The Science of Ranking and Rating

Dan Lanoue

October 13, 2014
The Goal

Mathematical models for rating and predictions in sports.

Figure 1: NBA Western Conference Power Rankings, 2013-14 [5]
Ranking vs. Ratings

For our purposes, the distinction is as follows:
For our purposes, the distinction is as follows:

- **Rating**: Quantifying the ‘strength’ of a team.
Ranking vs. Ratings

For our purposes, the distinction is as follows:

- **Rating**: Quantifying the ‘strength’ of a team.
  - ex. The Lakers have an Off. Efficiency rating of 101.9.
Ranking vs. Ratings

For our purposes, the distinction is as follows:

- **Rating**: Quantifying the ‘strength’ of a team.
  - ex. The Lakers have an Off. Efficiency rating of 101.9.
- **Ranking**: Teams ordered by their rating.
For our purposes, the distinction is as follows:

- **Rating**: Quantifying the ‘strength’ of a team.
  - **ex.** The Lakers have an Off. Efficiency rating of 101.9.

- **Ranking**: Teams ordered by their rating.
  - **ex.** The Lakers have the 10th best offense.

In general ratings carry more information than their corresponding rankings, so we will focus primarily on them.
For our purposes, the distinction is as follows:

- **Rating**: Quantifying the ‘strength’ of a team.
  - *ex*. The Lakers have an Off. Efficiency rating of 101.9.

- **Ranking**: Teams ordered by their rating.
  - *ex*. The Lakers have the 10th best offense.

In general ratings carry more information than their corresponding rankings, so we will focus primarily on them.

Note that this distinction in terminology is not universal (e.g. PageRank not PageRating).
Ultimately our goal is to **evaluate teams** and **predict future games**. The most basic rating is simply a team’s win/loss record. There are a few problems though:
Ultimately our goal is to **evaluate teams** and **predict future games**. The most basic rating is simply a team’s win/loss record. There are a few problems though:

- Short or incomplete seasons mean a small data set.
Ultimately our goal is to **evaluate teams** and **predict future games**. The most basic rating is simply a team’s win/loss record. There are a few problems though:

- Short or incomplete seasons mean a small data set.
- Not all wins are equal - wins and losses don’t tell the whole story. Primary here is the notion of **strength of schedule**.
Ultimately our goal is to **evaluate teams** and **predict future games**. The most basic rating is simply a team’s win/loss record. There are a few problems though:

- Short or incomplete seasons mean a small data set.
- Not all wins are equal - wins and losses don’t tell the whole story. Primary here is the notion of **strength of schedule**.

In fact, as we’ll see in an example, records can be woefully inadequate in finding the best teams.
The Bowl Championship Series (BCS) rankings determined who were the ‘best’ college football teams over a season and who got to play in the prestigious (and hugely profitable) BCS Bowl games and for a National Championship.
The Bowl Championship Series (BCS) rankings determined who were the ‘best’ college football teams over a season and who got to play in the prestigious (and hugely profitable) BCS Bowl games and for a National Championship. The BCS has been replaced with a playoff system starting this Fall, but the same ranking issues continue.
The main issue is that there are too many FBS teams (126 as of 2013) relative to how many games each team plays (≈ 12) to simply rank teams by the standings.
The main issue is that there are too many FBS teams (126 as of 2013) relative to how many games each team plays (≈ 12) to simply rank teams by the standings.

Teams also play in different conferences with wildly different strengths so comparing strength of schedule is nearly impossible.
To illustrate let’s look at picking two teams for the National Championship for the 2008 season.

<table>
<thead>
<tr>
<th>TEAM</th>
<th>RECORD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Florida</td>
<td>12-1</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>12-1</td>
</tr>
<tr>
<td>Texas</td>
<td>11-1</td>
</tr>
<tr>
<td>Alabama</td>
<td>12-1</td>
</tr>
<tr>
<td>USC</td>
<td>11-1</td>
</tr>
<tr>
<td>Penn State</td>
<td>11-1</td>
</tr>
<tr>
<td>Utah</td>
<td>12-0</td>
</tr>
<tr>
<td>Texas Tech</td>
<td>11-1</td>
</tr>
<tr>
<td>Boise State</td>
<td>12-0</td>
</tr>
<tr>
<td>Ball State</td>
<td>12-1</td>
</tr>
</tbody>
</table>

**Figure 2:** Undefeated and one loss teams, 2008[7]
2008 Season

On record alone Utah and Boise state, as the only undefeated teams, should compete for the championship.
2008 Season

On record alone Utah and Boise state, as the only undefeated teams, should compete for the championship.

However (at the time) Utah was in the Mountain West Conference (MWC) and Boise State the Western Athletic Conference (WAC), both considered substantially weaker than the AQ conferences such as the SEC which Florida had just won.
Ultimately (and controversially) Florida and Oklahoma were selected for the BCS Championship, with Florida winning 24 – 14.
Ultimately (and controversially) Florida and Oklahoma were selected for the BCS Championship, with Florida winning 24 – 14.

Utah ended up in the Sugar Bowl and beat Alabama 31 – 17.
Ultimately (and controversially) Florida and Oklahoma were selected for the BCS Championship, with Florida winning 24 – 14.

Utah ended up in the Sugar Bowl and beat Alabama 31 – 17. Boise State was left out of the BCS Bowls all together and lost to 16 – 17 to TCU (also non-AQ) in the Poinsettia Bowl.
The BCS Rankings were given by averaging three other rankings:
How did the BCS Rankings work?

The BCS Rankings were given by averaging three other rankings:

- The Harris Interactive Poll of former players, coaches, administrators and sports writers.
How did the BCS Rankings work?

The BCS Rankings were given by averaging three other rankings:

- The Harris Interactive Poll of former players, coaches, administrators and sports writers.
- The Coaches Poll.
How did the BCS Rankings work?

The BCS Rankings were given by averaging three other rankings:

- The Harris Interactive Poll of former players, coaches, administrators and sports writers.
- The Coaches Poll.
- The Computer Ranking. Actually an average of 4 out of 6 computer rankings.
How did the BCS Rankings work?

The BCS Rankings were given by averaging three other rankings:

- The Harris Interactive Poll of former players, coaches, administrators and sports writers.
- The Coaches Poll.
- The Computer Ranking. Actually an average of 4 out of 6 computer rankings.

For our purposes the computer rankings are the most relevant.
Since 2004 these rankings were forbidden to consider factors like margin of victory. Some of the ranking methods used [11] were:

- Sagarin's ranking, an Elo method.
- Colley Matrix, based on Bayesian inference.
- Massey Rating, which computes an offensive and defensive ranking based on points scored and allowed.

Figure 3: Artist's Rendering of the ‘BCS Computer’
BCS Computer Rankings

Since 2004 these rankings were forbidden to consider factors like margin of victory. Some of the ranking methods used [11] were:

- Sagarin’s ranking, an Elo method.

Figure 3: Artist’s Rendering of the ‘BCS Computer’
BCS Computer Rankings

Since 2004 these rankings were forbidden to consider factors like margin of victory. Some of the ranking methods used [11] were:

- Sagarin’s ranking, an Elo method.
- Colley Matrix, based on Bayesian inference.

Figure 3: Artist’s Rendering of the ‘BCS Computer’
BCS Computer Rankings

Since 2004 these rankings were forbidden to consider factors like margin of victory. Some of the ranking methods used [11] were:

- Sagarin’s ranking, an Elo method.
- Colley Matrix, based on Bayesian inference.
- Massey Rating, which computes an offensive and defensive ranking based on points scored and allowed.

Figure 3: Artist’s Rendering of the ‘BCS Computer’
Even given a larger number of games relative to the number of teams (and ‘irreducibility’), win/loss record still may not tell the whole story.

For example a teams current record may correlate poorly with its future record, compared to other metrics.
An Example: Pythagorean Record

Figure 4: The Bill James Baseball Abstract, 1982

Baseball is all about scoring runs (or possibly pitching/hitting, etc.). But sometimes the timing of runs is just as important as how many. Wins and losses it turns out are less meaningful predictively than runs scored.
An Example: Pythagorean Record

Say that in a 3 game week:

The Red Sox win 2−1, lose 0−6 and win 4−2.
The Yankees lose 2−3, lose 2−3 and lose 2−3.

Which team was better?
Say that in a 3 game week:

- The Red Sox win 2 – 1, lose 0 – 6 and win 4 – 2.
Say that in a 3 game week:

- The Red Sox win 2 – 1, lose 0 – 6 and win 4 – 2.
- The Yankees lose 2 – 3, lose 2 – 3 and lose 2 – 3.
An Example: Pythagorean Record

Say that in a 3 game week:

- The Red Sox win 2 – 1, lose 0 – 6 and win 4 – 2.
- The Yankees lose 2 – 3, lose 2 – 3 and lose 2 – 3.

Which team was better?
An Example: Pythagorean Record

Bill James estimated a team’s ‘expected win percentage’ as

\[
\text{Win} = \frac{\text{Runs Scored}^2}{\text{Runs Scored}^2 + \text{Runs Allowed}^2}.
\]
Bill James estimated a team’s ‘expected win percentage’ as

\[ \text{Win} = \frac{\text{Runs Scored}^2}{\text{Runs Scored}^2 + \text{Runs Allowed}^2}. \]

Actually an exponent of 1.83 turns out to be more accurate than 2. In other sports, other exponents \( e \) are commonly used: for Basketball \( e = 13.91 \), for Football \( e = 2.37 \), for hockey \( e \approx 2 \).
Bill James estimated a team’s ‘expected win percentage’ as
\[
Win = \frac{\text{Runs Scored}^2}{\text{Runs Scored}^2 + \text{Runs Allowed}^2}.
\]

Actually an exponent of 1.83 turns out to be more accurate than 2. In other sports, other exponents $e$ are commonly used: for Basketball $e = 13.91$, for Football $e = 2.37$, for hockey $e \approx 2$.

The difference between actual wins and expected wins is (in a first approximation) a team’s luck.
In this talk we will:
In this talk we will:

- Introduce two general rating methodologies: Elo and PageRank,
In this talk we will:

- Introduce two general rating methodologies: Elo and PageRank,
- Two very specific rating methods for the NFL: DVOA and EPA,
In this talk we will:

- Introduce two general rating methodologies: Elo and PageRank,
- Two very specific rating methods for the NFL: DVOA and EPA,
- Compare these methods and what that tells us about the science of ratings.
In this talk we will:

- Introduce two general rating methodologies: Elo and PageRank,
- Two very specific rating methods for the NFL: DVOA and EPA,
- Compare these methods and what that tells us about the science of ratings.

I’ve chosen the NFL as a recurring example I find the most interesting. In particular the NFL is a nice balance between the moneyball ‘science’ of baseball and the ‘Wild West’ of college football.
Elo Rating

Invented by the physicist Arpad Elo as a chess rankings system.
Elo Rating

Invented by the physicist Arpad Elo as a chess rankings system.

- Iterative, self-correcting rating.
Invented by the physicist Arpad Elo as a chess rankings system.

- Iterative, self-correcting rating.
- Made famous in *The Social Network* as the precursor to Facebook, known as Facemash.
Elo Rating

The basic idea:
Elo Rating

The basic idea:

- Players (unknown, intrinsic) ratings are normally distributed.
The basic idea:

- Players (unknown, intrinsic) ratings are normally distributed.
- Estimate these ratings, then refine over time after observing results.
How the Elo Rating works:

\[
E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}
\]

\[
E_B = \frac{1}{1 + 10^{(R_A - R_B)/400}}
\]

Note that always \(E_A + E_B = 1\).
How Elo Works

How the Elo Rating works:

- Each player $A$ has a current rating $R_A$
How the Elo Rating works:

- Each player $A$ has a current rating $R_A$
- If player $A$ has rating $R_A$ and player $B$ has rating $R_B$ then $A$’s **expected score** is

$$E_A = \frac{1}{1 + 10^{(R_B-R_A)/400}}$$

and similarly for $B$. 
How Elo Works

How the Elo Rating works:

- Each player A has a current rating \( R_A \)
- If player A has rating \( R_A \) and player B has rating \( R_B \) then A's expected score is

\[
E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}
\]

and similarly for B.

- Note that always

\[
E_A + E_B = 1.
\]
How Elo Works

When $A$ and $B$ play, let $S_A$ represent what $A$ actually scored (in chess this is 0 for a loss, 1 for a win and $\frac{1}{2}$ for draw).
How Elo Works

When A and B play, let $S_A$ represent what A actually scored (in chess this is 0 for a loss, 1 for a win and $\frac{1}{2}$ for draw).

- Then A’s updated rating $R'_A$ is given by

$$R'_A = R_A + K(S_A - E_A),$$

for some factor $K$. 
How Elo Works

When A and B play, let $S_A$ represent what A actually scored (in chess this is 0 for a loss, 1 for a win and $\frac{1}{2}$ for draw).

- Then A’s updated rating $R'_A$ is given by

$$R'_A = R_A + K(S_A - E_A),$$

for some factor $K$.

- We similarly update player B’s rating.
How Elo Works

When $A$ and $B$ play, let $S_A$ represent what $A$ actually scored (in chess this is 0 for a loss, 1 for a win and $\frac{1}{2}$ for draw).

- Then $A$’s updated rating $R'_A$ is given by

$$R'_A = R_A + K(S_A - E_A),$$

for some factor $K$.
- We similarly update player $B$’s rating.
- Note that

$$R'_A + R'_B = R_A + R_B,$$

preserving the total average rating.
How Elo Works

Some more details:

We can start with any initial configuration, average rating or spread. The 400 here is arbitrary, but chosen so a 400 point advantage corresponds to a $10 \times$ win probability.

The $K$ factor (relative to the spread) does matter and can be chosen differently based on the sport and even for different types of matches within a sport.

In chess a typical $K$-factor is $K \approx 24$, but different rankings use varying $K$'s.

Elo is non-commutative, i.e. the order of the matches matters.
How Elo Works

Some more details:

- We can start with any initial configuration, average rating or spread. The 400 here is arbitrary, but chosen so a 400 point advantage corresponds to a $10 \times$ win probability.
How Elo Works

Some more details:

- We can start with any initial configuration, average rating or spread. The 400 here is arbitrary, but chosen so a 400 point advantage corresponds to a $10 \times$ win probability.

- The $K$ factor (relative to the spread) does matter and can be chosen differently based on the sport and even for different types of matches within a sport.
How Elo Works

Some more details:

- We can start with any initial configuration, average rating or spread. The 400 here is arbitrary, but chosen so a 400 point advantage corresponds to a $10 \times$ win probability.

- The $K$ factor (relative to the spread) does matter and can be chosen differently based on the sport and even for different types of matches within a sport.

- In chess a typical $K$-factor is $K \approx 24$, but different rankings use varying $K$’s.
Some more details:

- We can start with any initial configuration, average rating or spread. The 400 here is arbitrary, but chosen so a 400 point advantage corresponds to a $10 \times$ win probability.
- The $K$ factor (relative to the spread) **does** matter and can be chosen differently based on the sport and even for different types of matches within a sport.
- In chess a typical $K$-factor is $K \approx 24$, but different rankings use varying $K$’s.
- Elo is non-commutative, i.e. the order of the matches matters.
Pros/Cons of Elo

Pros:

Simple to implement/adjust.
Allows for a player’s intrinsic ratings to change over time.
Convergence of ratings over time.

Cons:
Single factor analysis, may be less accurate for team sports.
Even in sports where a ‘single factor’ is reasonable - e.g. chess - ratings tend to not actually be normally distributed.
In particular in chess, weaker players statistically tend to have higher win probabilities (and thus Elo rankings) than predicted.

Dan Lanoue
The Science of Ranking and Rating
Pros/Cons of Elo

Pros:

- Simple to implement/adjust.
Pros/Cons of Elo

Pros:

- Simple to implement/adjust.
- Allows for a player’s intrinsic ratings to change over time.

Cons:

- Single factor analysis, may be less accurate for team sports.
- Even in sports where a ‘single factor’ is reasonable - e.g. chess - ratings tend to not actually be normally distributed.

In particular in chess, weaker players statistically tend to have higher win probabilities (and thus Elo rankings) than predicted.
Pros/Cons of Elo

Pros:

- Simple to implement/adjust.
- Allows for a player’s intrinsic ratings to change over time.
- Convergence of ratings over time.
Pros/Cons of Elo

Pros:

- Simple to implement/adjust.
- Allows for a player’s intrinsic ratings to change over time.
- Convergence of ratings over time.

Cons:
Pros/Cons of Elo

Pros:

- Simple to implement/adjust.
- Allows for a player’s intrinsic ratings to change over time.
- Convergence of ratings over time.

Cons:

- Single factor analysis, may be less accurate for team sports.
Pros/Cons of Elo

Pros:

- Simple to implement/adjust.
- Allows for a player’s intrinsic ratings to change over time.
- Convergence of ratings over time.

Cons:

- Single factor analysis, may be less accurate for team sports.
- Even in sports where a ‘single factor’ is reasonable - e.g. chess - ratings tend to not actually be normally distributed.
Pros/Cons of Elo

Pros:

- Simple to implement/adjust.
- Allows for a player’s intrinsic ratings to change over time.
- Convergence of ratings over time.

Cons:

- Single factor analysis, may be less accurate for team sports.
- Even in sports where a ‘single factor’ is reasonable - e.g. chess - ratings tend to not actually be normally distributed.
  - In particular in chess, weaker players statistically tend to have higher win probabilities (and thus Elo rankings) than predicted.
Ratings Distribution for Chess

**Figure 5:** Ratings Distribution from USCF Golden Database, 2014

Dan Lanoue
The Science of Ranking and Rating
NFL Elo Rankings

An example of Elo methods applied to the NFL can be seen at the FiveThirtyEight blog [6]. The details of FiveThirtyEight’s system, as far as are publicly available, are this:
NFL Elo Rankings

An example of Elo methods applied to the NFL can be seen at the FiveThirtyEight blog [6]. The details of FiveThirtyEight’s system, as far as are publicly available, are this:

- Average Elo ratings are set at 1500
An example of Elo methods applied to the NFL can be seen at the FiveThirtyEight blog [6]. The details of FiveThirtyEight’s system, as far as are publicly available, are this:

- Average Elo ratings are set at 1500
- The spread of Elo ratings are set so that for every 25 points the two teams’ Elo ratings differ, the point spread increases by 1 (regular NFL) point.
An example of Elo methods applied to the NFL can be seen at the FiveThirtyEight blog [6]. The details of FiveThirtyEight’s system, as far as are publicly available, are this:

- Average Elo ratings are set at 1500
- The spread of Elo ratings are set so that for every 25 points the two teams’ Elo ratings differ, the point spread increases by 1 (regular NFL) point.
- For example: If the Patriots have a rating of 1575 and the Falcon’s rate at 1475 then (on a neutral field) the Patriots would be a 4 point favorite.
NFL Elo Rankings

The full details of FiveThirtyEight’s methodology do not seem to be publicly available. While their listed algorithm is either incorrect or incomplete, the basic idea is that team $A$ beats team $B$ then the update in team $A$’s rating is

$$r^\text{new}_A - r_A \propto \ln(|\text{margin of victory}| + 1),$$

with $r_B$ decreasing by the same amount.
Although short on math, FiveThirtyEight is long on colorful graphics and witty commentary!

**FiveThirtyEight’s NFL Elo Ratings**  
Oct. 8, 2014

<table>
<thead>
<tr>
<th>TEAM</th>
<th>RATING</th>
<th>ELO CHANGE</th>
<th>RANK CHANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seattle</td>
<td>1679</td>
<td>-8</td>
<td>-</td>
</tr>
<tr>
<td>Denver</td>
<td>1646</td>
<td>+22</td>
<td>-</td>
</tr>
<tr>
<td>San Francisco</td>
<td>1622</td>
<td>+10</td>
<td>-</td>
</tr>
<tr>
<td>San Diego</td>
<td>1619</td>
<td>+11</td>
<td>-</td>
</tr>
<tr>
<td>New England</td>
<td>1610</td>
<td>+29</td>
<td>+2</td>
</tr>
<tr>
<td>Cincinnati</td>
<td>1578</td>
<td>-29</td>
<td>-1</td>
</tr>
<tr>
<td>Arizona</td>
<td>1574</td>
<td>-22</td>
<td>-1</td>
</tr>
<tr>
<td>Indianapolis</td>
<td>1564</td>
<td>+10</td>
<td>+2</td>
</tr>
<tr>
<td>Dallas</td>
<td>1563</td>
<td>+6</td>
<td></td>
</tr>
<tr>
<td>Philadelphia</td>
<td>1555</td>
<td>+11</td>
<td>+1</td>
</tr>
<tr>
<td>Baltimore</td>
<td>1549</td>
<td>-18</td>
<td>-3</td>
</tr>
<tr>
<td>Green Bay</td>
<td>1546</td>
<td>+24</td>
<td>+4</td>
</tr>
<tr>
<td>Pittsburgh</td>
<td>1543</td>
<td>+11</td>
<td></td>
</tr>
<tr>
<td>Carolina</td>
<td>1542</td>
<td>+14</td>
<td></td>
</tr>
<tr>
<td>New Orleans</td>
<td>1536</td>
<td>-9</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6:** Week 6 Elo Rankings [6]
Perhaps most interesting, using the built-in future predictions of Elo rankings FiveThirtyEight is able to simulate the rest of the season and give predictions for each team.

![Figure 7: Future Predictions using Week 6 Elo Rankings](image-url)
The TrueSkill Rankings

Developed for Xbox Live at Microsoft Research to track the rankings of players and match them competitively using Bayesian inference.
The TrueSkill Rankings

Developed for Xbox Live at Microsoft Research to track the rankings of players and match them competitively using Bayesian inference.

- TrueSkill tracks both a player’s rating $\mu$ as well as a degree of uncertainty $\sigma$ in that rating, which decreases over time.
The TrueSkill Rankings

Developed for Xbox Live at Microsoft Research to track the rankings of players and match them competitively using Bayesian inference.

- TrueSkill tracks both a player’s rating $\mu$ as well as a degree of uncertainty $\sigma$ in that rating, which decreases over time.
- Allows for multi-player matches.
TrueSkill Example

Before Eric and Natalie play:

Figure 8: TrueSkill rating distribution [3]
TrueSkill Example

After the match:

Figure 9: TrueSkill rating distribution [3]

Dan Lanoue
The Science of Ranking and Rating
How TrueSkill Works

TrueSkill is essentially the Elo system, with a variable $K$-factor based on each player’s $\sigma$.

\[
\mu_{\text{winner}} \leftarrow \mu_{\text{winner}} + \frac{\sigma^2_{\text{winner}}}{c} \cdot \nu \left( \frac{\mu_{\text{winner}} - \mu_{\text{loser}}}{c}, \frac{\varepsilon}{c} \right)
\]

\[
\mu_{\text{loser}} \leftarrow \mu_{\text{loser}} - \frac{\sigma^2_{\text{loser}}}{c} \cdot \nu \left( \frac{\mu_{\text{winner}} - \mu_{\text{loser}}}{c}, \frac{\varepsilon}{c} \right)
\]

\[
\sigma^2_{\text{winner}} \leftarrow \sigma^2_{\text{winner}} \cdot \left[ 1 - \frac{\sigma^2_{\text{winner}}}{c^2} \cdot w \left( \frac{\mu_{\text{winner}} - \mu_{\text{loser}}}{c}, \frac{\varepsilon}{c} \right) \right]
\]

\[
\sigma^2_{\text{loser}} \leftarrow \sigma^2_{\text{loser}} \cdot \left[ 1 - \frac{\sigma^2_{\text{loser}}}{c^2} \cdot w \left( \frac{\mu_{\text{winner}} - \mu_{\text{loser}}}{c}, \frac{\varepsilon}{c} \right) \right]
\]

\[
c^2 = 2\beta^2 + \sigma^2_{\text{winner}} + \sigma^2_{\text{loser}}
\]

Figure 10: TrueSkill Algorithm [9]
Google’s PageRank

At the core of Google’s original search algorithm is a rating method for websites based on Markov chain theory. Called PageRank, the algorithm rates websites based on the number of incoming links.
Heuristically, a page’s rank (or rating) can be thought of as the probability of a ‘random web surfer’ winding up on the page by following random links.
Google’s PageRank: Heuristics

Heuristically, a page’s rank (or rating) can be thought of as the probability of a ‘random web surfer’ winding up on the page by following random links.

Note that while the number of incoming links to a page $A$ are important, so is the probability of even being at the origin of those links; i.e. so are the number of links to pages that link to $A$, and so on recursively.
Google’s PageRank: Math

The PageRank $P(u)$ of a page $u$ is given by

$$P(u) = \sum_{v \in B_u} \frac{P(v)}{\text{deg}(v)}.$$
The PageRank $P(u)$ of a page $u$ is given by

$$P(u) = \sum_{v \in B_u} \frac{P(v)}{\deg(v)}.$$ 

Here $B_u$ is the set of all pages linking to $u$ (with multiplicity) and $\deg(v)$ is the degree, i.e. number of outgoing links, of a page $v$. These equations can either be solved recursively/iteratively or by linear algebra.
Google’s PageRank: Math

The PageRank $P(u)$ of a page $u$ is given by

$$P(u) = \sum_{v \in B_u} \frac{P(v)}{\text{deg}(v)}.$$ 

Here $B_u$ is the set of all pages linking to $u$ (with multiplicity) and $\text{deg}(v)$ is the degree, i.e. number of outgoing links, of a page $v$. These equations can either be solved recursively/iteratively or by linear algebra.

In general $P(u)$ is normalized so that

$$\sum_u P(u) = 1.$$
Google’s PageRank: Math

The PageRank $P(u)$ of a page $u$ is given by

$$P(u) = \sum_{v \in B_u} \frac{P(v)}{\text{deg}(v)}.$$  

Here $B_u$ is the set of all pages linking to $u$ (with multiplicity) and $\text{deg}(v)$ is the degree, i.e. number of outgoing links, of a page $v$. These equations can either be solved recursively/iteratively or by linear algebra. In general $P(u)$ is normalized so that

$$\sum_u P(u) = 1.$$  

In applications a ‘dampening’ factor can be introduced to account for sources and sinks. We can also weight edges non-uniformly.
Google’s PageRank: An Example

Figure 12: An example of PageRank for a small graph [10]
Mathematically, finding the PageRank for a graph is equivalent to finding the stationary distribution of the *Markov Chain* given by the graph.
Mathematically, finding the PageRank for a graph is equivalent to finding the stationary distribution of the Markov Chain given by the graph.

Given some technical conditions (irreducibility, aperiodicity) standard theory ensures the existence of a unique stationary distribution.
To apply PageRank to a sports season, we construct a graph based on the seasons results so far:
To apply PageRank to a sports season, we construct a graph based on the seasons results so far:

- The vertex set are the teams.
To apply PageRank to a sports season, we construct a graph based on the seasons results so far:

- The vertex set are the teams.
- For every game that team $A$ beats team $B$, add a directed edge from $B$ to $A$. 
To apply PageRank to a sports season, we construct a graph based on the seasons results so far:

- The vertex set are the teams.
- For every game that team $A$ beats team $B$, add a directed edge from $B$ to $A$.
- The edges can be weighted by margin of victory, etc.
Pros and Cons

Pros:
Pros and Cons

Pros:

- Simple and easy to calculate,
Pros and Cons

Pros:

- Simple and easy to calculate,
- Commutative, i.e. the order of games doesn’t matter
Pros:

- Simple and easy to calculate,
- Commutative, i.e. the order of games doesn’t matter

Cons:

- Single factor rankings,
Pros and Cons

Pros:

- Simple and easy to calculate,
- Commutative, i.e. the order of games doesn’t matter

Cons:

- Single factor rankings,
- Subject to wild swings in a teams ranking, sometimes based on results having nothing to do with that team.
PageRank NFL Example

Figure 13: Super Bowl XLIV

Let’s look at the PageRank graph for the first few weeks of the 2009 NFL season. The data/visualization is from the Web Science and Digital Libraries Research Group at Old Dominion U. [1]
PageRank NFL Example
PageRank NFL Example

Week 2

Dan Lanoue
The Science of Ranking and Rating
PageRank NFL Example
PageRank NFL Example

Week 5

Dan Lanoue
The Science of Ranking and Rating
PageRank NFL Example

Week 6
PageRank NFL Example
PageRank NFL Example
PageRank NFL Example
PageRank NFL Example

![PageRank NFL Example Diagram]

The Basics
PageRank in Sports
PageRanking the NFL
PageRank for Passing Networks in Soccer

Dan Lanoue
The Science of Ranking and Rating
PageRank NFL Example
PageRank NFL Example
While not strictly a ranking/prediction method, a very interesting application of PageRank is passing networks within a soccer team.[12]
While not strictly a ranking/prediction method, a very interesting application of PageRank is passing networks within a soccer team.\[12\]

The idea is to evaluate the strategy and robustness of a soccer team by looking at the role of each player in moving the ball, potentially uncovering weaknesses as well as determining each player’s relative importance.
Passing Networks

Figure 14: Dutch and Spanish national teams, 2010 [8]
The PageRank of a player can be thought of as the probability they have the ball at any given time.
The PageRank of a player can be thought of as the probability they have the ball at any given time.

Analysis of ‘closeness’ and ‘clustering’ - basically the balance and connectivity of the network - shows a correlation with the overall performance of the team.[12]
The Elo and PageRank are essentially single factor methods, based only on margin of victory. Let’s take a look at a significantly more complicated statistic specific to the NFL known as Defense Adjusted Value over Average, or DVOA.[4]
The basic idea is to rate every single play in every game.
The basic idea is to rate every single play in every game.

- Each play is given a ‘success score’, which takes into account whether the play succeeded or not as well as the game situation.
The basic idea is to rate every single play in every game.

- Each play is given a ‘success score’, which takes into account whether the play succeeded or not as well as the game situation.
  - For ex. an 8 yard run on 3rd and 10 might be worth +0.54 points.
The basic idea is to rate every single play in every game.

- Each play is given a `success score`, which takes into account whether the play succeeded or not as well as the game situation.
  - For ex. an 8 yard run on 3rd and 10 might be worth +0.54 points.
- A bad outcome would get negative score; for instance a fumble is worth between −1.7 and −4.0 points depending on the situation.
The basic idea is to rate every single play in every game.

- Each play is given a ‘success score’, which takes into account whether the play succeeded or not as well as the game situation.
  - For ex. an 8 yard run on 3rd and 10 might be worth +0.54 points.
- A bad outcome would get negative score; for instance a fumble is worth between −1.7 and −4.0 points depending on the situation.
- Positive ratings are good for the offense, negative are good for the defense.
The Basics

Then:
Then:

- For every play, each player/team’s score is compared with the league average score on similar plays.
Then:

- For every play, each player/team’s score is compared with the league average score on similar plays.
- These relative scores are then combined to get the non-adjusted DVOA.
Then:

- For every play, each player/team’s score is compared with the league average score on similar plays.
- These relative scores are then combined to get the non-adjusted DVOA.
- Finally, adjustments are made for strength of competition, i.e. strength of schedule.
Recall that a greater negative score on defense is good.
A similar family of statistics to DVOA are Advanced Football Analytics’ Expected Points Added (EPA) and Win Probability Added (WPA). We’ll focus on EPA.
For each field position and down and distance, an offense has an expected number of points they will score based on historical data.
The Basics

EPA is a differential statistic. Here’s how it works:

Let’s say the Patriots have a 1st and 10 from their own 20 yard line, which has an expectation of 0
0.5 points.
Tom Brady throws a 30 yard pass and it’s now 1st and 10 from the 50, which is worth an expectation of 2 points.
Thus that pass has just added an expectation of +1
0.5 points.

Breaking this up among individual players can be complicated, but it’s easy to see how this turns into a rating for a team’s offense and defense.
EPA is a differential statistic. Here’s how it works:

- Lets say the Patriots have a 1st and 10 from their own 20 yard line, which has an expectation of 0.5 points.
The Basics

EPA is a differential statistic. Here’s how it works:

- Let’s say the Patriots have a 1st and 10 from their own 20 yard line, which has an expectation of 0.5 points.
- Tom Brady throws a 30 yard pass and it’s now 1st and 10 from the 50, which is worth an expectation of 2 points.
EPA is a differential statistic. Here's how it works:

- Let's say the Patriots have a 1st and 10 from their own 20 yard line, which has an expectation of 0.5 points.
- Tom Brady throws a 30 yard pass and it's now 1st and 10 from the 50, which is worth an expectation of 2 points.
- Thus that pass has just added an expectation of +1.5 points.
EPA is a differential statistic. Here’s how it works:

- Let’s say the Patriots have a 1st and 10 from their own 20 yard line, which has an expectation of 0.5 points.
- Tom Brady throws a 30 yard pass and it’s now 1st and 10 from the 50, which is worth an expectation of 2 points.
- Thus that pass has just added an expectation of +1.5 points.

Breaking this up among individual players can be complicated, but its easy to see how this turns into a rating for a team’s offense and defense.
EPA Ratings

Figure 16: EPA Rankings, Week 5 2014 [2]
Win Probability Graphs

Connected to this type of analysis are win probability graphs.

EPA (or more rightly its cousin WPA) can be thought of as assigning credit for the variation in win probability. More on this in your lecture next week!
An Observation

Statistics like DVOA and EPA are most interesting when they both ‘confirm’ some of our beliefs but also contradict them in surprising ways.

Let’s take a look at Detroit Lion WR’s historic 2012 season where he set the all time record for receiving yards in a season.
According to DVOA, Calvin Johnson’s 2012 season not only wasn’t the greatest of all time, but only the 20th best in just 2012! First place belonged to the (fairly anonymous) Danario Alexander.
An Observation

So does this mean all the pundits are wrong? Or is DVOA flawed? There isn’t necessarily a simple answer here.
An Observation

So does this mean all the pundits are wrong? Or is DVOA flawed? There isn’t necessarily a simple answer here.

A bit more can be said though:
An Observation

So does this mean all the pundits are wrong? Or is DVOA flawed? There isn’t necessarily a simple answer here.

A bit more can be said though:

- DVOA masks an important issue in ratings, efficiency vs. usage.
An Observation

So does this mean all the pundits are wrong? Or is DVOA flawed? There isn’t necessarily a simple answer here.

A bit more can be said though:

- DVOA masks an important issue in ratings, efficiency vs. usage.
- Players (and teams) that are asked to do more, tend to do so less efficiently.
An Observation

So does this mean all the pundits are wrong? Or is DVOA flawed? There isn’t necessarily a simple answer here.

A bit more can be said though:

- DVOA masks an important issue in ratings, efficiency vs. usage.
- Players (and teams) that are asked to do more, tend to do so less efficiently.
- Another way to put this is that consistently performing at league average is itself an ‘above average’ skill.
An Observation

So does this mean all the pundits are wrong? Or is DVOA flawed? There isn’t necessarily a simple answer here.

A bit more can be said though:

- DVOA masks an important issue in ratings, efficiency vs. usage.
- Players (and teams) that are asked to do more, tend to do so less efficiently.
- Another way to put this is that consistently performing at league average is itself an ‘above average’ skill.

Taking this (and some other factors) into account, Football Outsiders introduced DYAR.
An Observation

According to DYAR, Calvin Johnson’s season is back to being historic. But are we just picking a statistic that confirms what we believed?

<table>
<thead>
<tr>
<th>Player</th>
<th>Team</th>
<th>DYAR</th>
<th>Rk</th>
<th>YAR</th>
<th>Rk</th>
<th>DVOA</th>
<th>Rk</th>
<th>VOA</th>
<th>Passes</th>
<th>Yards</th>
<th>EYds</th>
<th>TD</th>
<th>Catch Rate</th>
<th>FUM</th>
<th>DPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.Johnson</td>
<td>DET</td>
<td>451</td>
<td>1</td>
<td>437</td>
<td>2</td>
<td>14.7%</td>
<td>20</td>
<td>13.8%</td>
<td>203</td>
<td>1,964</td>
<td>2,008</td>
<td>5</td>
<td>60%</td>
<td>3</td>
<td>5/102</td>
</tr>
<tr>
<td>A.Johnson</td>
<td>HOU</td>
<td>415</td>
<td>2</td>
<td>464</td>
<td>1</td>
<td>19.6%</td>
<td>16</td>
<td>23.4%</td>
<td>163</td>
<td>1,598</td>
<td>1,650</td>
<td>4</td>
<td>60%</td>
<td>0</td>
<td>2/24</td>
</tr>
<tr>
<td>D.Thomas</td>
<td>DEN</td>
<td>399</td>
<td>3</td>
<td>401</td>
<td>3</td>
<td>21.4%</td>
<td>14</td>
<td>21.7%</td>
<td>141</td>
<td>1,430</td>
<td>1,539</td>
<td>10</td>
<td>67%</td>
<td>2</td>
<td>5/97</td>
</tr>
<tr>
<td>E.Decker</td>
<td>DEN</td>
<td>397</td>
<td>4</td>
<td>401</td>
<td>4</td>
<td>27.2%</td>
<td>8</td>
<td>27.6%</td>
<td>123</td>
<td>1,071</td>
<td>1,397</td>
<td>13</td>
<td>69%</td>
<td>0</td>
<td>3/54</td>
</tr>
<tr>
<td>L.Moore</td>
<td>NO</td>
<td>368</td>
<td>5</td>
<td>344</td>
<td>7</td>
<td>31.3%</td>
<td>5</td>
<td>28.6%</td>
<td>104</td>
<td>1,041</td>
<td>1,217</td>
<td>6</td>
<td>63%</td>
<td>0</td>
<td>1/17</td>
</tr>
<tr>
<td>M.Crabtree</td>
<td>SF</td>
<td>344</td>
<td>6</td>
<td>336</td>
<td>9</td>
<td>21.9%</td>
<td>13</td>
<td>21.1%</td>
<td>127</td>
<td>1,113</td>
<td>1,318</td>
<td>9</td>
<td>68%</td>
<td>0</td>
<td>3/57</td>
</tr>
<tr>
<td>M.Colston</td>
<td>NO</td>
<td>338</td>
<td>7</td>
<td>339</td>
<td>8</td>
<td>19.8%</td>
<td>15</td>
<td>19.9%</td>
<td>130</td>
<td>1,154</td>
<td>1,346</td>
<td>10</td>
<td>64%</td>
<td>4</td>
<td>3/33</td>
</tr>
<tr>
<td>D.Bryant</td>
<td>DAL</td>
<td>336</td>
<td>8</td>
<td>352</td>
<td>6</td>
<td>18.3%</td>
<td>17</td>
<td>19.8%</td>
<td>138</td>
<td>1,382</td>
<td>1,379</td>
<td>12</td>
<td>67%</td>
<td>2</td>
<td>1/14</td>
</tr>
<tr>
<td>R.White</td>
<td>ATL</td>
<td>334</td>
<td>9</td>
<td>361</td>
<td>5</td>
<td>16.3%</td>
<td>18</td>
<td>18.7%</td>
<td>143</td>
<td>1,351</td>
<td>1,431</td>
<td>7</td>
<td>64%</td>
<td>1</td>
<td>1/1</td>
</tr>
<tr>
<td>M.Floyd</td>
<td>SD</td>
<td>324</td>
<td>10</td>
<td>303</td>
<td>12</td>
<td>36.0%</td>
<td>3</td>
<td>32.8%</td>
<td>84</td>
<td>814</td>
<td>1,022</td>
<td>5</td>
<td>67%</td>
<td>0</td>
<td>2/37</td>
</tr>
<tr>
<td>J.Jones</td>
<td>ATL</td>
<td>300</td>
<td>11</td>
<td>320</td>
<td>10</td>
<td>16.0%</td>
<td>19</td>
<td>17.9%</td>
<td>129</td>
<td>1,198</td>
<td>1,293</td>
<td>10</td>
<td>61%</td>
<td>0</td>
<td>3/45</td>
</tr>
<tr>
<td>R.Cobb</td>
<td>GB</td>
<td>300</td>
<td>12</td>
<td>304</td>
<td>11</td>
<td>24.0%</td>
<td>9</td>
<td>24.5%</td>
<td>104</td>
<td>954</td>
<td>1,107</td>
<td>8</td>
<td>77%</td>
<td>0</td>
<td>1/23</td>
</tr>
</tbody>
</table>
There isn’t necessarily a simple conclusion here. We’ve definitely learned something - both about Calvin Johnson and about measuring efficiency. These are the sort of issues that come up when we look at ‘advanced stats’.
There isn’t necessarily a simple conclusion here. We’ve definitely learned something - both about Calvin Johnson and about measuring efficiency. These are the sort of issues that come up when we look at ‘advanced stats’.

It’s one thing to make up a statistic from a data set, interpreting it can be a whole other story. Even more complicated still is using these statistics to make concrete predictions.
An Observation: Conclusion

There isn’t necessarily a simple conclusion here. We’ve definitely learned something - both about Calvin Johnson and about measuring efficiency. These are the sort of issues that come up when we look at ‘advanced stats’.

It’s one thing to make up a statistic from a data set, interpreting it can be a whole other story. Even more complicated still is using these statistics to make concrete predictions.

We’ll conclude by looking at how some of these methods perform in prediction.
Ratings vs. Vegas

The barometer of sports predictions are the Vegas odds. It’s one thing to assign teams ratings based on a cute mathematical formula, but if the predictions don’t add any actual predictive power then the ratings are meaningless.

Figure 17: A sports book odds board
So perhaps the most important question we can ask about these ratings is:
So perhaps the most important question we can ask about these ratings is:

How do they perform when compared with Vegas (or market) predictions?
So perhaps the most important question we can ask about these ratings is:

How do they perform when compared with Vegas (or market) predictions?

Unfortunately, the answer is not well.
There are of course some confounding factors here:
There are of course some confounding factors here:

- A rating that consistently beats Vegas would never be made public.
There are of course some confounding factors here:

- A rating that consistently beats Vegas would never be made public.
- Even if it was, it would quickly be factored into the Vegas odds.
There are of course some confounding factors here:

- A rating that consistently beats Vegas would never be made public.
- Even if it was, it would quickly be factored into the Vegas odds.

With this in mind, let’s take a look at how some of the ratings we’ve talked about do compared to Vegas and each other.
First, a brief reminder of what Vegas point spreads mean. Clearly each game is not an even match - almost every game has a favorite. The goal is to give a possible outcome of the margin of victory for which each side is equally likely.
Vegas Point Spreads

First, a brief reminder of what Vegas point spreads mean. Clearly each game is not an even match - almost every game has a favorite. The goal is to give a possible outcome of the margin of victory for which each side is equally likely.

Oakland is ‘getting’ 7 points, i.e. the median outcome is San Diego winning by 7 points.
Vegas Point Spreads

Some important facts about betting on spreads:
Some important facts about betting on spreads:

- Spreads on average pay -$110, that is to win $100 you need to bet $110.
Vegas Point Spreads

Some important facts about betting on spreads:

- Spreads on average pay $-110$, that is to win $100$ you need to bet $110$.
- So to make money, we need to correctly predict games more than $52.38\%$ of the time.
Some important facts about betting on spreads:

- Spreads on average pay $-110, that is to win $100 you need to bet $110.
- So to make money, we need to correctly predict games more than 52.38% of the time.
- This of course ignores all other expenses as well as the opportunity cost of betting.
FiveThirtyEight Elo vs. the Spread

There’s only been 5 weeks to test these ratings so far, but if you had been betting on FiveThirtyEight’s Elo picks (as they are quick to point out) you would be seriously in the red.
There’s only been 5 weeks to test these ratings so far, but if you had been betting on FiveThirtyEight’s Elo picks (as they are quick to point out) you would be seriously in the red.

So far the Elo rankings are 33-38-3 against Vegas for a win percentage of 46.4%.
FiveThirtyEight Elo vs. the Spread

There’s only been 5 weeks to test these ratings so far, but if you had been betting on FiveThirtyEight’s Elo picks (as they are quick to point out) you would be seriously in the red.

So far the Elo rankings are 33-38-3 against Vegas for a win percentage of 46.4%.

Of course the point of Elo rankings are long-term convergence, so perhaps they will be more likely to do better as the season goes on.
According to Football Outsiders [4], their ‘Premium Picks’ record against the spread is:
According to Football Outsiders [4], their ‘Premium Picks’ record against the spread is:

- 2010 Regular season: 141-108-7, 56.1%
DVOA vs. the Spread

According to Football Outsiders [4], their ‘Premium Picks’ record against the spread is:

- 2010 Regular season: 141-108-7, 56.1%
- 2011 Regular season: 127-117-12, 52%
According to Football Outsiders [4], their ‘Premium Picks’ record against the spread is:

- 2010 Regular season: 141-108-7, 56.1%
- 2011 Regular season: 127-117-12, 52%
- 2012 Regular season: 144-104-8, 57.8%
According to Football Outsiders [4], their ‘Premium Picks’ record against the spread is:

- 2010 Regular season: 141-108-7, 56.1%
- 2011 Regular season: 127-117-12, 52%
- 2012 Regular season: 144-104-8, 57.8%
- 2013 Regular season: 119-130-7, 47.9%
According to Football Outsiders [4], their ‘Premium Picks’ record against the spread is:

- 2010 Regular season: 141-108-7, 56.1%
- 2011 Regular season: 127-117-12, 52%
- 2012 Regular season: 144-104-8, 57.8%
- 2013 Regular season: 119-130-7, 47.9%

Total: 531-459-34, 53.6%.
According to Football Outsiders [4], their ‘Premium Picks’ record against the spread is:

- 2010 Regular season: 141-108-7, 56.1%
- 2011 Regular season: 127-117-12, 52%
- 2012 Regular season: 144-104-8, 57.8%
- 2013 Regular season: 119-130-7, 47.9%

Total: 531-459-34, 53.6%.

Over 4 years and 1024 games, 53.6% is 12 games above the break even point.
Let’s compare this to CBS’s expert picks.

<table>
<thead>
<tr>
<th>NFL WEEK</th>
<th>Pete Prisco</th>
<th>Jason La Canfora</th>
<th>Will Brinson</th>
<th>Josh Katzowitz</th>
<th>Ryan Wilson</th>
<th>John Breech</th>
<th>Dave Richard</th>
<th>Jamey Eisenberg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Totals</td>
<td>1-0.0</td>
<td>1-0.0</td>
<td>1-0.0</td>
<td>1-0.0</td>
<td>1-0.0</td>
<td>1-0.0</td>
<td>0-1.0</td>
<td>1-0.0</td>
</tr>
<tr>
<td>42-35-0</td>
<td>44-33-0</td>
<td>30-47-0</td>
<td>33-44-0</td>
<td>39-38-0</td>
<td>39-38-0</td>
<td>40-37-0</td>
<td>44-33-0</td>
<td></td>
</tr>
<tr>
<td>Last Season</td>
<td>126-141-0</td>
<td>119-148-0</td>
<td>134-133-0</td>
<td>119-148-0</td>
<td>117-150-0</td>
<td>130-137-0</td>
<td>122-145-0</td>
<td>130-137-0</td>
</tr>
</tbody>
</table>

That’s a 50.7% win percentage this year so far (compare with 46.6% last year).
Does this mean ratings are worthless? No!
Does this mean ratings are worthless? No!

It’s possible that over the long run Vegas odds are set ‘perfectly’ - i.e. an efficient market - and there is no money to be made. Perhaps there are small edges to be found on the margin.
Conclusion

Does this mean ratings are worthless? No!

It’s possible that over the long run Vegas odds are set ‘perfectly’ - i.e. an efficient market - and there is no money to be made. Perhaps there are small edges to be found on the margin.

But team sports (and the NFL in particular) are complicated, multi-faceted, and have far more luck involved than we may realize. We tend to build narratives after the fact (teams of destiny, etc.) that make us feel like we could have predicted things, if we could have only seen things more clearly.
Some project ideas:

- **Which rating model is the best?** Compare different models’ performance in a sport (or sports) over a larger time frame using historical data.
Some Ideas

Some project ideas:

- **Which rating model is the best?** Compare different models’ performance in a sport (or sports) over a larger time frame using historical data.

- **Can you beat the house?** Construct your own rating/prediction model and win (us) lots of money.
Some Ideas

Some project ideas:

- **Which rating model is the best?** Compare different models’ performance in a sport (or sports) over a larger time frame using historical data.

- **Can you beat the house?** Construct your own rating/prediction model and win (us) lots of money.

- Many other ideas for your favorite sport.
Some project ideas:

- **Which rating model is the best?** Compare different models’ performance in a sport (or sports) over a larger time frame using historical data.

- **Can you beat the house?** Construct your own rating/prediction model and win (us) lots of money.

- Many other ideas for your favorite sport.

**Thanks for listening!**


A network theory analysis of football strategies.  
_ArXiv e-prints_, June 2012.