Social Sequence Analysis: Assignment 1

Alec McGail

December 12, 2017

1 Methods

My analysis for this homework is based on American football, as inspired by Ben Cornwell's comment that the sequence of pitches in MLB games would be interesting. I think the sequence of plays the teams choose is interesting, and I wonder about the internal structure of these decisions. Admittedly this isn't so social dataset, at most an interacting dyad, but it is a fun one to test my skills on!

The data[1] represents a play-by-play recounting of all NFL games played 2009 - 2017. I focus on *PlayType*, a variable which can take the following values:

> Kickoff, Pass, Run, Timeout, Extra Point, Sack, Quarter End, Punt, No Play,
> Two Minute Warning, Spike, Field Goal, QB Kneel, End of Game, Half End

These each denote a clearly differentiated quality of a well-defined play in a football game. Except for *Sack*, *Quarter End*, *Two Minute Warning*, *End of Game*, and *Half End*, the states represent explicit intent on the part of the offensive team, usually their coach.

Each state does carry with it, in the dataset, the exact amount of time it takes, although I will be neglecting it as part of this analysis. Speaking generally, the sequence is recurrent except a few special cases. *Extra Point*, *Half End*, and *Quarter End* must always be followed by *Kickoff*. *End of Game* cannot have any state follow.

I will use exclusively data from 2016, which includes 256 NFL games in their entirety. Figure 1 shows the frequency distribution of the length, in states, of these games. The sequences are dominated by the states *Pass* and *Run*, as shown in the beautiful figure 2.

[1] https://github.com/ryurko/nflscrapR-data

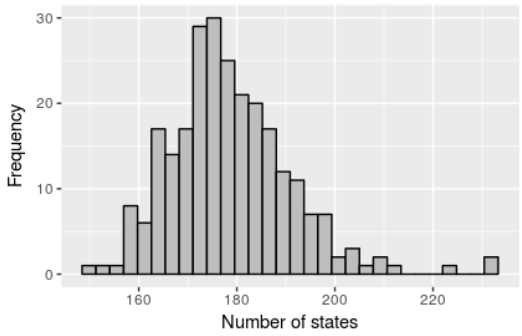

Figure 1: Distribution of length of sequences

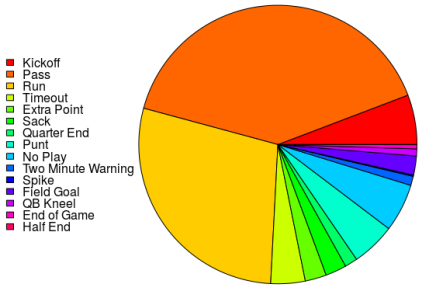

Figure 2: Distribution of states across all sequences

2 Results

My transition matrix is rather large, with 15 different possible states (I could've narrowed it down, but I feel I'd have lost some interesting information, and 15 isn't too much to handle), so I've split it into tables 1 and 2.

One can see that *QB Kneel* leads typically to another *QB Kneel*, or to *EndQ* or *EndG*. *Kickoff* and *Punt* lead very strongly to *Run* or *Pass*, more than any other states. *Sack* leads to *Punt* (more than any other state) or *Pass*, but less so to *Run*. Each of these common transitions correspond to well known situations in a football game, and are quite reasonable. We can also see which states are recurrent by examining the diagonal. Specifically, *Timeout*

(0.02), *Sack* (0.02), *Run* (0.30), *QB Kneel* (0.31), *Pass* (0.39), and *No Play* (0.07) are recurrent.

	EndG	Extra Point	Field Goal	EndH	Kickoff	No Play	Pass
EndG	0.00	0.00	0.00	0.36	0.00	0.00	0.00
Extra Point	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Field Goal	0.00	0.00	0.00	0.00	0.86	0.01	0.08
EndH	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Kickoff	0.00	0.00	0.00	0.00	0.00	0.05	0.44
No Play	0.00	0.01	0.01	0.00	0.00	0.07	0.50
Pass	0.00	0.04	0.04	0.00	0.01	0.06	0.39
Punt	0.00	0.00	0.00	0.00	0.01	0.05	0.43
QB Kneel	0.40	0.00	0.00	0.00	0.01	0.00	0.00
EndQ	0.01	0.00	0.01	0.00	0.28	0.06	0.33
Run	0.00	0.03	0.01	0.00	0.00	0.06	0.45
Sack	0.00	0.00	0.05	0.00	0.02	0.06	0.31
Spike	0.00	0.00	0.09	0.00	0.00	0.07	0.66
Timeout	0.00	0.00	0.06	0.00	0.00	0.07	0.47
2MinWarn	0.00	0.00	0.01	0.00	0.00	0.07	0.50

Table 1: Part 1 of transition matrix

A simple and interesting look at the data is garnered by a frequency plot over time, see figure 3. We can see that the 2-minute warning occurs around T72, and at that time there is a relative explosion in the use of timeouts. At the same time passing becomes much more common than running. The same is true near T184, although we see more variance on the location of the 2-minute warning for the end of the game, as one might imagine. The red near the end of the time series shows where the end of the game typically happens.

	Punt	QB Kneel	EndQ	Run	Sack	Spike	Timeout	2MinWarn
EndG	0.00	0.00	0.27	0.27	0.00	0.00	0.09	0.00
Extra Point	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Field Goal	0.00	0.01	0.00	0.05	0.00	0.00	0.00	0.00
EndH	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Kickoff	0.00	0.03	0.01	0.42	0.02	0.00	0.01	0.00
No Play	0.03	0.00	0.01	0.31	0.03	0.00	0.02	0.01
Pass	0.09	0.00	0.01	0.28	0.03	0.00	0.04	0.01
Punt	0.00	0.01	0.00	0.48	0.02	0.00	0.00	0.00
QB Kneel	0.00	0.31	0.23	0.00	0.00	0.00	0.04	0.00
EndQ	0.03	0.00	0.00	0.25	0.02	0.00	0.00	0.00
Run	0.02	0.00	0.02	0.30	0.03	0.00	0.06	0.02
Sack	0.24	0.01	0.04	0.14	0.02	0.00	0.08	0.02
Spike	0.00	0.00	0.00	0.06	0.03	0.00	0.10	0.00
Timeout	0.08	0.01	0.00	0.26	0.03	0.00	0.02	0.00
2MinWarn	0.05	0.02	0.00	0.33	0.02	0.00	0.00	0.00

Table 2: Part 2 of transition matrix

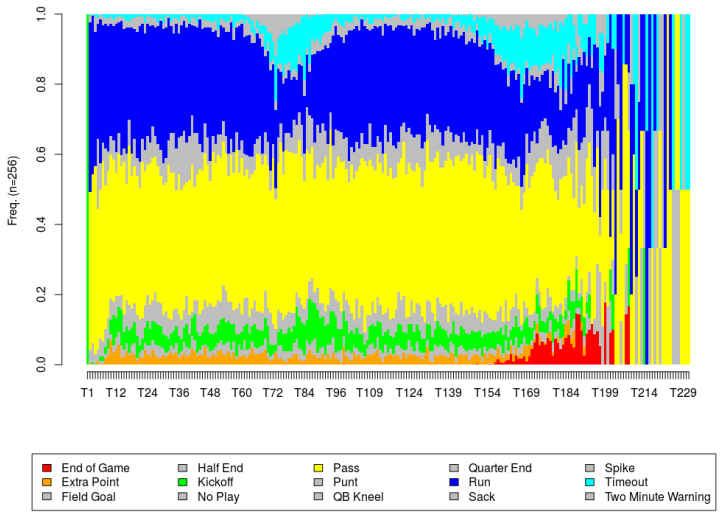

Figure 3: Frequency distribution over time of all games

It shouldn't be too shocking that there is no modal sequence as such. Most sequences don't have the same length, and the sequence is built from an alphabet of 15 states, making the number of possible combinations quite large (about $180^{15} = 6.7 * 10^{33}$, an unimaginably large number). We can, however think of modal subsequences. Specifically, we can

reasonably cut each game into separate "drives," or "possessions," defined as a continuous string of plays which one team carries on, before losing possession. Figure 4 displays the modal sequences, once the data has been cut so. This diagram explains a lot about the game itself. For example, after a turnover a team is most likely to execute the sequence **(Kickoff) Run Pass Pass Punt**. Immediately less likely is the sequence **(Kickoff) Pass Run Pass Punt**. The most common sequence ending in a goal is **Pass FieldGoal**.

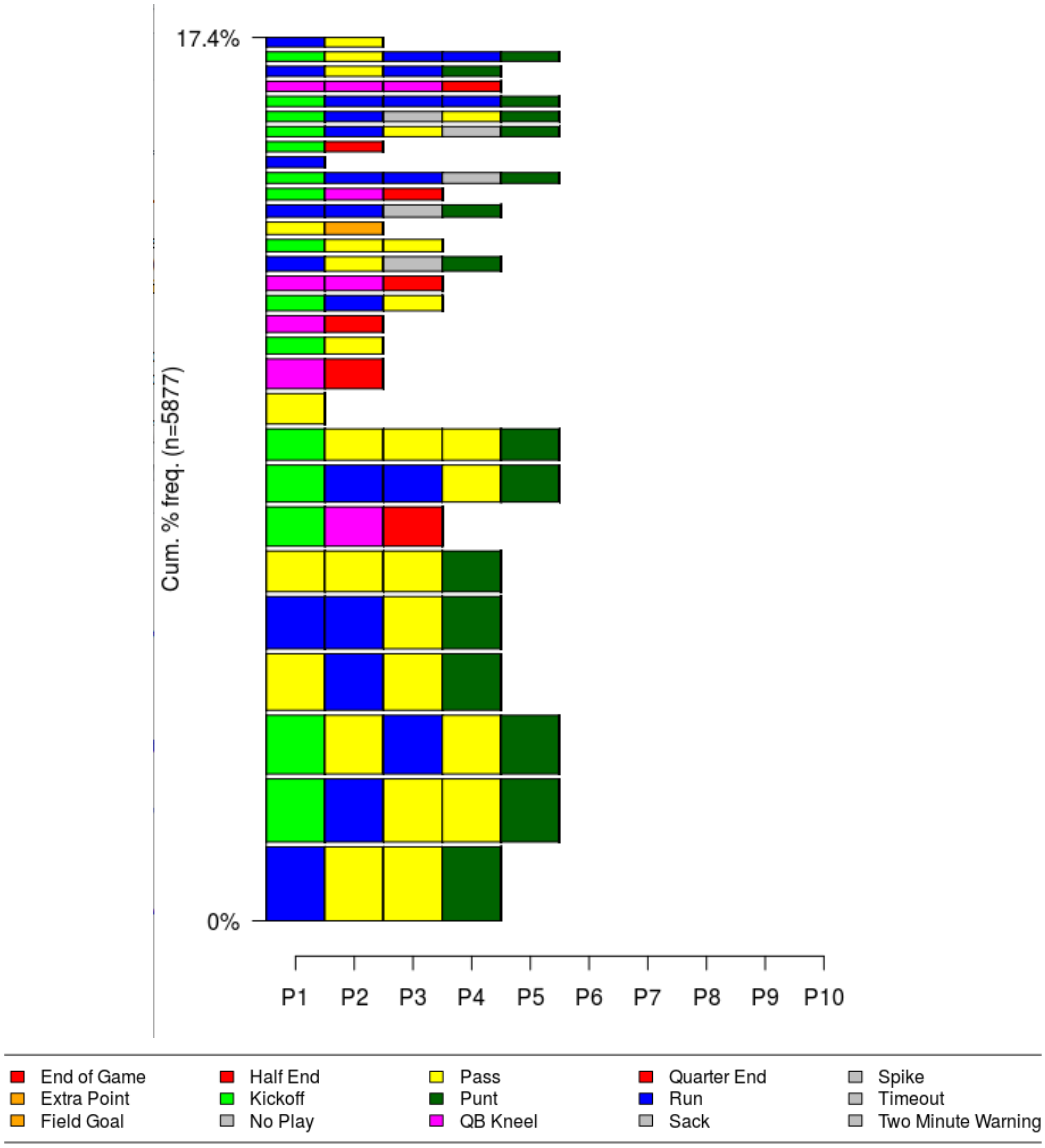

Figure 4: Modal sequences when looking just at sequences inside

It's quite unlikely to have more than 5 consecutive plays (including a kickoff) without turning the ball over, something I intuitively find when watching football.

Some insight can be garnered by examining state distribution graphs, split by the team's first state, shown in Figure 5. For example, it seems more probable to get a touchdown within the first 8 flays if the turnover wasn't by Kickoff, which is intuitively reasonable. If the team starts with a run, they are more likely to choose run in the future, but are less likely to choose it as their second play, than possessions in which the team chose pass first.

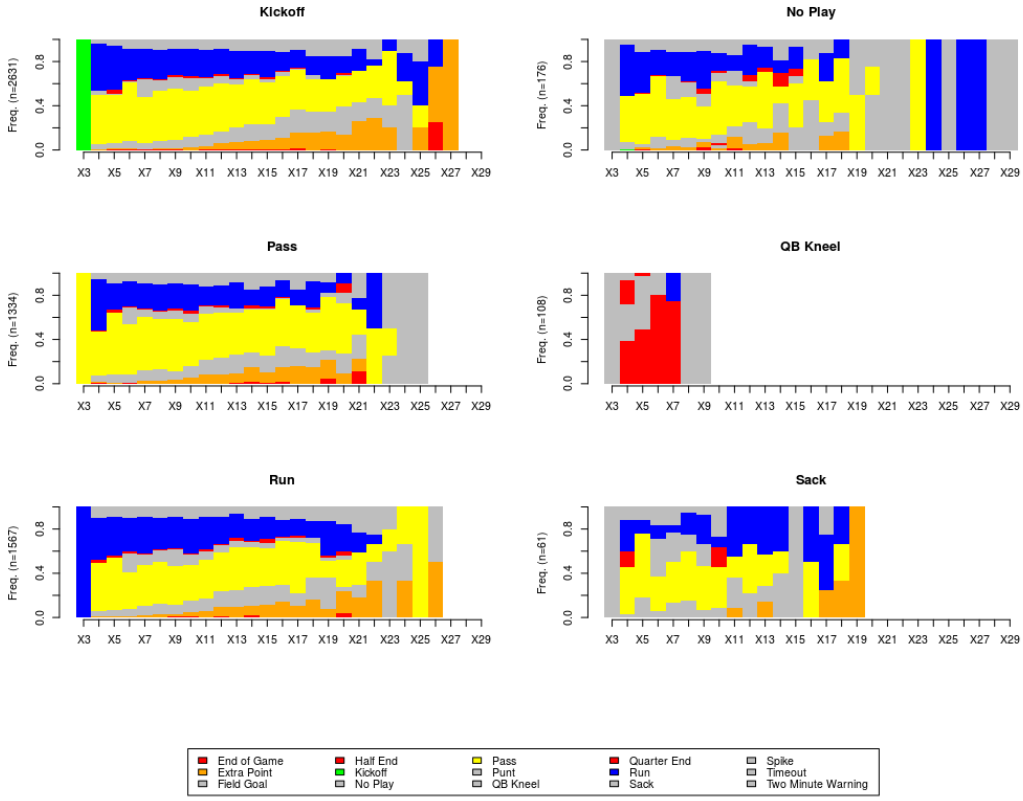

Figure 5: State distribution graph, depending on the first move of the team.

3 Next steps

I'll briefly recount some analysis I wish I did but couldn't get to.

One piece of information which is crucial, but missing, in the current coding of these

states is where the team was on the field when they made the decision, and where they ended up as a result. Multiple methods could be tried for coding this (and I'd likely have to greatly expand my dataset, including 10 years). First, keeping each state as an action, I can code them along with the number of yards gained, binning the number of yards gained by its quartiles. I would only do this for **Pass** and **Run**, bringing the total number of states up to $15 + 8 = 23$.