

The Price of Anarchy in Football

Brian Skinner¹ and Bradley P. Carlin²

¹Fine Theoretical Physics Institute, School of Physics and Astronomy, University of Minnesota,
Minneapolis, MN 55455

²Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN
55455

¹bskinner@physics.umn.edu, ²brad@biostat.umn.edu

A team with a highly-skilled, “superstar” offensive player has a large tactical advantage over its opponent. So why do teams with superstars so often underperform? The answer to this question lies in understanding how a superstar player is best used. A team with a superstar must find the proper balance between putting the ball in the hands of its best player and maintaining enough unpredictability in the offense to keep the defense off-balance. In this chapter we show that some insight into this optimization problem comes from a surprising source: the study of traffic networks. We explain how optimizing the performance of a football offense is analogous to planning the flow of traffic through city streets. The analogy highlights a particular form of short-sightedness that arises in football strategy, and it helps to explain why sometimes a team will *improve* after losing its best player.

Consider the following hypothetical situation, which, on the face of it, has nothing to do with football.

Imagine that in some particular city the suburbs and the downtown are connected by a single, convenient freeway. Every morning and evening the freeway fills up with commuters traveling to and from the city, each of whom (correctly) understands that taking the freeway is the quickest way to get to and from work.

Now imagine that one day the freeway is closed due to construction. The commuters brace themselves for an unusually slow and congested commute, but, to their surprise, they find that by and large their commutes are actually *shorter*. That is, closing down the highway has apparently made the traffic *improve*.

How is this possible?

Now consider a completely different scenario, this time one that is very near to the heart of many college football fans.

Imagine that some particular college football team that has one extraordinarily-talented, “superstar” offensive player – say, the quarterback. Naturally, the offense calls for their star quarterback to throw the ball on the great majority of plays, understanding (correctly) that a throw by their quarterback represents the most effective way to gain yards.

Now imagine that one day the team loses its star quarterback, perhaps to injury or to the NFL draft, and is forced to play the star’s considerably less talented backup. The team and its fan base brace themselves for a serious decline in performance, but, to their surprise, they find

that the team's offense starts playing *even better*, gaining more yards on the average play than they did when their star QB was playing. That is, removing the team's best player has apparently caused the team to *improve*.

How is this possible?

In this chapter we will show that both of these counterintuitive hypothetical scenarios are completely possible, and what's more, that they can happen for the same reason. This reason has nothing to do with the psychology of drivers or of football players facing low expectations. Rather, it relates to the decisions that drivers and football teams make about usage: how drivers use roads, and how football teams use plays. We will show in this chapter that a formal analogy can be made between a football offense and a traffic network. The analogy highlights a particular kind of short-sightedness that arises in football strategy, and it sheds new light on what it really means to give your team the best chance of winning.

Before getting into college football, it's worth spending a bit more time talking about freeway traffic.

In the spring of 1990, New York City's Transportation Commissioner decided to close 42nd Street, one of the city's most perennially crowded roads, in observation of Earth Day¹. The idea, of course, was to encourage New Yorkers to learn to commute without driving by intentionally making the traffic worse for one day. Much to everyone's surprise, however, the traffic on Earth Day was actually significantly *less* congested than average. So apparently New York drivers were better off without 42nd Street at all, giving new meaning to the slogan "every day should be Earth Day."

¹ Gina Kolata, "What if They Closed 42d Street and Nobody Noticed?" [New York Times](#), December 25, 1990.

Actually, the strange improvement of New York City traffic was just one instance of a larger scientific phenomenon called Braess's Paradox, named after the German mathematician who first described it². Stated succinctly, Braess's Paradox says that, under certain conditions, adding capacity to a network can reduce the network's overall performance, and conversely, that removing elements from the network can cause the performance to improve. In terms of traffic, Braess's Paradox means that you can't necessarily expect to make traffic better just by adding new roads, and closing roads doesn't necessarily make traffic worse.

Braess's Paradox isn't just an academic curiosity. It has been observed in a number of cities in addition to New York, including Stuttgart, Seoul, and San Francisco³, and it has important implications for how we route information through computer networks⁴. For these reasons Braess's Paradox continues to be an active topic of study among scientists and mathematicians. It was most recently observed in Minneapolis, where traffic was seen to deteriorate after the re-opening of the collapsed I-35W bridge⁵.

But what causes Braess's Paradox? Actually, the origin of Braess's paradox can be understood by looking at a pair of very simple examples, first suggested by Dietrich Braess in 1968.

Imagine first that ten cars have to get from point A to point B, and that connecting A and B are two roads, as shown in Figure 1. The upper road (labeled 1) is a wide and indirect highway. It takes ten minutes to traverse regardless of the number of cars on it. In other words,

² D. Braess, "Über ein Paradoxon aus der Verkehrsplanung," *Unternehmensforschung* 12 (1969): 258–268.

³ Y. Youn, M.T. Gastner, and H. Jeong, "Price of anarchy in transportation networks: efficiency and optimality control," *Physical Review Letters* 101 (2008): 128701.

⁴ Tim Roughgarden, *Selfish Routing and the Price of Anarchy* (Cambridge: MIT Press, 2005).

⁵ Shanjiang Zhu, David Levinson, and Henry Liu, "Measuring winners and losers from the new I-35W Mississippi River Bridge," in *Transportation Research Board 89th Annual Meeting Compendium of Papers* (Washington: Transportation Research Board).

the commute time L_1 of a driver on road 1 is a constant, $L_1 = 10$. The lower road (labeled 2) is a narrow and direct alley. It generally provides a shorter commute, but it cannot accommodate much traffic without becoming significantly congested. Suppose that on this road the duration of a commute depends on the number of cars x_2 that take the road according to $L_2 = x_2$. That is, if only one car takes the alley, then that car has a 1 minute commute; if two cars take the alley, then they each have a 2 minute commute; and so on.

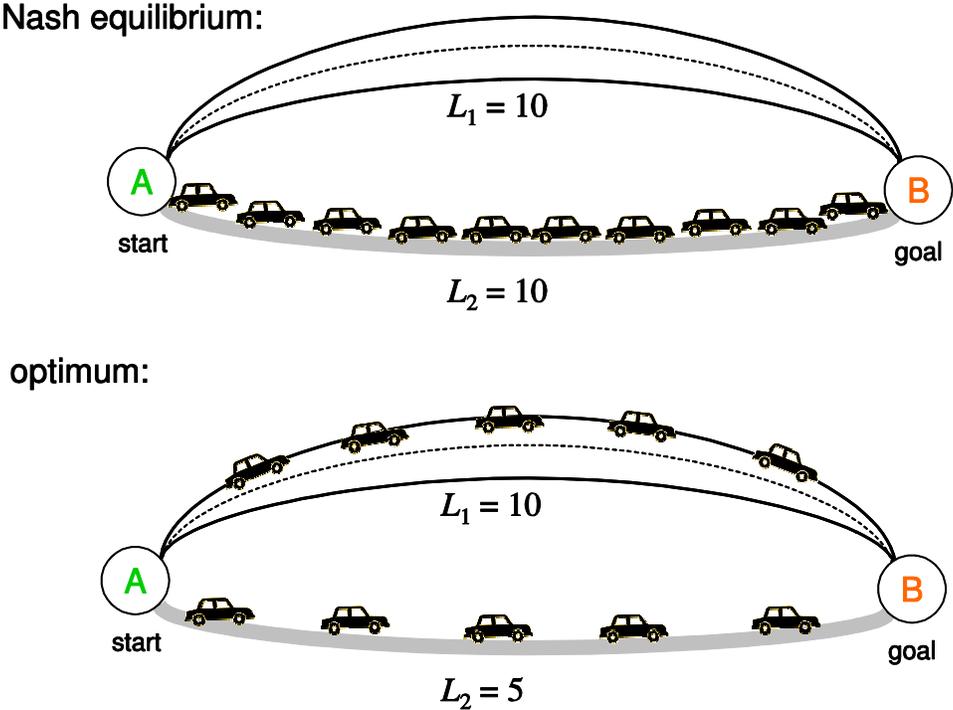


Figure 1: A pair of parallel roads by which drivers travel from A to B. In the Nash equilibrium (top), only the narrow, direct alley is used, and the average commute time is above. In the true optimal distribution (bottom), both roads are used equally. Transferring

any number of cars from one road to another increases the average commute time.

In this simple example, where there are only ten drivers, there is apparently no reason for anyone to take the indirect highway. At worst, driving through the alley can take ten minutes, while the highway always takes ten minutes. It is therefore in the best interest of all ten drivers to take the alley. This situation results in an average commute time of 10 minutes, and is called the “Nash equilibrium,” after the Nobel Prize-winning mathematician and economist John Nash (the subject of the Oscar-winning movie “A Beautiful Mind”). In general, the Nash equilibrium is the state where no individual can improve his situation by making a different choice. In this particular case, it is the state where each driver has the same length of commute (in traffic networks, this state is sometimes also called the “Wardropian User Equilibrium.”)

However, for the drivers in this example, the Nash equilibrium is not actually the best possible outcome. Imagine, for example, what would happen if the town’s road usage was dictated by some hypothetical benevolent monarch. That benevolent monarch could mandate that five of the ten drivers must take the indirect highway, while the other five are free to take the convenient alley. In this case, the five highway drivers would have a ten minute commute, while the alley drivers would have a much shorter five minute commute. The result would be a reduction in the average commute time from 10 minutes to 7.5 minutes. In this way the monarch could alleviate congestion and improve traffic overall, by mandating an optimum distribution among roads.

Of course, this optimal distribution would not be stable in a social sense. In the absence of the benevolent monarch, the drivers will immediately revert to the Nash equilibrium: each

highway driver will immediately want to switch back to taking the alley, and traffic will go back to its previous congested state where everyone has a 10 minute commute. The difference between the “every driver for himself” Nash equilibrium and the true societal optimum is called “the price of anarchy.” In this example, the price of anarchy is an average 2.5 minutes of extra commute time, or 33%. As we will show below, in football it is the coaching staff that must play the role of the “benevolent monarch,” directing the team in such a way as to reduce the “price of anarchy” on the field. But what exactly is meant by “the price of anarchy in football” is not immediately clear.

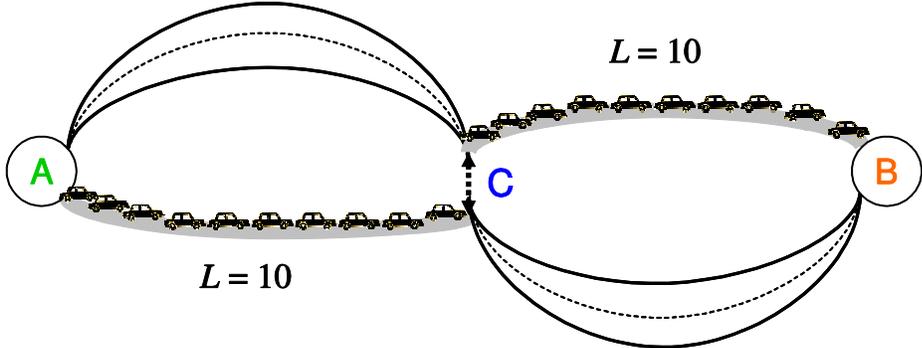
The simple example of Figure 1 is interesting, but its message is not completely counterintuitive: when individuals make sacrifices, the average welfare of the group can improve. Under certain conditions, however, it is possible for something even more dramatic to happen. Sometimes, forcing individuals to act against their own immediate best interest can improve *everyone's* welfare.

Take, for example, the network shown in Figure 2. This example is similar to that of Figure 1: ten drivers commute from A to B by way of indirect “highways” and direct “alleys”. In this example, however, the drivers can choose to switch roads halfway to their destination by using an on/off ramp located at Point C, midway between points A and B.

What will drivers do in this situation? As you would expect, in the absence of any centralized instruction, drivers will look for the quickest path from A to B. Just like in Figure 1, this path involves taking the alley the whole way. The result is that all ten drivers crowd onto the alley for the first half of the commute, then all take the on/off ramp in order to continue taking the alley for the second half of the commute. And, as a result, all drivers end up with 20 minute commutes.

However, as we saw in the previous example, the real global optimum solution is for the alleys and highways to be used equally. If five of the drivers would take the upper roads (highway then alley) and the other five would take the lower roads (alley then highway), then *everyone* would reduce their commute from 20 minutes to 15 minutes. However, as soon as drivers are given the freedom to choose their own routes, then they immediately abandon highways in favor of alleys, traffic again becomes congested, and everyone pays the “price of anarchy.”

Nash equilibrium:



optimum:

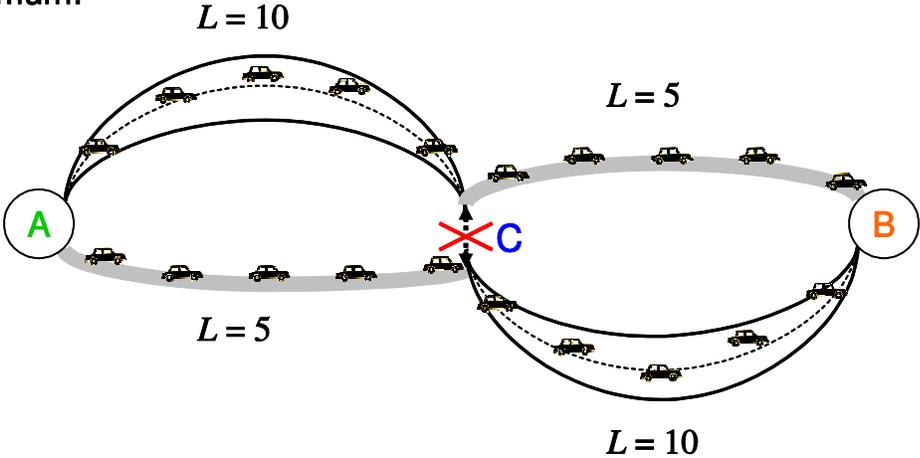


Figure 2: An illustration of Braess’s Paradox. When the central road is open (top), all drivers take the narrow alleys and have 20 minute commutes. When the central road is closed (bottom), the roads are used equally and all drivers have 15 minute commutes.

If you were the hypothetical transportation commissioner in the hypothetical world of Figure 2, all this anarchy would present something of a practical dilemma. What can be done in a democratic society to make drivers cooperate and reduce the price of anarchy? Well, in this particular case, there is actually a surprisingly simple solution: just close the on/off ramp. This closure would immediately reduce the choices for traveling from A to B to only two equivalent options: either drive on the upper roads (highway then alley) or drive on the lower roads (alley then highway). Since these two paths are identical, drivers will naturally split themselves equally between the two, and everyone's commute will be reduced to a happy 15 minutes.

This, in essence, is the origin of Braess's Paradox. It has everything to do with the selfish decisions of individual drivers. When every driver thinks only of minimizing his own commute time, it places a cost on the roadways that is shared by all drivers. Closing a road can inadvertently change how drivers distribute themselves among the roads, and in doing so can place a lower overall cost on the roadways and improve traffic for all.

For researchers, civil engineers, and city planners, reducing the "price of anarchy" on the highway is very much a real problem. Every city has a particular ideal way to distribute drivers among its different roads, but this ideal plan is not necessarily what drivers will naturally follow. So, somehow, drivers must be coerced, or "coached", into driving a particular way, so that the city as a whole will pay a smaller price of anarchy.

And this is where our discussion naturally returns to football.

In football, one can think that the goal of the offense is to acquire yards. Every time the ball is snapped, the offense looks to move the ball as far down the field as possible. To do so, the offense has a number of distinct options. They can try for a long pass down the sideline, or a short pass across the middle, or a handoff to the running back, or a quarterback option. Each of these plays will have a different success rate, as measured by the number of yards it can be expected to produce on average. What's more the success rate of each play can be expected to decline with usage, for reasons that will be discussed below.

In this way running a football offense is a bit like trying to commute to work. Plays are like roads: they represent distinct options (or "paths") for trying to get where you're going. The tricky part is that the more a particular path is used, the less effective it becomes. For this reason one can actually make a very direct analogy between a traffic network and a football offense. Individual play selections are like cars: each one tries to move down the field as efficiently as possible, and must choose carefully which route to take. In principle, all of the strange phenomena associated with traffic patterns can be expected to have some manifestation in football as well.

In traffic, we saw that strange phenomena arise because congested traffic makes roads get worse with increased usage. But what about football plays? Why should they behave like roads? Why, for example, should the halfback draw produce diminishing returns the more often it is used?

Actually, there are a number of reasons why football plays decline in effectiveness. For example, the running back might get fatigued after 20+ rushing attempts. But the biggest reason almost surely has to do with the defense. When a team decides to use a particular play very frequently, it allows their opponents to focus more particularly on stopping that play. If play

effectiveness didn't decline with use, then optimizing the performance of a football offense would be easy: just find the offense's single best play and run it on every single down. But real football strategy is complicated, just like city planning is complicated. In general, you can't expect that having one very good play will be enough to carry you to victory. Running the same play very often gives the defense increased ability to predict what the offense will do, and causes the effectiveness of the play to decline. When football teams say that they need to "keep the defense honest," this is what they mean.

As an illustration, consider the passing efficiency of one of football's most accomplished quarterbacks: Peyton Manning. Over his thirteen-year NFL career to date, Manning has averaged a remarkable 7.6 yards per passing attempt. If this efficiency is broken down by season, however, a clear trend emerges⁶. In seasons where Manning threw the ball a large number of times per game, his passing efficiency was lower than in seasons where he threw the ball less often. This trend is shown in Figure 3. Fitting a straight line to these data (a reasonable approximation here), Manning's yards per passing attempt y is seen to decline with the number of passes x that he throws per game according to the relation $y = 12.4 - 0.138x$. This relation is shown as the dashed line in Figure 3.

While the function $y(x)$ cannot be taken very literally (it fails to control for the quality of Manning's team during a given season, for example, and is unlikely to hold for very large or very small x), it illustrates the general nature of the usage-efficiency tradeoff in football, and we can use it as an example. The problem we want to answer is: what is the best way to use a quarterback with an efficiency function $y(x)$? In other words, if your team has a quarterback like Peyton Manning, how often should you be throwing the ball?

⁶ Data available at ESPN.com: http://espn.go.com/nfl/player/stats/_/id/1428/peyton-manning (2012).

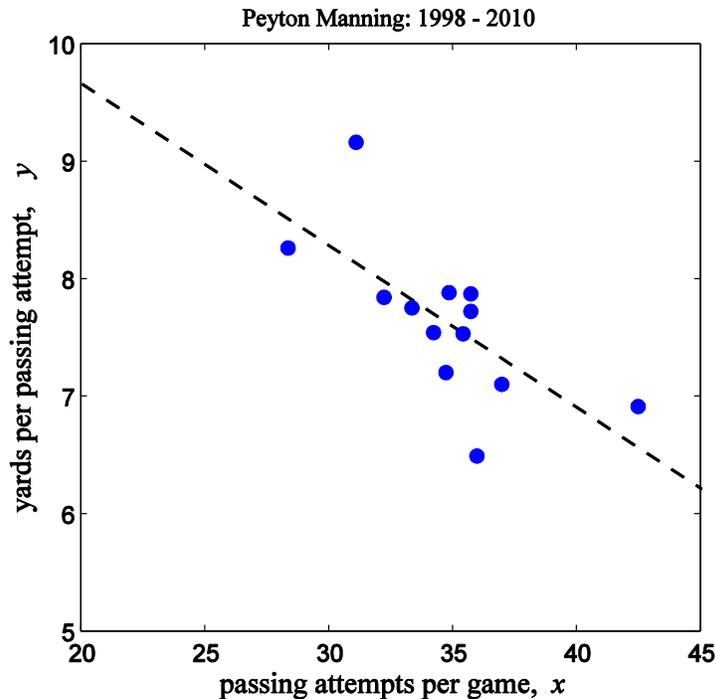


Figure 3: Peyton Manning's passing efficiency as a function of his number of passes per game. Each data point represents one season from Manning's career.

The answer, of course, depends on how good the rest of your team is. Let's think, for simplicity, that your team has only two options on offense: pass the ball or run the ball. To make the problem even simpler, we can think that the running option produces a constant average of 4.5 yards per play. Assuming that you want to maximize the number of yards gained per play, how often should you ask your quarterback to throw?

The most obvious strategy is to say that the quarterback should throw the ball whenever the pass is expected to produce more yards than the run. That is, if the team finds that the pass is generating on average more yards per play than the run, then this strategy calls for the team to call more passing plays and fewer running plays. This process continues until the efficiency of

the pass and run options become equal, at which point the team has supposedly found the optimal balance between the pass and the run.

But is that balance *really* optimal? Consider the problem of Peyton Manning paired with an offense that can produce 4.5 yards per running attempt⁷. If we assume that Manning should throw the ball as long as his efficiency is higher than 4.5 yards/attempt, then by setting $y(x) = 4.5$, we find that his usage rises to $x = 57$ passing attempts per game. This is an enormous number of passes, and corresponds to almost 94% of the team's plays! In contrast, the real Peyton Manning averages only about 35 attempts per game, or about 57% of his team's plays. Clearly, NFL teams are not following the "obvious" strategy. But why not? What is wrong with the logic of "always run the play that is most likely to succeed"?

If we think back on the problem of optimal allocation of cars on city streets, the answer becomes a little clearer. The strategy of always choosing the most effective play is like the Nash equilibrium in a traffic network. In a traffic network, each driver takes the path that is in his/her best interest, and this results in congested traffic and slower-than-optimal commutes. In a football game, when each *play* takes the path that is in *its* best interest, the most effective plays become "congested" and the offense as a whole suffers.

In fact, the optimal balance of running versus passing does not come by always running the most effective play, but by purposefully withholding the team's most effective plays to a certain extent in order to keep the defense off-balance. In this way the team's best plays maintain their high efficiency against the defense, and the offense retains a degree of unpredictability that allows it to run optimally. Mathematically, the optimal pass/run balance

⁷ In general, of course, the running game will also become less effective the more it is used. This is ignored for the sake of simplicity in the present example. For a more detailed treatment, see the following references: Brian Skinner, "The Price of Anarchy in Basketball", *Journal of Quantitative Analysis in Sports*, Vol. 6, Iss. 1, Art. 3 (2010), and Brian Skinner, "Scoring Strategies for the Underdog: A general quantitative method for determining optimal sports strategies," *Journal of Quantitative Analysis in Sports*, Vol. 7, Iss. 4, Art. 11 (2011).

can be found by looking for the maximum of the total number of yards gained per game, which is described by the function

$$Y = x \times y(x) + x_{\text{run}} \times y_{\text{run}}(x_{\text{run}}).$$

In this equation, $y_{\text{run}}(x_{\text{run}})$ stands the average number of yards gained per play when the team runs the ball x_{run} times (in this example, $y_{\text{run}}(x_{\text{run}}) = 4.5$ for all x_{run}). Generally speaking, the sum $x + x_{\text{run}}$ is constrained by the total number of plays available in the game; the NFL average⁸ is $x + x_{\text{run}} \approx 61$.

As it turns out, the function Y has its maximum when the number of passing and running plays are roughly equal⁹: $x = 29$ and $x_{\text{run}} = 32$. At first sight, this seems a bit hard to believe: the quarterback in this example is exceptionally good, and if called upon to pass only 29 times per game will produce 8.3 yards/attempt. The team's running game, on the other hand, is very average, and produces only 4.5 yards/attempt. In this sense it seems obvious that the quarterback should be throwing the ball more. However, the result of limiting the quarterback's throws, and thereby keeping the defense from focusing too intently on the passing game, pays off, and the team performs better as a whole than if the quarterback were to throw more. This can be seen clearly in the graph of Figure 4. Having Peyton Manning throw only 50% of the time (the true optimum) results in the team gaining an average of 6.4 yards per play (388 total), while if he were throwing 94% of the time (the "Nash equilibrium") then the team would gain only 4.5 yards per play.

⁸ Data from teamrankings.com : <http://www.teamrankings.com/nfl/team-stats/> (2012).

⁹ This can be seen by substituting $x_{\text{run}} = 61 - x$ into the expression for Y and taking the derivative of Y with respect to x .

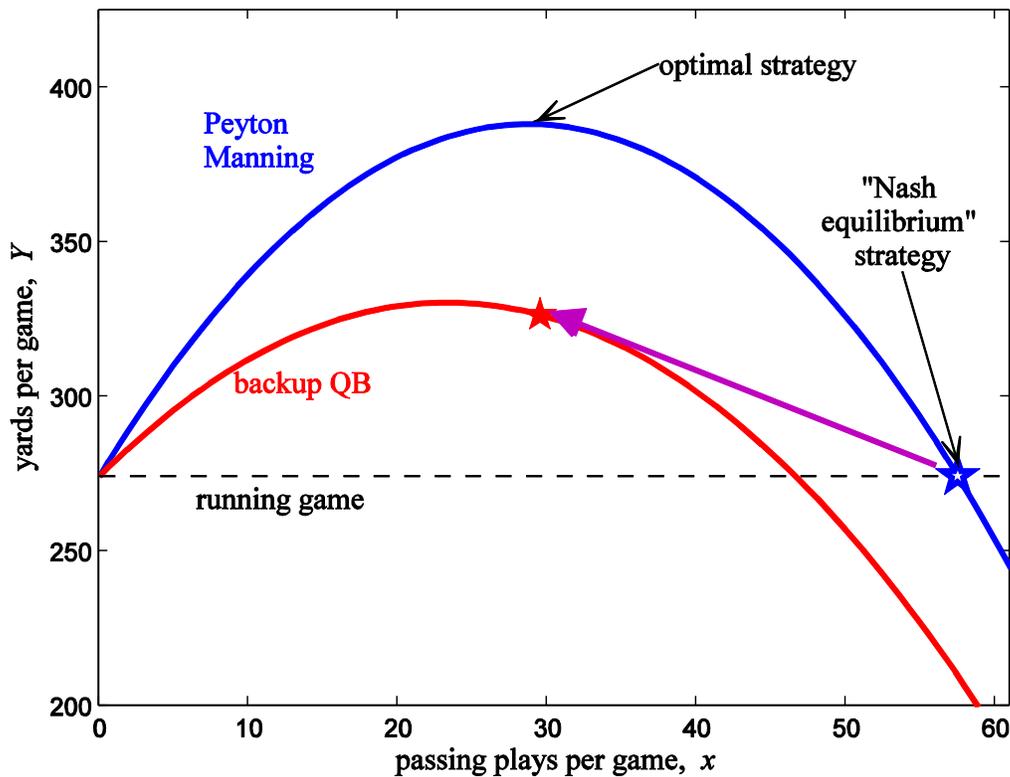


Figure 4: A team's total number of yards per game as a function of how many passing plays it calls. The upper curve shows the team's output with Peyton Manning as their quarterback (see Figure 3), and the lower curve shows the team with the backup quarterback, who is only 75% as effective as Manning. The team can improve when Manning is replaced by his backup if the team's use of the passing play shifts away from the "Nash equilibrium" strategy and closer to the true optimum.

This is an extremely important conclusion, one that seems counterintuitive at first, but is probably not terribly surprising to football coaches. It can be stated in a general way like this:

playing the best possible game is not the same as making the most high-percentage play at each down. Sometimes, in order to maximize its chance of winning, a team needs to *intentionally refrain from running its best plays*, with a view toward keeping the defense from slowing them down. This reasoning may help to explain why, both in college and the NFL, passing plays consistently gain more yards per attempt than running plays. This imbalance has been called “the passing premium puzzle¹⁰,” but one could also say that perhaps football teams and their coaches have learned to largely avoid paying the “price of anarchy” that plagues “selfish” drivers.

As shown above, when a team with a star quarterback does not realize that they need to intentionally limit their passing attack, the team performs sub-optimally. This sub-optimal behavior allows for strange and unanticipated effects associated with changing the offense. The most dramatic effect is an analogue of Braess’s Paradox, which in sports is often called the “Ewing Theory.”

As an example, imagine that some particular team with Peyton Manning as quarterback is generally following the “Nash equilibrium” strategy of simply running the highest-percentage play at each down. On such a team, Peyton Manning will rack up about 260 passing yards per game, but his team will gain only 4.5 yards per play. This state is shown as the open star in Figure 4. Now imagine that Peyton Manning gets injured or traded away, and is replaced by his significantly less skilled backup. Suppose that this backup quarterback can only produce 75% as many yards as Manning at a given level of usage, i.e. his efficiency function $y(x)$ is 25% lower than Manning’s. It is likely that a team with such a new quarterback will use him somewhat

¹⁰ Benjamin C. Alamar, “The Passing Premium Puzzle,” Journal of Quantitative Analysis in Sports Vol. 2, Iss. 4, Art. 5 (2006).

tentatively, limiting his total number of passing attempts and calling more running plays¹¹. If the team lowers their number of passes to, say, only 30 per game, they are likely to find something very surprising. Their total number of passing yards per game will fall, from 260 to 186, but the effectiveness of the pass will shoot up, from 4.5 to 6.2 yards per attempt. As a consequence, the entire offense will perform better, even though the team has replaced their (suboptimally-used) star quarterback with his unambiguously less talented backup.

In sports, the idea that a team can improve after losing its best player is often called the “Ewing theory”, a term coined by the sportswriter/humorist Bill Simmons¹² and named after the perennial all-star NBA center Patrick Ewing, whose team always seemed to perform better when he was sitting out¹³. The Ewing theory is usually explained as a psychological phenomenon, also called the “no one believed in us!” effect. It is likely, however, that many instances of the Ewing theory have a completely non-psychological explanation, one that is rooted more in network theory than sports psychology. Namely, when a star player sits out, it can have the effect of changing how the team uses its plays, and may unintentionally push the team closer to the true optimal strategy.

As for Patrick Ewing, it is probably just a matter of coincidence that on Earth Day, 1990, while New Yorkers were enjoying the unintended positive benefits of a closed-down 42nd Street, Ewing was the leading scorer for the New York Knicks in a loss to the underdog Cleveland Cavaliers¹⁴.

¹¹ This situation is reminiscent of Tim Tebow’s “miracle” year with the Denver Broncos.

¹² Bill Simmons, “Ewing Theory 101”, ESPN.com : <http://sports.espn.go.com/espn/page2/story?page=simmons/010509a> (2001).

¹³ Of course, sports history is rife with plenty of other examples of “Ewing Theory” players, including Wilt Chamberlain, Ken Griffey Jr., and Peyton Manning himself after his graduation from the University of Tennessee.

¹⁴ basketball-reference.com : <http://www.basketball-reference.com/boxscores/199004220CLE.html> (2012).

In closing, it should be said that the arguments we've made in this chapter can't be taken too literally. They are, for the most part, extremely simplified, and they neglect a number of important aspects that are important for real football games. For one thing, we have ignored the important effects associated with risk in football strategy. For example, a passing play might generate more yards on average than a running play, but it also carries a substantially higher risk associated with throwing an interception or an incomplection (i.e., gaining no yards at all)¹⁵. There is also risk associated with fatigue or injury to a player who is overused, which can cause a team to limit the use of its best players for reasons other than optimizing the offense. Most crucially, perhaps, the arguments in this chapter do not duly account for the unique four-down structure of football. In football, different yardage situations (say, 3rd and 1 at your opponent's 7 yard line, as opposed to 1st and 10 at midfield) call for different risk/reward balances, and the goal of winning the game cannot be reduced simply to optimizing the number of yards gained per game.

Still, the parallels between football and traffic networks are compelling enough that there is likely a significant amount of insight to be gained by looking at these two apparently very disparate subjects in the same light. It may well be that in the future a "traffic network" model will be adapted to create quantitative rules that football coaches can actually use when calling plays, although it will of course require a more careful approach than we have presented here. For instance, it might be that such rules could be developed for three separate settings: when the ball is inside the team's own 20 yard line (and thus turnover prevention is paramount), when it is between the 20s, and when it is inside the opponent's 20 (the famous "red zone", where an incomplection on 3rd down is likely to bring on the field goal kicker, rather than the punter).

¹⁵ As Woody Hayes, the legendary former coach of Ohio State, once supposedly said, "There are three things that can happen when you pass the ball, and two of them are bad."

If nothing else, this chapter can perhaps stand as a reminder that in sports strategy, as in science and mathematics, obvious-sounding statements often have surprising consequences. In the study of traffic patterns, the obvious statement is this: selfish driving does not produce optimal traffic. Its surprising consequence is that closing a road can make traffic improve.

In football, the obvious statement goes like this: you can't just run your best play every time; you've got to keep the defense honest. Its surprising consequence is this: if you're trying to optimize each play instead of trying to optimize the game as whole, then your team can improve when a good player is replaced by a worse one.

For sports fans and science nerds alike, what's perhaps most surprising is that both of these strange phenomena occur for the same reason.