Social Network Analysis of NFL Coaches, 1980-2013

Nathan E Dire

November 2014

1 Introduction

Coaching in the United States' National Football League is a high stakes business. NFL coaches domainate the top paid coaches in U.S. sports [3]. The best coaches are paid in excess of $7 million per year. In addition, there is relatively high turnover, with many of the 32 head coaches being fired each year.

The group of NFL coaches is a relatively tight circle. Head coaches are often hired from the pool of offensive and defensive coordinators. There is a general perception that influential coaches pass on their knowledge to assistants, who then become successful head coaches themselves. This is typically associated wiht a particular philosophy, such as the "west coast offense" or the 3-4 defense. Figure 1 shows the coaching family tree for Bill Walsh, a successful coach from the 1980s.

This report considers a two main questions about NFL coaches who have worked together. First, does centrality in the social network align with overall coaching success? Second, are there distinct communities in the network, perhaps aligned with some particular style or philosophy?

2 Methodology

2.1 Data Acquisition

Data on NFL coaching staffs was gathered from the site `http://www.pro-football-reference.org` by reading the pages and parsing the listed coaching staff. For example, this page `http://www.pro-football-reference.com/teams/det/2010.htm` lists the coaches for the 2010 Detroit Lions. The team page generally lists the head coach, offensive coordinator, and defensive coordinator. Only the years 1980-2013 were considered. The code used to produce this data is available at [4].

2.2 Network Construction

The relationship between members of the coaching staff is subject to interpretation. Within this data, the coach may be both a coordinator and head coach.

For this analysis, I consider two constructions of the social network:

1. The first model attempts to follow the "coaching family tree" perception, and only has directed edges from head coach to coordinator. This approach attempts to model the influence of successful coaches through the success of their assistants. Edges are between the head coach and the coordinators; coordinators have no connection to each other. This model will be

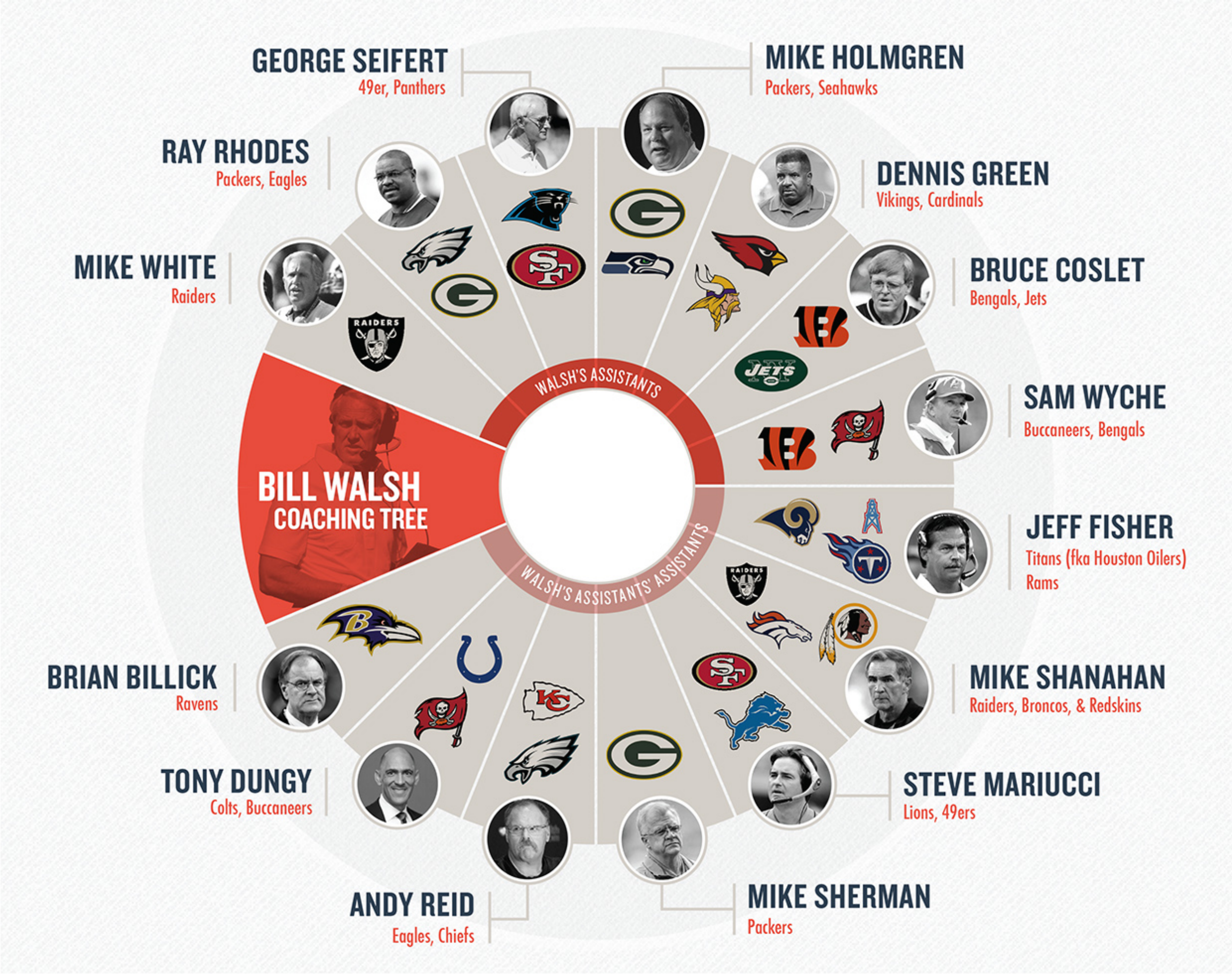

Figure 1: Image: The Bill Walsh Coaching Tree. Source: http://blog.hubspot.com/marketing/paypal-mafia-bill-walsh-nfl-infographic

referred to as the *tree* model. The tree model is used for centrality measures since this report is looking for the centrality successfull *head* coaches. The file is `coaches_tree.gml`.

2. The second model considers each member of the staff to be connected with every coach they served with, with increasing weight for each year served. This approach attempts to identify coaching "communities" who may tend serve together. This model will be referred to as the *peer* model. The file is `coaches_peer.gml`.

2.3 Analysis

Network analysis is done in the R programming language with the `igraph` package. The calculations are included in this document using the `knitr` package.

For centrality measurement, the validation metric is based on alignment with super-bowl winning coaches since this data was easy to gather. Due to multiple career wins, there are only 23 super-bowl winning coaches since 1980. Ideally the evaluation should be based on a more continuous metric applicable to all coaches, such as win percentage, which can be used to calculate correlation.

For community finding, it's difficult to pick a validation metric. The main evaluation metric is subjective based on the infographic in figure 1.

2.4 Limitations

There are several important limitations of this anlaysis:

- Only the head coach and coordinators are considered. This represents the only clean data available from http://pro-football-reference.com. It also avoids the complexities resulting from teams having different names for the various position coaches. This has the downside of not recognizing cases where head coaches and coordinators take jobs as position coaches.
- Only the years 1980-2013 are considered. The consistent reporting of offensive and defensive coordinators appears to start around 1980. This results in the network being truncated for coaches employed before and after 1980.
- Coaches are considered to serve the entire year. In reality, teams may occasionally fire coaches mid-year, and coaches may take a leave of absence.

3 Network Properties

The network includes 427 nodes. The tree model has 777 edges; the peer model has 1244. In both networks, the possible number of edges is constrained due to the realities of the coaching profession. It's nearly impossible for a node to have 0 edges since no coach serves all three roles. The maximum degree is also relatively limited due to coaching careers being finite. It's rare for a head coach or coordinator to be under 30 or over 70.

The degree distribution for the tree model is calculated as follows and shown in figure 2.

```
tree_g = read.graph("coaches_tree.gml",format="gml")
summary(tree_g)
```

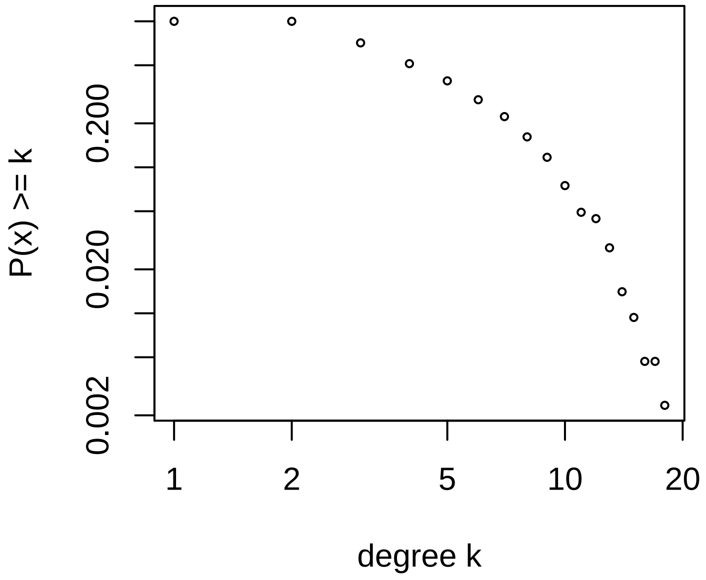

Figure 2: Cumulative degree distribution for tree model

```
## IGRAPH U--- 427 777 --
## attr: href (v/c), guid (v/c), label (v/c), id (v/n), value (e/n)

tree_deg = degree.distribution(tree_g, cumulative = TRUE)
plot(tree_deg,log="xy",xlab="degree k",ylab="P(x) >= k",cex=0.5)
```

The degree distribution for the peer model is calculated as follows and shown in figure 2.

```
peer_g = read.graph("coaches_peer.gml",format="gml")
summary(peer_g)

## IGRAPH U--- 427 1244 --
## attr: href (v/c), guid (v/c), label (v/c), id (v/n), value (e/n)

peer_deg = degree.distribution(tree_g, cumulative = TRUE)
plot(peer_deg,log="xy",xlab="degree k",ylab="P(x) >= k",cex=0.5)
```

A Fruchterman-Reingold layout of the tree graph is show in figure 4. There graph appears tighly clustered in general, without an obvious community structure.

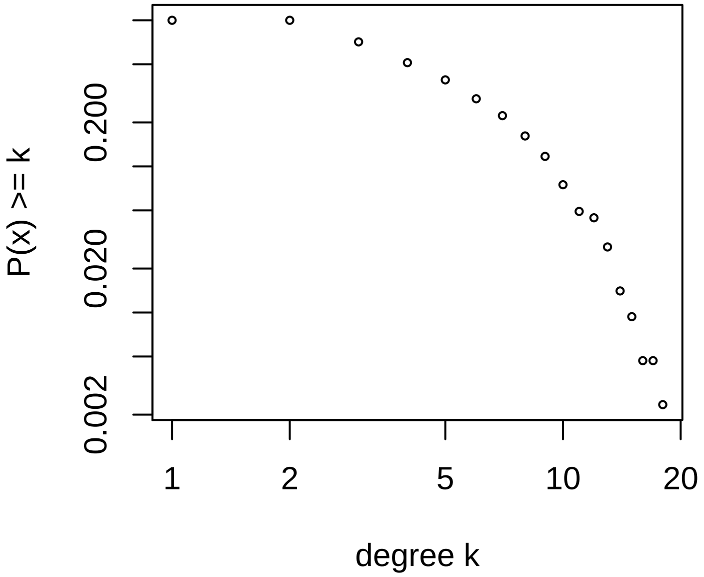

Figure 3: Cumulative degree distribution for peer model

```
plot.igraph(tree_g, layout=layout.fruchterman.reingold, vertex.label=NA, vertex.size=4)
```

4 Centrality

There are several measures of centrality; it's unclear which is most useful for this network. This section considers *betweenness*, *closeness*, *page rank*, and *authority*. The tree model is used in order to stress the influence of the head coach.

4.1 Degree

The simplest centrality measure is degree centrality. The following snippet selects the top 20 nodes by degree:

```
td = degree(tree_g)
top = V(tree_g)$label[order(td, decreasing=TRUE)[1:20]]
intersect(top, sb_winners)

## [1] "Mike Shanahan"  "Mike Holmgren"  "Pete Carroll"   "Tom Coughlin"
## [5] "George Seifert" "Bill Belichick" "Tony Dungy"
```

There are 7 Super Bowl winners in the top 20 nodes by degree centrality.

4.2 Betweenness

The first measure of centrality to be evaluated is betweeness, which measures how many shortest paths traverse the node. The following snippet calculates betweenness for all nodes in the tree model and selects the 20 nodes with the top betweenness.

```
bb = betweenness(tree_g, v=V(tree_g), directed=TRUE)
top = V(tree_g)$label[order(bb, decreasing=TRUE)[1:20]]
intersect(top, sb_winners)

## [1] "George Seifert" "Tom Coughlin"   "Bill Cowher"    "Mike Shanahan"
## [5] "Pete Carroll"   "Mike Holmgren"
```

There are 6 super-bowl-winning coaches in the top 20. Since there are 427 coaches in the graph and 23 Super Bowl winners, this is a promising result.

4.3 Closeness

Another measure of centrality is closeness, which represents the distance to other nodes in the network. The following snippet applies the closeness calculation.

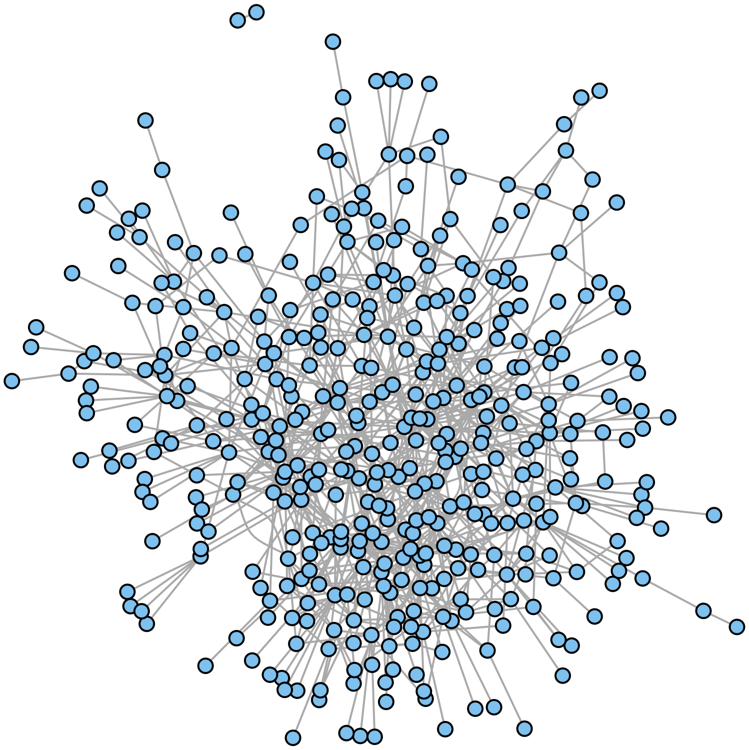

Figure 4: Visualization of the tree model

```
cc = closeness(tree_g, v=V(tree_g), mode="out")
top = V(tree_g)$label[order(cc, decreasing=TRUE)[1:20]]
intersect(top, sb_winners)

## [1] "Bill Cowher"    "Mike Holmgren"  "George Seifert" "Mike Shanahan"
```

The result is 4 Super Bowl winners in the top 20, better than chance, but not as strong as degree.

4.4 Page Rank

Page Rank is an algorithm developed at Google for ranking web pages based on hyperlink structure [6]. Each page's rank is based on the rank of the pages that link to it. This algorithm may help identify successful and influential coaches in a way that other centrality measures don't. Page Rank is calculated in R as follows:

```
pr  = page.rank(tree_g)$vector
top = V(tree_g)$label[order(pr, decreasing=TRUE)[1:20]]
intersect(top, sb_winners)

## [1] "Tom Coughlin"   "Pete Carroll"   "Mike Shanahan"  "Bill Belichick"
## [5] "Mike Holmgren"  "George Seifert"
```

Again we find 6 Super Bowl winners in the top 20, indicating page rank is reflective of coaching success.

4.5 Kleinburg Authority/Hub

Kleinberg's authority score is also based on web page hyperlinks [7]. Like Page Rank, it may be a more accurate centrality measure that more crude measures. It is calculated in R as follows:

```
as = authority.score(tree_g)$vector
top = V(tree_g)$label[order(as, decreasing=TRUE)[1:20]]
intersect(top, sb_winners)

## [1] "Mike Shanahan"  "Mike Holmgren"  "Bill Cowher"    "George Seifert"
## [5] "Mike McCarthy"
```

The top 20 nodes by Kleinberg authority score include only 5 Super Bowl winners.

5 Community Structure

In addition to centrality, the community structure of the NFL coaches network may reflect intuitions around coaching 'families'.

```
clq = largest.cliques(as.undirected(peer_g))
V(peer_g)$label[clq[[1]]]

## [1] "Mike Pettine"         "Rex Ryan"             "Brian Schottenheimer"
## [4] "Dennis Thurman"
```

The largest clique has only 4 members. A less restrictive definition of community is k-cores:

```
cn = graph.coreness(as.undirected(peer_g))
cn[which.max(cn)]

## [1] 5

length(cn[cn == 5])

## [1] 116
```

The largest k-core is only 5 nodes, though 116 nodes belong to a 5-core. It appears that coaches shuffle enough that it is difficult to identify communities. This mirrors the appearance of the graph in figure 4

6 Conclusion

This report analyzed the social network of NFL coaches serving as head coach or coorindator in the years 1980-2013. The two primary questions answered were regarding the alignment of centrality measures with coaching success and the existence of communities revolving around particular coaching styles and/or philosophies.

With respect to centrality, this report finds that multiple centrality measures were well-aligned with coaching success. In the tree model, the top 20 nodes by degree centrality contained 7 Super Bowl winning coaches. In addition, for both betweenness and page rank, the top 20 vertices contained 6 super-bowl-winning coaches. This should not be surprising since winning coaches will tend to have a long career and have their assistants hired away for other jobs, leading to more edges. We haven't gained much insight here; future work should include a full correlation measure between centrality and a more complete metric like winning percentage, as well as normalizing by career length.

With respect to communities, the results were less clear. The visual graph and the k-core measure did not show a significant presence of communities. This part of the analysis might be improved by including more assistant coaches and characterizing head coaches as offensive or defensive based on their previous coordinator positions.

The content used to generate this report can be found at [5].

References

[1] Harrison, C.K. & Associates (2013). Coaching Mobility (Volume I in the Good Business Series). A Report for the NFL Diversity and Inclusion Series.

[2] Harrison, C.K. & Bukstein, S. (2014). NFL Occupational Mobility Patterns (Volume III). A report for the NFL Diversity and Inclusion âĂIJGood BusinessâĂİ Series.

[3] http://www.forbes.com/sites/chrissmith/2013/05/22/the-highest-paid-coaches-in-us-sports/

[4] https://github.com/ndire/pfr-scraper

[5] https://github.com/ndire/sna-nfl-coaches

[6] Sergey Brin and Larry Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the 7th World-Wide Web Conference, Brisbane, Australia, April 1998.

[7] J. Kleinberg. Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. Extended version in Journal of the ACM 46(1999). Also appears as IBM Research Report RJ 10076, May 1997.