferences. We can also apply this model to calculate the fair odd for game betting. While we apply the techniques to predict outcomes, it can be used on any situation which could be represented as a weighted directed graph.

3.1 Problem Formulation

We formulate the problem of predicting the outcomes as a graph network problem where the nodes represent the players and node A points to node B if player B won a match between A and B with weight equal to 1. In case of a tie, there will be two directional edges connecting A and B in both directions with a half weight each. If there are multiple games played between A and B, the total weight will be computed by summing up all the weights of the edge from A to B. Hence, all the edges will have positive weights. The goal is to predict the direction of the edge as well as its probability between two disconnected nodes. We survey and evaluate several network analysis methods that could be used for prediction. First, we introduce a ranking algorithm where we applies ideas from the PageRank algorithm to rank players based on their results. Second, we analyze the relations between players in the tournament and use triads to make predictions on the direction of edges.

3.2 Dataset

The datasets we use are real historical data provided by the world chess federation (FIDE) available through Kaggle competition. They contain 54k chess players (nodes) over 11 year period and over 1.84 million games (edges). The datasets are in a nice format. Each row contains a gameID, two player IDs, month, and result. In addition, they provide the ratings of a subset of the 54k chess players that we can use as a baseline to compare our methods against.

3.3 Methods and Algorithms

- **Baseline Model** - This is a widely used simple model that we will use as a baseline to compare other models against. We will first compute the percentage of game wins for each player from the previous results. When we want to predict the outcome of a game between any two players, we will compare the players ratings and the player with a higher rating will be the one that this model predicts to win.

- **PageRank Algorithm** - We will use the original PageRank Algorithm with possibly some modifications to compute the ranking of players from the previous results. Hence, we can determine the stronger team based on the ranking. The player with a higher rank is predicted to beat the one with the lower rank. Another possible method which we have not thought out completely is to do ranking of players similar to PageRank, but using a centrality measure to rank instead. For instance, intuitively, Katz centrality seems like an appropriate measure. One nice thing about having score assigned to each player is that we can use this to calculate the probability model by computing the score difference between two players and with this difference, how much likely that the game will be based on historical data. This can be done by doing regression between score and proportion of wins, draws and losses.

- **Bradley-Terry model** - This is an appropriate model particularly for a chess game because it supports ties in paired comparisons. It also tells us the probability of the outcomes. However, because we have to deal with ties, we cannot use the simplified Bradley-Terry equation with logit, since $\Pr(j > i) \neq 1 - \Pr(i > j)$. Potential parameters we could use individually or combined for merit π include win percentage, raw number of wins, win-loss ratio (different from win percentage because of draws) and number of months played. One other variation in our data that might not make the Bradley-Terry model as straightforward as we expect is the fact that players can play each other more than once, yet they do not play every player the same amount of times, whereas in the Rao & Kupper paper each pair that was judged was judged exactly r times

- **Signed Network Prediction** - In our datasets, the edges are directed but do not have positive or negative values, but we could offset the data so that a win by player i over player j is a positive edge and loss is a negative edge from i to j. In the paper on status theory in signed networks [3], the analysis operates over all possible triad configurations for directed signed edges, but our analysis will require the introduction of the neutral edge to represent chess matches that resulted in ties. Counting up the triads of each configuration in the network, we can use the normalized frequency of a particular triad configuration as a probability measure of what kind of closing edge we expect to see, given the contextualized links. In the context of chess matches, the closing edge is the unseen chess match whose outcome we are trying to predict.

4 Evaluation

We will hold out the last two months of matches containing about 10% of the data for validation set and another 10% for test set. Validation set will be mainly used for hyperparameter tuning. We then compare our predictions with ground truth. We can then vary the decision criterion and plot the ROC curve of the predictive algorithm. Multiple binary classification measures will be used to compare the performances of the models. For instance, Balanced Error Rate (BER),

References

[1] D. Gleich. PageRank Beyond the Web. *arXiv:1407.5107*, 2014

[2] A. Govan, C. Meyer, R. Albright. Generalizing Googles PageRank to Rank National Football League Teams. *SAS Global Forum 2008, Data Mining and Predictive Modeling*, 2008

[3] J. Leskovec, D. Huttenlocher, J. Kleinberg. Signed Networks in Social Media. *In Proc. CHI*, 2010.

[4] F. Radicchi. Who Is the Best Player Ever? A Complex Network Analysis of the History of Professional Tennis *Journal: Plos One*, 2011

[5] P. V. Rao and L. L. Kupper. Ties in Paired-Comparison Experiments: A Generalization of the Bradley-Terry Model *Journal of the American Statistical Association* Vol. 62, No. 317, pp. 194-204, 1967.