**The Math of Networks**

|  |  |  |
| --- | --- | --- |
|  |  |  |



Mathematics and Statistics

hugh.chipman@acadiau.ca

 PageRank Graphic by Felipe Micaroni Lalli (micaroni@gmail.com), licensed CC-By-SA

A network:

****

* Could be city streets (many one way)
* Or webpages with links (see “school webpage” example)
* Or a social network

Mathematicians call this a **graph**. It has **vertices** (A, B, ...) and directed **edges** going from one vertex to another.

**Webpages as networks** (an example)

Suppose you’re visiting the website of the ABC school.

* Every page has links leading to other pages
* You click links to navigate around the website
* (look at example in separate file now)

**Webpages as networks** (an example)

This website can be thought of as a graph

|  |  |
| --- | --- |
| * Pages are vertices
* Links between pages are directed edges
 |  |

|  |  |
| --- | --- |
| **A question on the 1st network:**Suppose I am at vertex “A”, and I randomly choose one of the available paths (which are….?) .I then choose another random path for my next step.What is the chance that… |  |

* I am at vertex “B” after my first step?
* After taking 2 steps, I am at vertex B? At vertex D?
* After taking 3 steps, I am at vertex F? At vertex C?

How about the long run?

We could do an experiment to figure this out.

1. Start at vertex “A”
2. Use dice and the chart to take 10 steps.
3. Take 10 more steps after that, marking **each of these 10 steps** in your picture.

When finished, mark your 10 counts on the board



The actual probabilities, after many steps, are:



How were these calculated?

There is a more systematic way to do the calculations. Let’s try a simple example:

|  |  |
| --- | --- |
|  | We can make a probability tree with each level corresponding to a step. Here’s an example (on the next page) starting at “A”: |

(Note: instead of working with probabilities, we’ll imagine a group of 16 people, dividing equally at each branch)

|  |  |
| --- | --- |
|  | Example calculation, starting at A |

|  |  |
| --- | --- |
|  | Now you try it, starting at “B” and considering all possible paths, out to 4 “steps”. As before, start with 16 people at B, dividing them equally at each branch. |

Here’s what we get, taking 4 steps starting from B:

|  |  |
| --- | --- |
|  | This gives…8 at A (8/16 = 50%)3 at B (3/16 = 18.75%)5 at C (5/16 = 31.25%) |

These calculations can be tedious to do by hand, so we’ll use a computer.

* After each step, the computer program groups together all cases of each letter.
* Calculations can be organized using “matrix multiplication”.
* We’ll skip the details

This random walk around the network is an example of a **Markov chain**.

… A sequence of “states” (here A, B, C), observed over time, where the probability of the next state depends only on the current state.

Calculations on a computer:

> g1 = graph.formula(A++B,A++C,B-+C)

> my\_pagerank(g1)

 A B C

0.4444 0.2222 0.3333

|  |  |
| --- | --- |
|  | What do you notice? |

(Note: the “pagerank” values are probabilities after an infinite number of steps)

What are some things you notice?

* At the early steps, probabilities “jump around”
* The probabilities eventually stabilize.
* We spend more time in some vertices than others (why?).

Cool fact: Google uses the same idea to decide which webpages are more “important”

* **Random surfer** model: at each webpage, randomly follow a link. Keep going forever. What percentage of visits are to a particular webpage?
* Google calls this the “PageRank algorithm”

**Here are probabilities for the first graph:**

****

****

**Here are probabilities for the “ABC School website” graph:**

****

 **HOME CALENDAR CLUBS ABOUT CONTACT STAFF TEACHERS**

 **0.366 0.122 0.122 0.122 0.068 0.053 0.085**

 **ATWOOD CAMUS NIGHTINGALE DARWIN CURIE KAHLO**

 **0.011 0.011 0.011 0.011 0.011 0.011**

**Can you predict what will happen for these graphs?**



Divide into 4 groups and discuss / experiment.

* What will be the “steady state” after many steps?
* What vertex will get the most/least visits?

Either try random steps or draw a tree.

RESULTS







Graphs 3, 4 and 5 all have “problems”



Graphs 3, 4 and 5 all have “problems”

* 3 is “periodic” (repeating nonrandom pattern)
* 4 has an “absorbing state” (stuck in C)
* 5 has “disconnected components”

All are problematic for Google’s PageRank

Which of the 3 problems (periodic, absorbing state, disconnected components) occurs in modified graph below?

|  |  |
| --- | --- |
|  | Original graph |
|  | Modified graph |

**Google’s “fix”:**

* With some small probability, we randomly jump to any location, without following the arrows.
* Google uses a 15% chance of random jump at each step.

****

 **A B C D E F G H**

**original 0.060 0.067 0.030 0.068 0.098 0.202 0.180 0.295**

**No B-->D 0.000 0.999 0.000 0.000 0.000 0.000 0.000 0.000**

**rand jump 0.078 0.097 0.062 0.029 0.113 0.182 0.174 0.265**

**Calculating (and using) Google PageRank**

* There are billions of webpages on the internet.
* PageRank is calculated like we did it.
* These calculations are done offline and stored.
* When you search, Google identifies matching pages, and uses saved PageRank to give “top picks”

