Markov Chains in Pop Culture

DARTMOUTH

VOX CLAMANTIS IN DESERTO

1769

Lola Thompson

November 29, 2010

Introduction

There are many examples of Markov Chains used in science and technology. Here are some applications in "pop culture:"

Introduction

There are many examples of Markov Chains used in science and technology. Here are some applications in "pop culture:"

- Modeling the action in a game of Monopoly (for example, we can predict how many times you'll land on a given square)

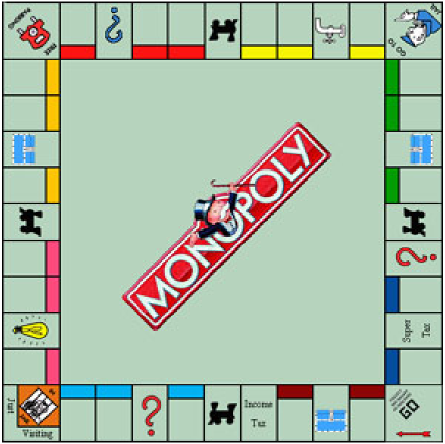
Introduction

There are many examples of Markov Chains used in science and technology. Here are some applications in “pop culture:”

- Modeling the action in a game of Monopoly (for example, we can predict how many times you’ll land on a given square)
- Google’s PageRank Algorithm (Google’s secret to becoming the most successful search engine on the web)

Rules of Monopoly

A Monopoly board has 40 positions:



Rules of Monopoly

- All players start at position 0 (“Go”). At each turn, a player rolls 2 dice and moves according to the sum of their values.

Rules of Monopoly

- All players start at position 0 (“Go”). At each turn, a player rolls 2 dice and moves according to the sum of their values.
- If the player rolls doubles, he rolls again after taking his turn. If he rolls three doubles in a row, then he lands in jail.

Rules of Monopoly

- All players start at position 0 (“Go”). At each turn, a player rolls 2 dice and moves according to the sum of their values.
- If the player rolls doubles, he rolls again after taking his turn. If he rolls three doubles in a row, then he lands in jail.
- If the player lands on a property that is not owned by another player, he has the opportunity to purchase it. If the property is already owned by another player, then he has to pay a fee.

Rules of Monopoly

- All players start at position 0 ("Go"). At each turn, a player rolls 2 dice and moves according to the sum of their values.
- If the player rolls doubles, he rolls again after taking his turn. If he rolls three doubles in a row, then he lands in jail.
- If the player lands on a property that is not owned by another player, he has the opportunity to purchase it. If the property is already owned by another player, then he has to pay a fee.
- The goal is to accrue more money than the other players by the end of the game (which generally happens if you have a "monopoly" on the property on the game board).

Markov and Monopoly

Monopoly is ideally-suited to modeling with a Markov chain:

Markov and Monopoly

Monopoly is ideally-suited to modeling with a Markov chain:

- The states are the squares on the board and each turn corresponds to a step. The transition matrix is a 40×40 matrix containing the probabilities of moving from each square to each other square.

Markov and Monopoly

Monopoly is ideally-suited to modeling with a Markov chain:

- The states are the squares on the board and each turn corresponds to a step. The transition matrix is a 40×40 matrix containing the probabilities of moving from each square to each other square.
- The moves are discrete (determined by the roll of a die).

Markov and Monopoly

Monopoly is ideally-suited to modeling with a Markov chain:

- The states are the squares on the board and each turn corresponds to a step. The transition matrix is a 40×40 matrix containing the probabilities of moving from each square to each other square.
- The moves are discrete (determined by the roll of a die).
- Given a player's starting position at each turn, we can determine the probability that he will land on each of the other 39 positions (states) on the board. The probabilities associated with the steps will be nonnegative and sum to 1.

Markov and Monopoly

Monopoly is ideally-suited to modeling with a Markov chain:

- The states are the squares on the board and each turn corresponds to a step. The transition matrix is a 40×40 matrix containing the probabilities of moving from each square to each other square.
- The moves are discrete (determined by the roll of a die).
- Given a player's starting position at each turn, we can determine the probability that he will land on each of the other 39 positions (states) on the board. The probabilities associated with the steps will be nonnegative and sum to 1.
- For any fixed starting position, the probability of landing on each of the other 39 positions will always be the same (i.e. it will not change based on the number of times that you've landed on this square previously or based on the squares that you visited previously).

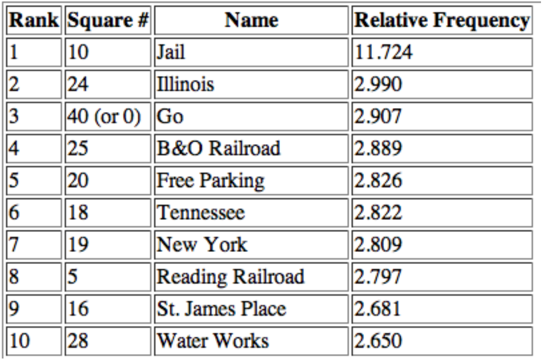
Why Model Monopoly?

Monopoly is a game that is worth modeling because there is some value in knowing how likely a player is to land on a given property. For example, if you're trying to determine a strategy of which properties to buy, it can be useful to know which ones the other players are likely to land on most frequently.

Candy Land, on the other hand, doesn't involve any real strategy so, even though you could assign probabilities to landing on each square, examining it as a Markov chain won't make us any more successful at playing the game.

Surprising Results

The ten most frequently-occurring squares that a player will visit:



Rank	Square #	Name	Relative Frequency
1	10	Jail	11.724
2	24	Illinois	2.990
3	40 (or 0)	Go	2.907
4	25	B&O Railroad	2.889
5	20	Free Parking	2.826
6	18	Tennessee	2.822
7	19	New York	2.809
8	5	Reading Railroad	2.797
9	16	St. James Place	2.681
10	28	Water Works	2.650

(Notice that Jail is #1!)

Strategy Tips to Glean from the Math

You can't own Jail, Go, or Free Parking, but you might as well try to purchase as many of the following as possible: Illinois Avenue, B&O Railroad, Tennessee Avenue, New York Avenue, Reading Railroad, St. James Place, Water Works, Pennsylvania Railroad.

If you're a Monopoly aficionado/a, you may notice that Tennessee Ave., New York Ave. and St. James Place are grouped together on the board (they're orange squares). The Pennsylvania Railroad, Illinois Ave. and B&O Railroad are very close to the orange cluster.

What's going on here???

A Possible Explanation for the Clustering Phenomenon

It's important to remember that Monopoly players spend a lot of time in Jail. Keeping that in mind, let's examine the number of spaces that it takes to go from Jail to some of these properties:

- Pennsylvania Railroad (5)

It's also important to remember that the most frequently-occurring sum of two die is 7, followed by 6 and 8, followed by 5 and 9.

A Possible Explanation for the Clustering Phenomenon

It's important to remember that Monopoly players spend a lot of time in Jail. Keeping that in mind, let's examine the number of spaces that it takes to go from Jail to some of these properties:

- Pennsylvania Railroad (5)
- St. James Place (6)

It's also important to remember that the most frequently-occurring sum of two die is 7, followed by 6 and 8, followed by 5 and 9.

A Possible Explanation for the Clustering Phenomenon

It's important to remember that Monopoly players spend a lot of time in Jail. Keeping that in mind, let's examine the number of spaces that it takes to go from Jail to some of these properties:

- Pennsylvania Railroad (5)
- St. James Place (6)
- Tennessee Ave. (8)

It's also important to remember that the most frequently-occurring sum of two die is 7, followed by 6 and 8, followed by 5 and 9.

A Possible Explanation for the Clustering Phenomenon

It's important to remember that Monopoly players spend a lot of time in Jail. Keeping that in mind, let's examine the number of spaces that it takes to go from Jail to some of these properties:

- Pennsylvania Railroad (5)
- St. James Place (6)
- Tennessee Ave. (8)
- New York Ave. (9)

It's also important to remember that the most frequently-occurring sum of two die is 7, followed by 6 and 8, followed by 5 and 9.

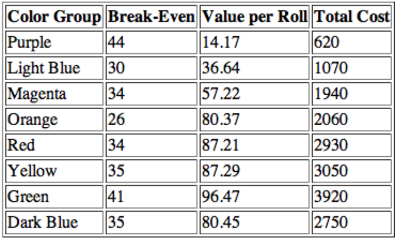
Secondary (and Tertiary, etc.) Effects

Of course, this doesn't tell the whole story, since rolling doubles results in getting a second turn (there are 6 possible pairs of doubles that you can get, making you just as likely to roll doubles as you are to roll a 7!), Jail is not unique in its "clustering" effect (the properties that are 5-10 spaces after Go are also fairly likely to be landed on),...

And then there's the fact that the properties themselves have different costs (to purchase) and demand different fees (from the other players)!

An “Expected Values” Approach

The following chart gives the break even point (i.e. number of rolls until you earn back what you spent on your investment) as well as the expected value per roll (in dollars) for each color group:



Color Group	Break-Even	Value per Roll	Total Cost
Purple	44	14.17	620
Light Blue	30	36.64	1070
Magenta	34	57.22	1940
Orange	26	80.37	2060
Red	34	87.21	2930
Yellow	35	87.29	3050
Green	41	96.47	3920
Dark Blue	35	80.45	2750

Notice that the orange group (St. James Place, Tennessee Ave. and New York Ave.) reaches break even point fastest, but the green group (Pennsylvania Ave., North Carolina Ave., Pacific Ave.) has the highest overall value per roll.

A More Lucrative Application

Sure, it's nice to win at Monopoly, but there's only so much happiness that you can get from paper money. Google has managed to turn Markov chains into fat stacks of legal U.S. tender. We'll discuss the brilliant (yet fairly simple) idea behind Google's powerful search algorithm, now famously called the PageRank algorithm.

The Search Engine Problem

The goal of a search engine is to help an internet user find a website that matches his or her needs as quickly as possible. How is a machine supposed to understand what the human user is actually looking for? With more than 150 million websites and blogs out there, how can a search engine distinguish a reputable website from some crackpot's online collection of personal rants?

Google's Clever Algorithm

The idea behind the PageRank algorithm is similar to the idea of the *impact factor* used to rank journals. The impact factor of a journal is defined to be the average number of citations per recently published paper in that journal. In general, the higher the impact factor, the more "important" the journal (at least, in the eyes of academics).

By regarding each web page as a journal and each link as a citation, Google attempted to measure the importance of a web page by the number of links leading to it. As with journal citations, having a more credible website (i.e. one affiliated with a major corporation or university or government agency) link to your website increases the likelihood that it is "important." As a result, the ranked list of search results that Google returns tends to have the websites with the highest number of credible links (TO their sites) listed first.

Markov Chain for PageRank: The Set-up

- Let N be the total number of web pages on the internet.

Markov Chain for PageRank: The Set-up

- Let N be the total number of web pages on the internet.
- Let k be the number of outgoing links of web page j. (i.e. we can think of web page j as a node and k as its *out-degree*).

Markov Chain for PageRank: The Set-up

- Let N be the total number of web pages on the internet.
- Let k be the number of outgoing links of web page j. (i.e. we can think of web page j as a node and k as its *out-degree*).
- Let P be the *hyperlink matrix*, with elements

$$p_{ij} = \begin{cases} \frac{1}{k} & \text{if webpage } i \text{ is an outgoing link of webpage } j \\ 0 & \text{otherwise} \end{cases}$$

The *hyperlink matrix* can be thought of as the transition matrix for our Markov Chain (as you proved on this week's proof assignment).

Markov Chain for PageRank: How it Works

We can regard a web surfer as a "random walker" and the web pages that the web surfer visits as the states of the Markov chain. We assume that pages that he revisits often must be "important," because they must be pointed to by many other important pages (since his surfing is random).

Assuming that the Markov chain is ergodic and aperiodic (i.e. the powers of P won't form a repeating cycle), then the limiting vector $w = (w_1, w_2, \cdots, w_N)$ exists.

Each w_i is the proportion of the time that the surfer clicks on the link for webpage i. The higher the value of w_i, the more "important" web page i will be. The *PageRank of webpage i* is defined by w_i.

PageRank for Non-ergodic Markov Chains

In the world of web surfing, the Markov chains won't always be ergodic. For example, It might be the case that a website doesn't have any external links (such websites are called "dangling nodes"). This is a common problem - for example, pdf files, image files and data tables are often dangling nodes.

Transition matrices have rows and columns consisting solely of 0's where these dangling nodes occur (hence, these Markov chains won't be ergodic). In order to get around this problem, Google developers perform a *stochasticity adjustment*, in which they simply replace the rows consisting of all 0's in the transition matrix with $\frac{1}{n}e^T$. This forces our new matrix to be stochastic. We'll call our new stochastic matrix S.

The Google Matrix

The new matrix S may be stochastic, but we still aren't guaranteed that it will have the desired convergence results. In order to guarantee convergence, Google developers make a second adjustment: while the random web surfer follows the hyperlink structure of the Web, at times he is bored and abandons the hyperlink method of surfing by entering a new destination into his browser's URL line.

By doing this, the random surfer "teleports" to a new page, where he begins hyperlink surfing again (until the next teleportation). To model this activity mathematically, Google invented a new matrix

$$G = \alpha S + (1 - \alpha)E,$$

where α is a scalar between 0 and 1 and $E = \frac{1}{n}ee^T$ is the "teleportation matrix". The matrix G is called the *Google matrix.*

Finishing Up

The good news is that G is an ergodic, regular, aperiodic Markov chain, so once again we can find a limiting vector.

It has been proven that this method converges on the 'limiting vector' for the transition matrix in approximately 50 iterations (when used over a web data set of over 80 million webpages!). So, it's pretty efficient!

Unfortunately, PageRank is a little outdated at this point (it was created circa 1997 and Googles competition has become stiffer since then). The current Google algorithm is a carefully-guarded secret.

More Information

There are many interesting websites, books and journal articles that discuss these topics in greater detail. A few good starting points include:

Langville, Amy N. and Meyer, Carl Dean. *Google's PageRank and Beyond: The Science of Search Engine Rankings.*

Ash, Robert B. and Bishop, Richard L. "Monopoly as a Markov Process." *Math. Mag.* **45** (1972): 26-29.

Peterson, Ivars. "Monopoly Dollars and Sense."
http://www.maa.org/mathland/mathland_6_9.html

Monopoly probabilities applet:
http://www.bewersdorff-online.de/amonopoly/

Other Courses at Dartmouth

If you found Markov chains interesting, you could also consider taking **Math 100: Markov Chain Monte Carlo** (taught by Dartmouth's own expert probabilist, Pete Winkler) this winter. The only pre-requisites for the course are Math 20 and Math 22.