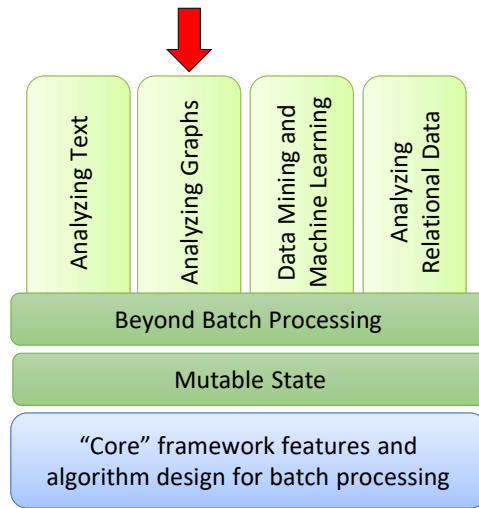


Data-Intensive
Distributed
Computing
CS431/451/631/651

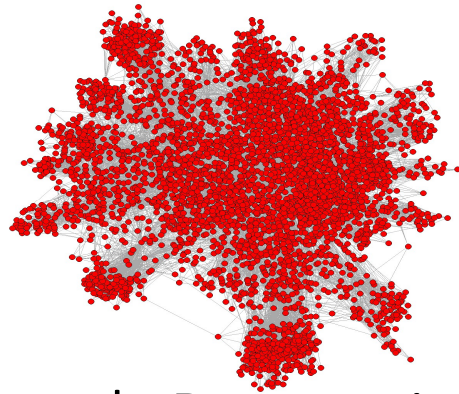
Module 10 – Graphs Redux



Structure of the Course



We're back to graphs!

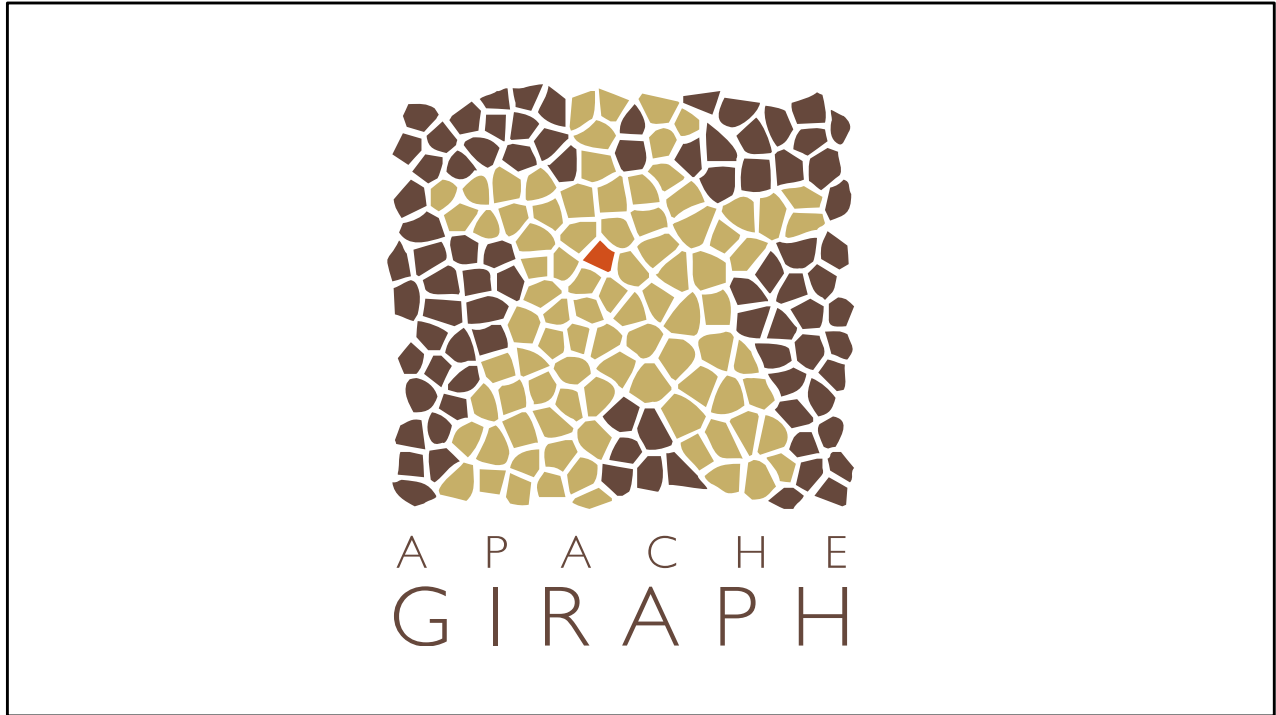


Graph Processing Frameworks

Graph Processing Frameworks

- Pregel
 - Google
- Apache Giraph
 - Based on Pregel
 - On Hadoop
- Spark GraphX





Giraph because:

1. Giraffe sounds like graph
2. Giraffe patterns are basically Voronoi diagrams!
3. Hadoop = animals

What is Apache Giraph

- Giraph performs iterative calculation on top of an existing Hadoop cluster

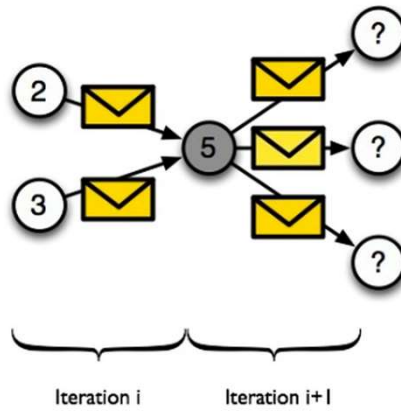


6

Giraph highjacks Hadoop mappers. All iterations are done in memory (and optionally spilled to disk). This is a mapper only job.

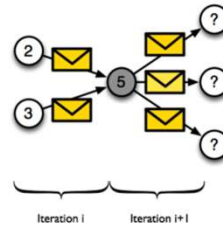
Bulk-Synchronous Parallel (BSP) Programming Model

Vertex-centric model

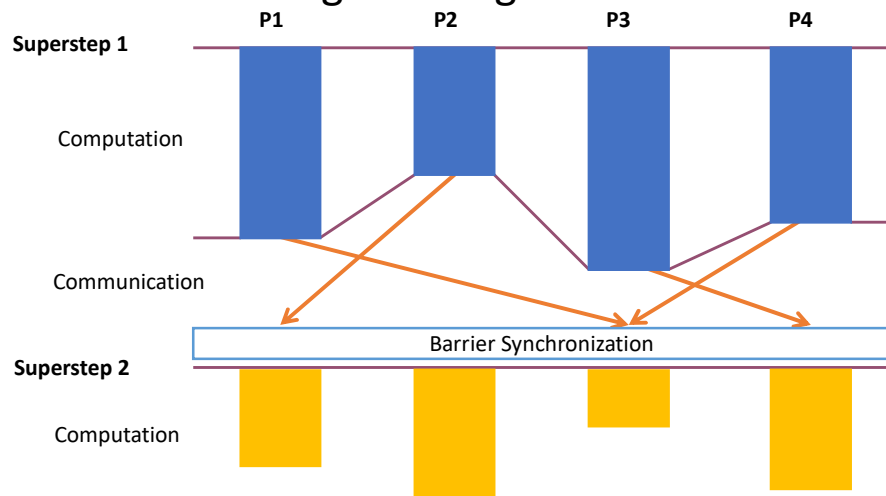


Vertex Centric Programming

- Vertex Centric Programming Model
 - ▶ Logic written from perspective on a single vertex.
 - Executed on all vertices.
- Vertices know about
 - ▶ Their own value(s)
 - ▶ Their outgoing edges



Bulk-Synchronous Parallel (BSP) Programming Model

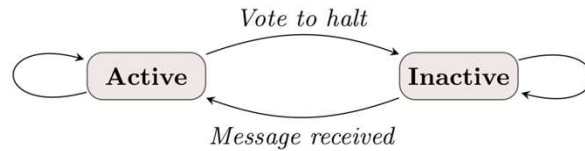


“Often *expensive* and should be used as sparingly as possible”

9

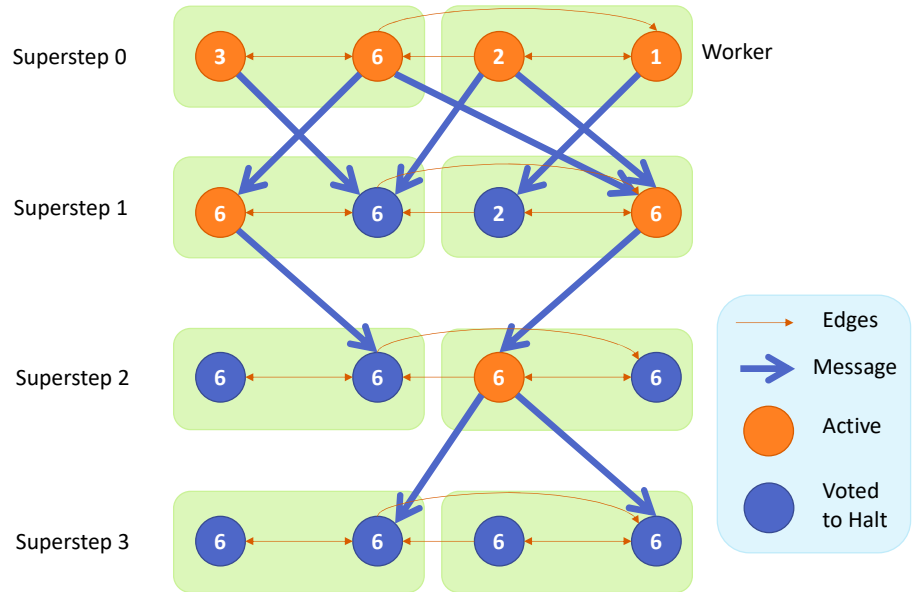
Synchronization is basically a shuffle. You're collecting pending messages for neighbours that lie in other partitions

Vertex State Machine



- In superstep 0, every vertex is in the **active state**.
- A **vertex deactivates itself** by voting to halt.
- It can be reactivated by receiving an (external) message.
- Algorithm termination is based on **every vertex voting to halt**.

Finding the Largest Value in a Graph



Finding the Largest Value in a Graph

```
public class MaxComputation extends BasicComputation<IntWritable, IntWritable,
    NullWritable, IntWritable> {
    @Override
    public void compute(Vertex<IntWritable, IntWritable, NullWritable> vertex,
        Iterable<IntWritable> messages) throws IOException
    {
        boolean changed = false;
        for (IntWritable message : messages) {
            if (vertex.getValue().get() < message.get()) {
                vertex.setValue(message);
                changed = true;
            }
        }
        if (getSuperstep() == 0 || changed) {
            sendMessageToAllEdges(vertex, vertex.getValue());
        }
        vertex.voteToHalt();
    }
}
```

Java again!

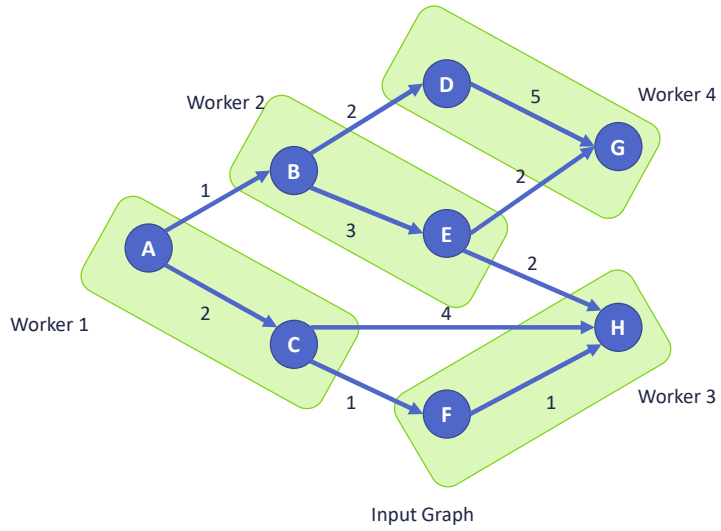
Much shorter than writing MapReduce graph code, though!

(However, for this particular problem, a single MapReduce pass could be done: Map -> smallest value in partition. Reduce -> smallest value overall)

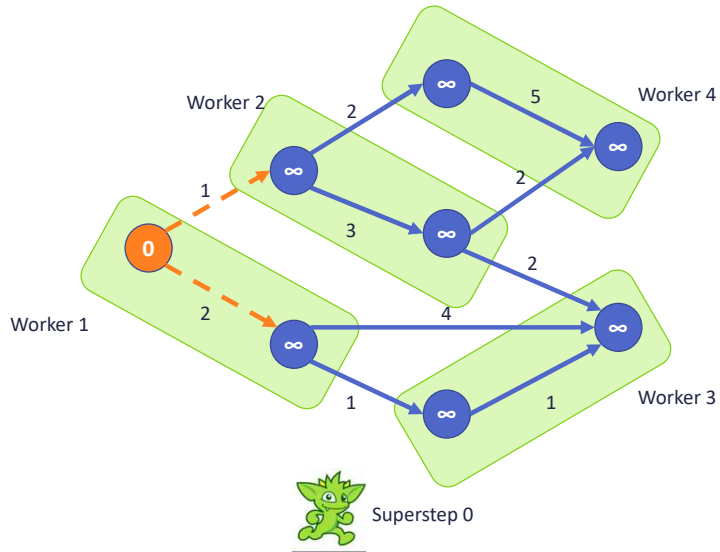
More Applications

Single Source Shortest path (SSSP)

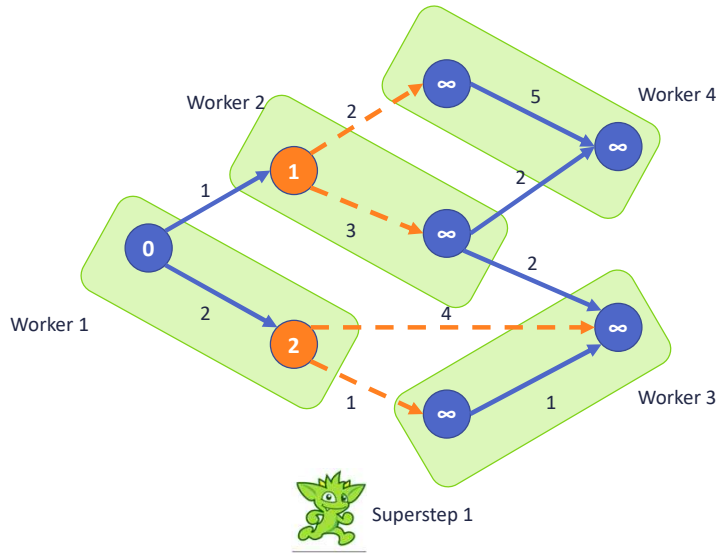
SSSP (1/6)



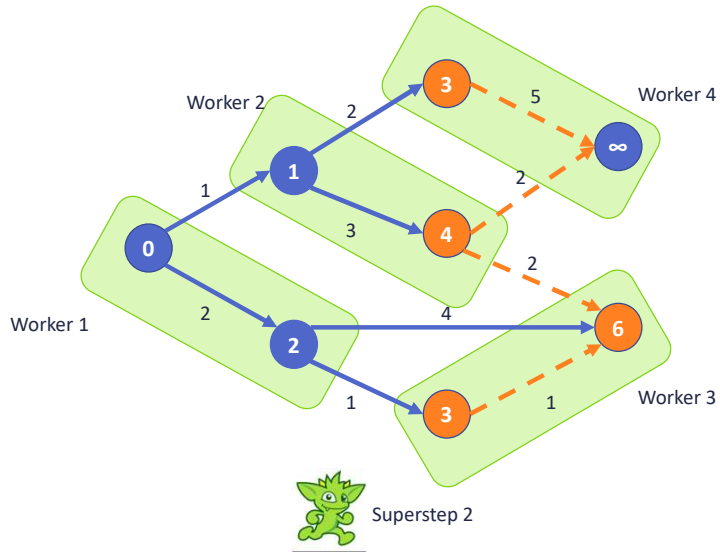
SSSP (2/6)



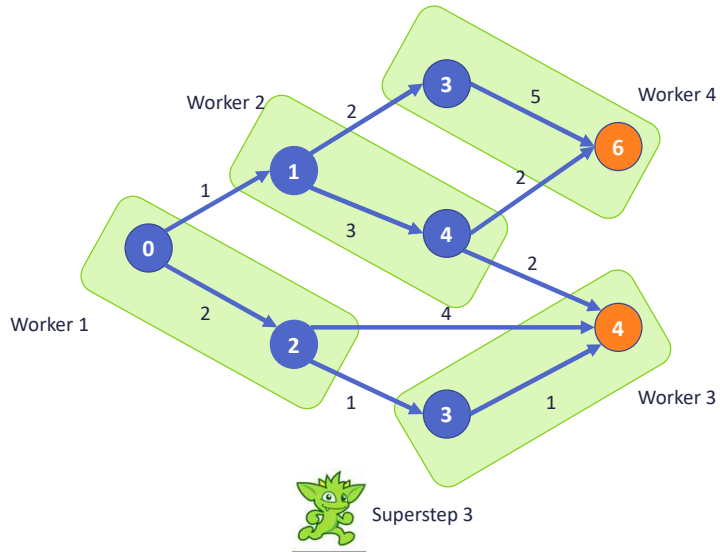
SSSP (3/6)



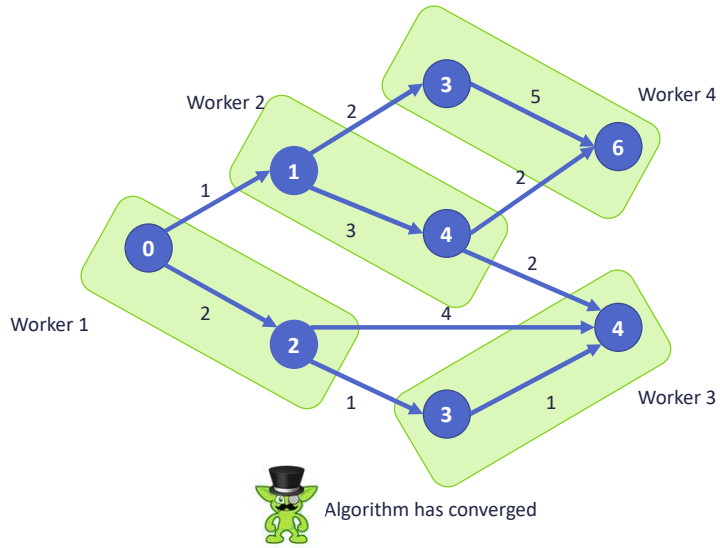
SSSP (4/6)



SSSP (5/6)



SSSP (6/6)



Single Source Shortest path

```
public void compute(Iterable<DoubleWritable> messages) {
    double minDist = Double.MAX_VALUE;
    for (DoubleWritable message : messages) {
        minDist = Math.min(minDist, message.get());
    }
    if (minDist < getValue().get()) {
        setValue(new DoubleWritable(minDist));
        for (Edge<LongWritable, FloatWritable> edge : getEdges()) {
            double distance = minDist + edge.getValue().get();
            sendMessage(edge.getTargetVertexId(), new DoubleWritable(distance));
        }
    }
    voteToHalt();
}
```

Reminder: Even if every node votes to halt, it will only stop if no node has incoming messages. If a node has unread messages, its vote to halt is not counted.

Page Rank

The PageRank algorithm is already formulated with message passing.
Simple to implement in Giraph

Page Rank

```
public void compute(Vertex<...> vertex, Iterable<DoubleWritable> messages) {
    if (getSuperstep() >= 1) {
        double sum = 0;
        for (DoubleWritable message : messages) sum += message.get();
        vertex.setValue(new DoubleWritable(ALPHA / getTotalNumVertices() + BETA * sum));
    }

    if (getSuperstep() < MAX_SUPERSTEPS) {
        int numEdges = vertex.getNumEdges();
        DoubleWritable message = new DoubleWritable(vertex.getValue().get() / numEdges);
        for (Edge<LongWritable, FloatWritable> edge: vertex.getEdges())
            sendMessage(edge.getTargetVertexId(), message);
    } else vertex.voteToHalt();
}
}
```

22

I had to prune a few things to get anywhere close to putting this on screen...

Page Rank – Dead Ends?

```
registerAggregator("missingMass", <aggregatorClass>)
```

Then, in compute:

```
double rank = vertex.getValue().get() +  
    getAggregatedValue("missingMass") / getTotalNumVertices();  
...  
if (numEdges == 0) {  
    aggregate("missingMass", vertex.getValue().get());  
}
```

Giraph vs Spark GraphX

BSP Algorithms are pretty simple with Spark already. GraphX makes them even easier.

According to Facebook:

- Giraph outperforms GraphX, even for small graphs.
- Giraph uses far less memory, meaning a large graph requires fewer workers.

Facebook does a LOT of graph algorithms to do friend suggestions, feed rankings, etc. so you can probably trust their judgement.

- User graph is 1.71 billion vertices, 200+ billion edges.



Remember Search?

Old Search Ranking – TF and DF and logarithms

Flaw: Term Spam.

Set div to not render: Spam spam spam spam spam spam spam spam ...

New Search Ranking – Page Rank

New Term Spam:

A SEO walks into a bar pub inn roadhouse saloon tavern alehouse beer house beer garden
public house drinkery beer ale draught wine...

26

You no longer get benefit from repeating a term, but do want to spam synonyms to increase the chance of containing a term that might be searched for

Solution?

Trust what others say about you, not what you say about yourself:

Use link text (and surrounding text) as terms, instead of contents of page

Remember “tragic love story” vs “starcrossed romance”? Solved.

It has its own problems, though

The screenshot shows a Google search interface with the query "miserable failure". The search results are displayed under the "Web" tab. The top result is "Biography of President George W. Bush" from the official White House website. The second result is "Welcome to MichaelMoore.com!". The third result is "BBC NEWS | Americas | 'Miserable failure' links to Bush". The fourth result is "Google's (and Inktomi's) Miserable Failure".

Web Images Groups News Froogle Local more »

Google miserable failure Search Advanced Search Preferences

Web Results 1 - 10 of about 969,000 for miserable failure. (0.06 seconds)

[Biography of President George W. Bush](#)
Biography of the president from the official White House web site.
www.whitehouse.gov/president/gwbbio.html - 29k - [Cached](#) - [Similar pages](#)
[Past Presidents](#) - [Kids Only](#) - [Current News](#) - [President](#)
[More results from www.whitehouse.gov »](#)

[Welcome to MichaelMoore.com!](#)
Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, ...
www.michaelmoore.com/ - 35k - [Sep 1, 2005](#) - [Cached](#) - [Similar pages](#)

[BBC NEWS | Americas | 'Miserable failure' links to Bush](#)
Web users manipulate a popular search engine so an unflattering description leads to the president's page.
news.bbc.co.uk/2/hi/americas/3298443.stm - 31k - [Cached](#) - [Similar pages](#)

[Google's \(and Inktomi's\) Miserable Failure](#)
A search for miserable failure on Google brings up the official George W. Bush biography from the US White House web site. Dismissed by Google as not a ...
searchenginewatch.com/sereport/article.php/3296101 - 45k - [Sep 1, 2005](#) - [Cached](#) - [Similar pages](#)

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets,
<http://www.mmds.org>

28

By going to forums, blogs, etc. and spamming some phrase you can make Google associate that phrase with the target website. If it already has high page rank, it'll become the top result for that phrase.

(I believe this has largely, if not entirely, been eliminated as a "hack")



Forum Spam /
Comment Spam

What if you go to every page that allows posting, and link to your webpage?

Now YouTube, Facebook, CBC News, etc. all link to you.

They have high rank => You have high rank

You also chose the link text, so you're picking your own terms again





Spam Farming

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets,
<http://www.mmds.org>

Spam Farming Techniques

“Spider traps” accumulate rank.

- Random jumps prevent them from accumulating ALL rank, but it’s still boosted by the topology

Technique:

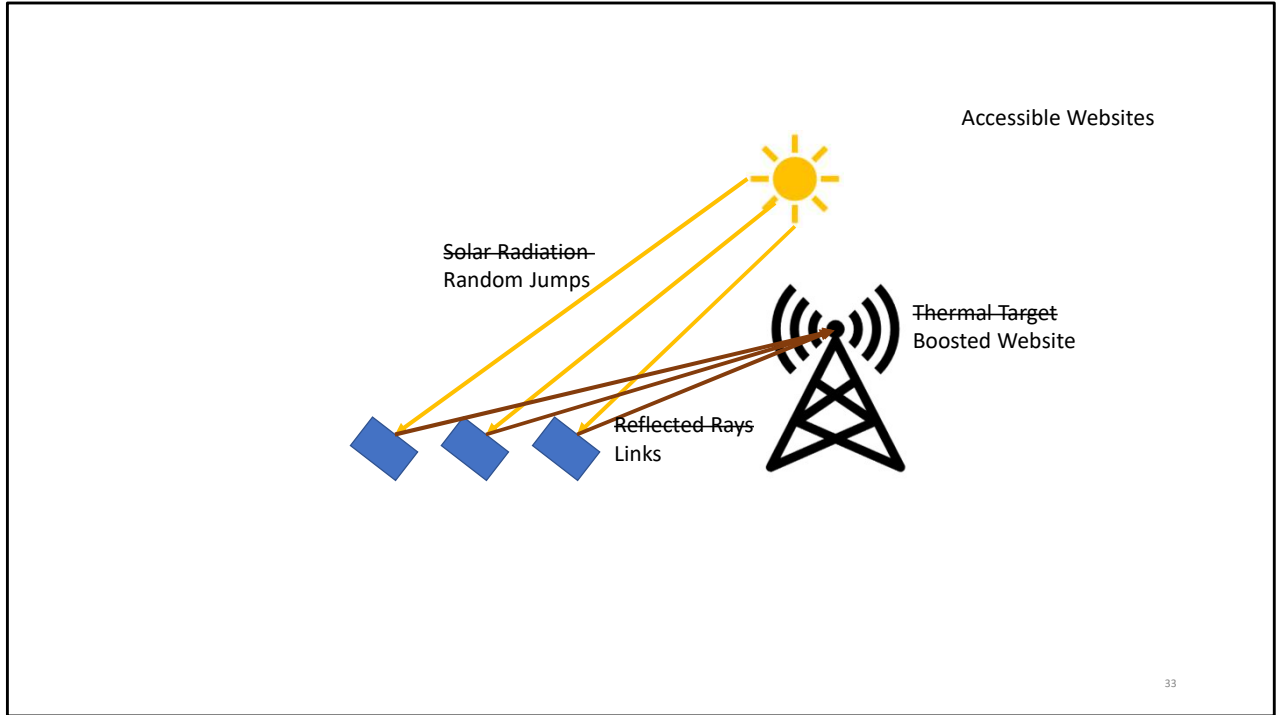
- Page you want to promote has millions of hidden links to farm pages
 - They all accumulate the random-jump weight
- Farm pages all link back to the page you want to promote
 - They send all their rank back to the page being boosted



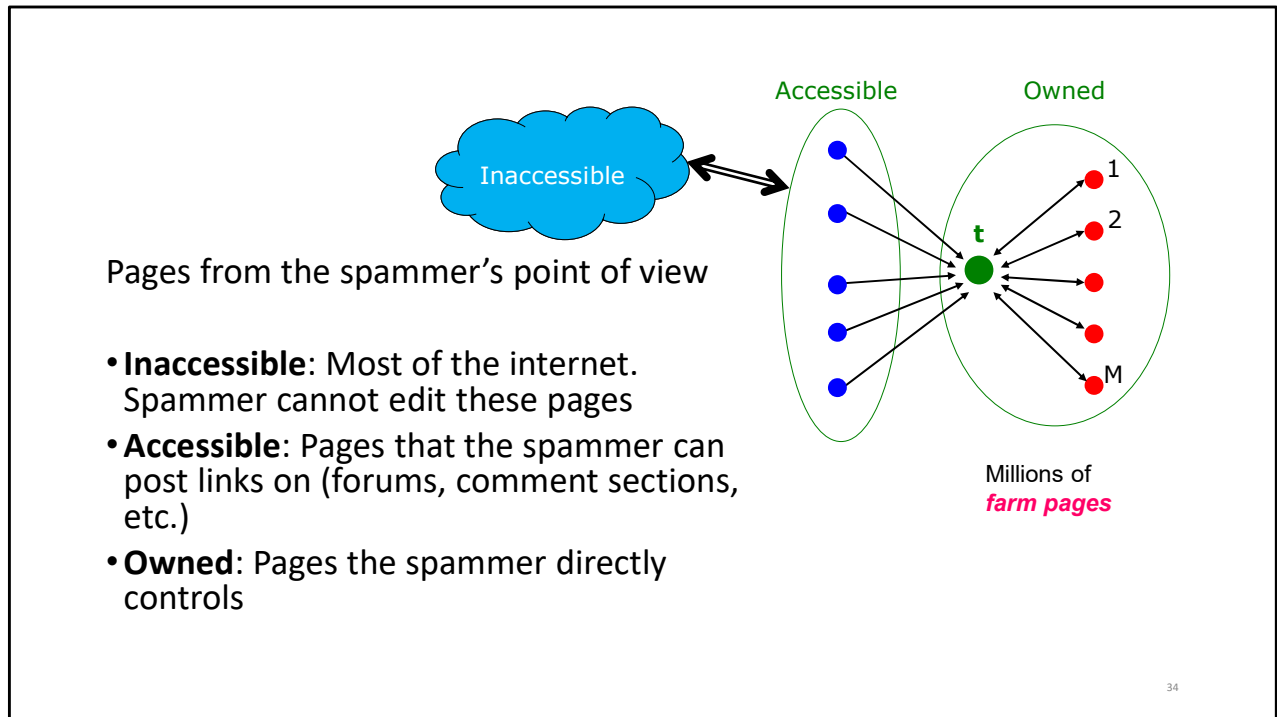
Bogus
Website

Link Farm

32



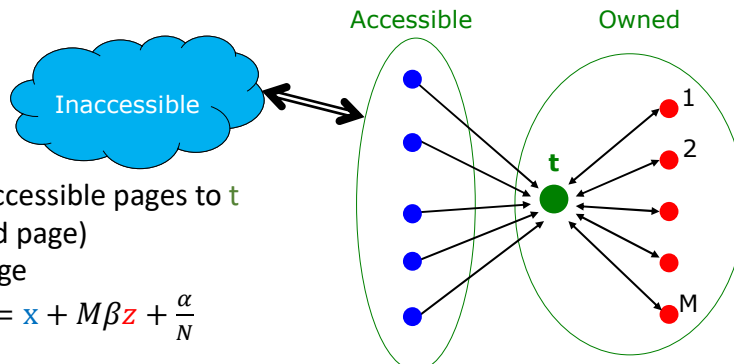
This is just a joke because it reminded me of concentrated solar thermal generators



Random jumps mean that the spider trap gets incoming rank even with no external links. The more websites you can insert links into, the better.

Also the joke solar diagram misses that the target has to link back to the millions of farm pages. That ensures that its accumulated rank stays within the trap. (Websites do of course link to external sites, but even a few dozen links will be minor when divided by millions of total links)

MATH!



x : total contribution of accessible pages to t

y : page rank of t (boosted page)

z : page rank of a farm page

$$z = \frac{\beta y}{M} + \frac{\alpha}{N} \quad y = x + M\beta z + \frac{\alpha}{N}$$

$$y = x + M\beta \left[\frac{\beta y}{M} + \frac{\alpha}{N} \right] + \frac{\alpha}{N}$$

This term is very small, so ignored

Millions of farm pages

3.6x for $\beta = 0.85$
Spider trap amplifies incoming links

$$y = \frac{x}{1 - \beta^2} + \frac{\beta M}{(1 + \beta)N}$$

For $\beta = 0.85$, $0.45M/N$
Grows linearly with M

35

Random jumps mean that the spider trap gets incoming rank even with no external links. The more websites you can insert links into, the better.

Also the joke solar diagram misses that the target has to link back to the millions of farm pages. That ensures that its accumulated rank stays within the trap. (Websites do of course link to external sites, but even a few dozen links will be minor when divided by millions of total links)

Solution to Link Spam

- Ignore links tagged as “nofollow”
- Convince forums, news sites, etc. to insert “nofollow” to all links posted in comments

Added Benefit: A researcher (university website, high rank) can link to a page (to use as an example of term spam) and not boost its ranking

Makes this 0.
Solved?

$$y = \frac{x}{1 - \beta^2} + \frac{\beta M}{(1 + \beta)N}$$

36

Not solved! The rank is directly proportional to M. The spammer can still boost their page rank arbitrarily high by increasing M (which is cheap, just add more dummy pages to the farm and link to them from t



How to Solve a Problem like Spam Farms

Thoughts? What can we do to prevent this sort of trickery?

Any trick to identify the “farm pages” will lead to cat and mouse.

Don't allow small pages to contribute? They'll make them large enough to count.

[Also now recipe pages are forced to include a multi-paragraph story or they'll be null rated]

Can we somehow ensure random jumps do not lead to the farm pages?

Solution to Link Farms



What's the solution?



You did it on the assignment!



In Personalized Page Rank, spam farms don't work. Why?

38

Why: Random jumps do not lead into the farm pages, so they're not accumulating very much mass.

What to use for “Source Nodes”

We should identify “trustworthy” pages

Easy to say...

What’s trustworthy?

Domains with strict entry requirements?

.edu, .gov, etc.

(UW doesn’t make the cut...)

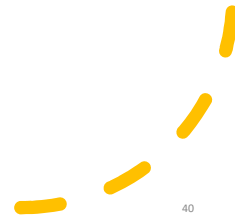


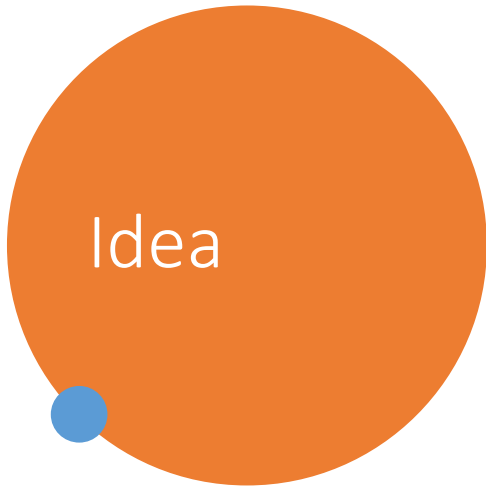
Idea

Collect a sampling of webpages (seed pages)

Oracle (Human) sorts the trustworthy from the spam.

- Expensive
- Keep the set as small as possible





Use the “good” pages as the source nodes for personalized page rank

Small change: Each page in trust set is initialized to 1:

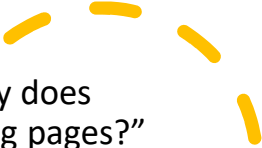
Trust sums to M instead of to 1

After iteration, all pages have a trust factor of between 0 and 1

Pick a threshold and mark all pages below that as spam



Justification

- 
- Mostly the same as “Why does Page Rank find interesting pages?”
 - Trustworthy pages mostly only link to other trustworthy pages
 - Spam pages mostly only link to other spam pages
 - By only teleporting to known good pages, only the “good” partition accumulates significant trust

Conflicting Interests

- The more seed pages there are, the most time and effort needed to curate
- The fewer seed pages there are, the less trust there is in the system
 - Threshold will need to be lower, more spam pages slip through
- Need to pick seed pages that are highly likely to point to “most” of the other good pages
 - Your Trust is roughly proportional to your link distance from a “good” seed page.

Picking Good Seeds

Pick the top k pages by Page Rank

- Assumption: even with link farms, bad pages won't be in the top k

Use Trusted Domains

- Can't get a .edu, .mil, .gov domain just by buying one!
- But, in fact, you can be trustworthy without being the US Government or a US University

Scandalous claim!

So what's the solution?

Bootstrapping Trust

- If your seed set is small, will miss a lot of trustworthy sites
 - Alternate View: You will only catch a small number of spam farmers
 - BUT: Anything that you find is probably pretty trustworthy
 - Candidates to be added to the trusted set.
1. Run PageRank
 2. Select top pages as seed (and verify trustworthiness)
 3. Run TrustRank
 4. Set threshold low enough to avoid false positives
 5. Remove spam pages from graph
 6. Goto Step 1

Alternative – Spam Mass

r_p = PageRank of Page P

r_p^+ = PageRank of Page P, but random jumps only lead to **Trusted** pages

$r_p^- = r_p - r_p^+$ = Contribution of spammers to Page P's rank

$S_p = \frac{r_p^-}{r_p}$ = Spam Mass (Fraction of P's rank that's due to spam)

The higher your Spam Mass, the more likely you are to be spam

46

Question: Can this be exploited?

Can you target a business and make them look like spammers?

No: If you target Wikipedia with a spam farm, almost all of its rank is coming from elsewhere. You can't form the proper spider trap topology without Wikipedia pointing to all of your farm pages (and hopefully editors won't let you do that)