Week 11: Pagerank & Final Review

Data Science Certificate Program

Ryerson University

Bora Caglayan

9 Apr, 2015

Outline

- Pagerank Algorithm
- Review for the Final
- Automation of Scripts using R

Duffy diaries long on party business, short on sober second thought

- **RELATED** Lawyer signals Duffy is ready to testify in his own defence

Shell-BG merger a game changer for B.C.'s LNG industry

- **RELATED** Deal could push others in oil patch to rethink growth

Top executives scrutinized as part of Amaya investigation

Canadian jets drop first bombs on Islamic State stronghold in Syria

Jury to decide if Boston Marathon bomber will face death penalty

New Zealand rejects Canadian's take of Bain's murder conviction

LIFE VIDEO Untapped real estate potential: How cities like Toronto are missing out by ignoring laneway housing

GLOBE EDITORIAL

Balanced budget bill: Great politics, bonehead economics

102

MARCUS GEE

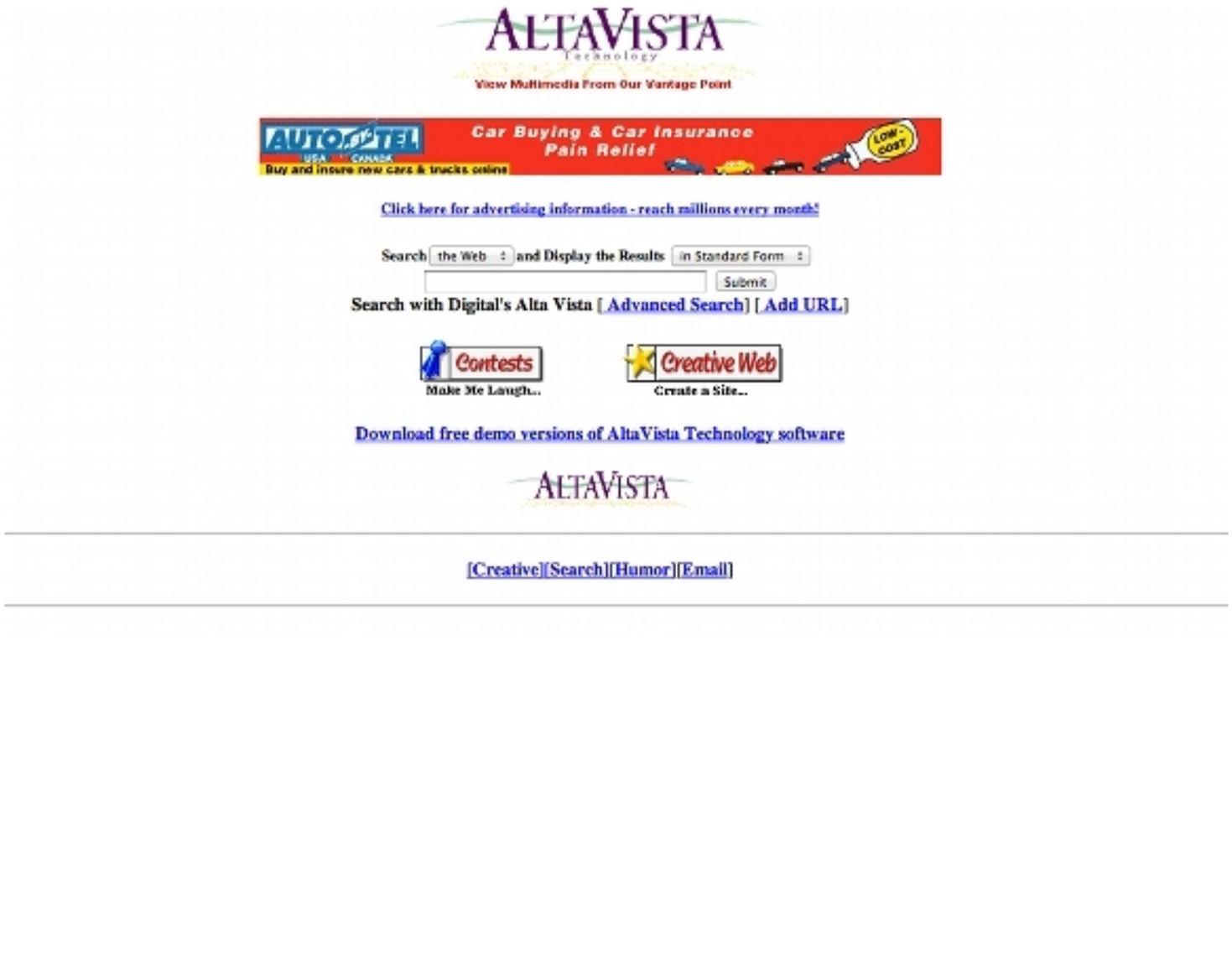
Why the TSO set a terrible precedent by barring pianist Valentina Lisitsa

190

Early Search

- **Arts** - - *Humanities*, *Photography*, *Architecture*, ...
- **Business and Economy** **[Xtra!]** - - *Directory*, *Investments*, *Classifieds*, ...
- **Computers and Internet** **[Xtra!]** - - *Internet*, *WWW*, *Software*, *Multimedia*, ...
- **Education** - - *Universities*, *K-12*, *Courses*, ...
- **Entertainment** **[Xtra!]** - - *TV*, *Movies*, *Music*, *Magazines*, ...
- **Government** - - *Politics* ***[Xtra!]***, *Agencies*, *Law*, *Military*, ...
- **Health** **[Xtra!]** - - *Medicine*, *Drugs*, *Diseases*, *Fitness*, ...
- **News** **[Xtra!]** - - *World* ***[Xtra!]***, *Daily*, *Current Events*, ...
- **Recreation and Sports** **[Xtra!]** - - *Sports*, *Games*, *Travel*, *Autos*, *Outdoors*, ...
- **Reference** - - *Libraries*, *Dictionaries*, *Phone Numbers*, ...
- **Regional** - - *Countries*, *Regions*, *U.S. States*, ...

Early Search



Ranking Nodes

- How can we rank the nodes based on their relative importance?

  - Degrees - Too simple

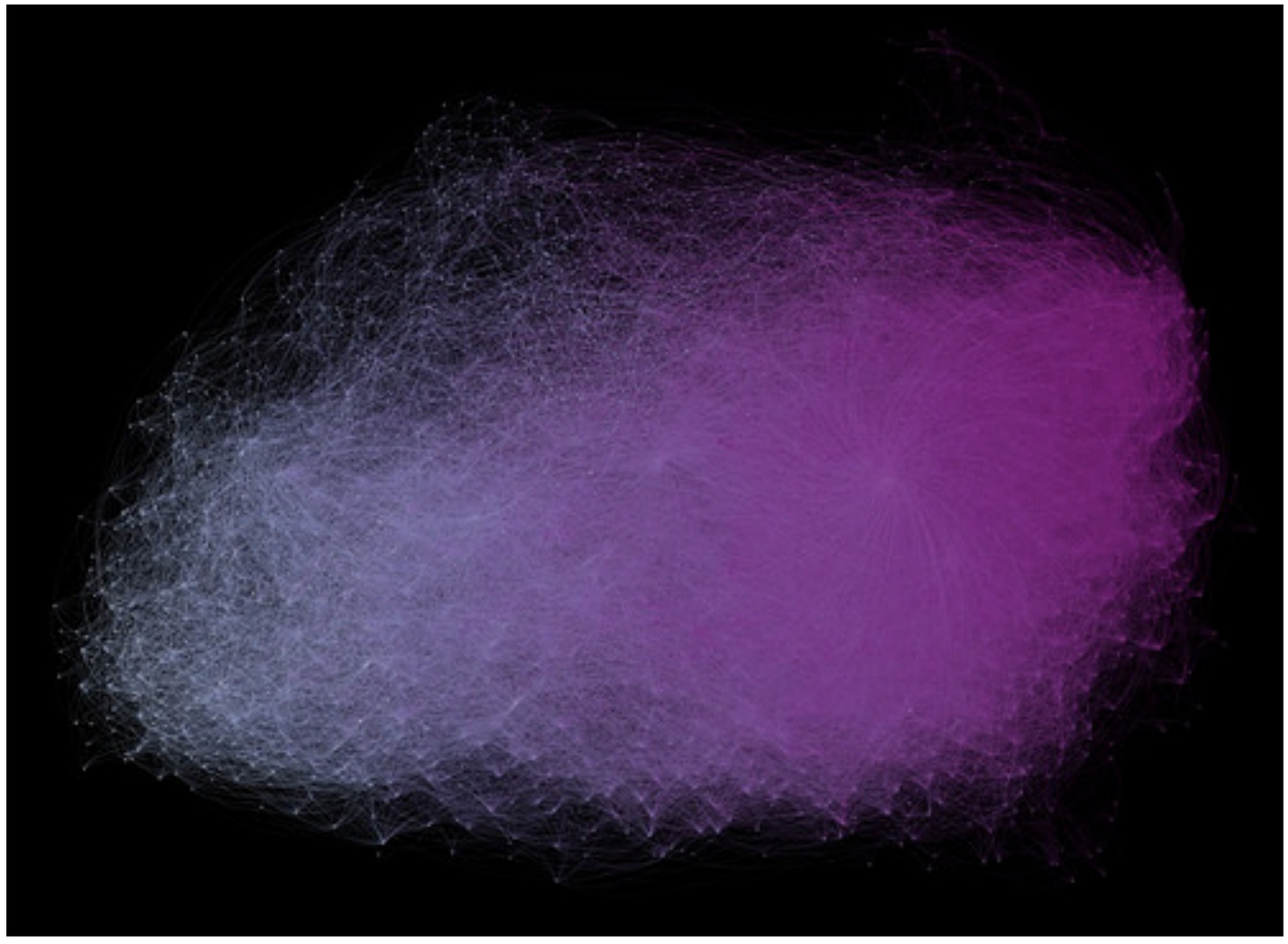
  - Betweenness values - Too hard to compute for large networks since it is calculated *globally*.

  - Closeness centrality - Again calculated globally.

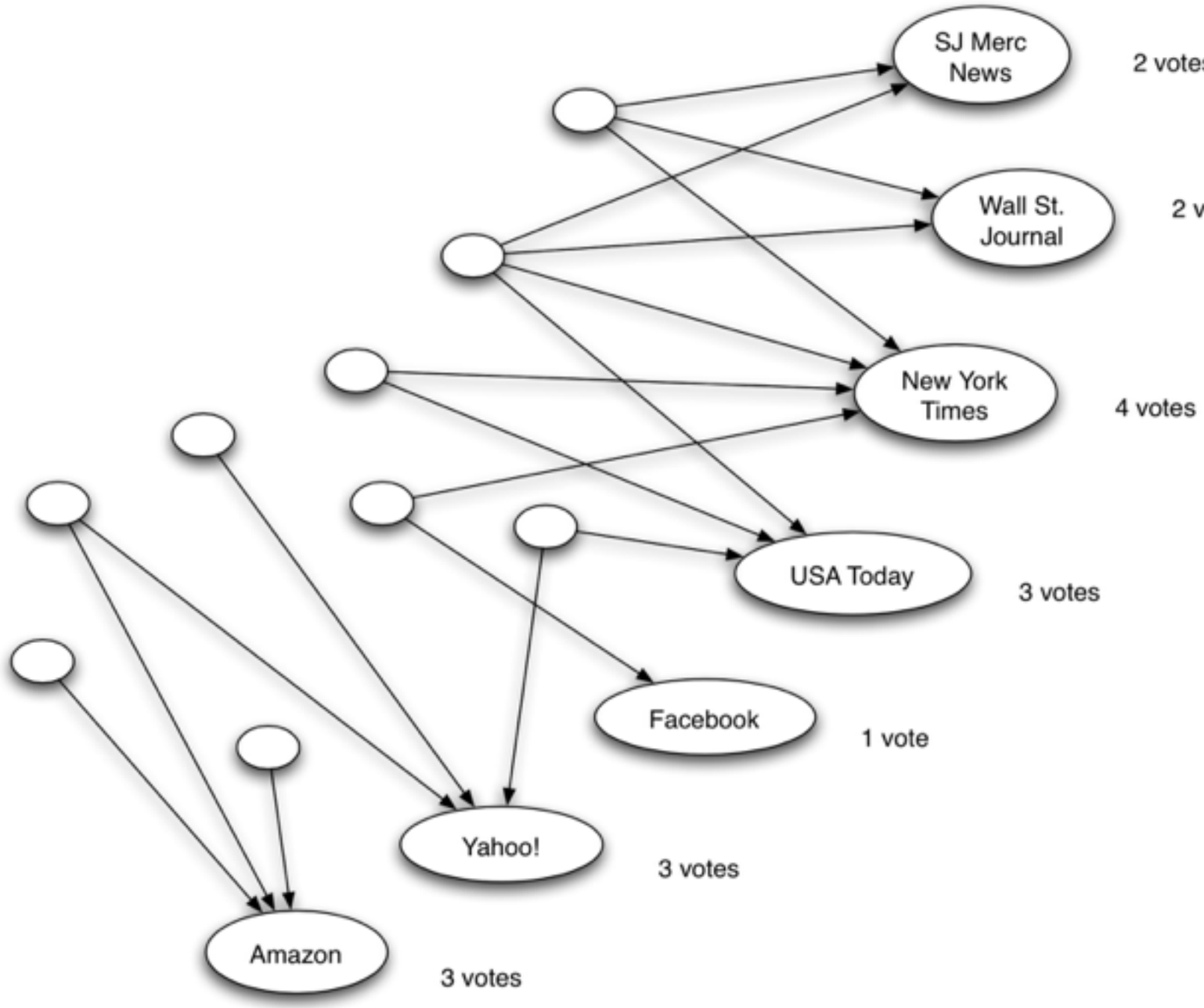
  - …

Ranking Nodes

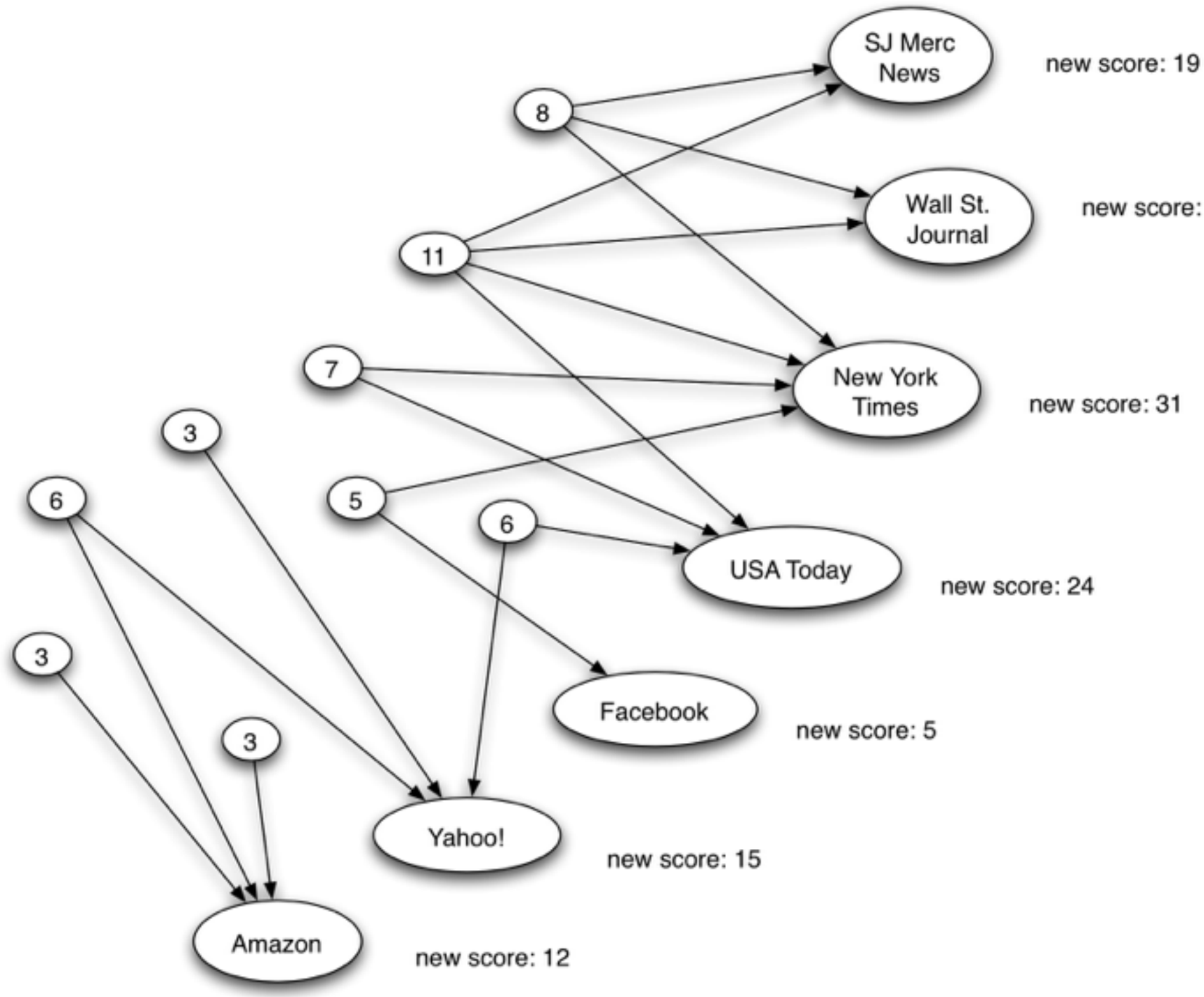
Tip: We have to use local algorithms to analyze massive networks.



Ranking Nodes - In Degree



Ranking Nodes - Weighting Nodes



Motivation for Page Rank

User scenario

- A web user either follows the links or hops to another unlinked webpage or stops web usage.
- Let's simplify this a bit.

Flow of Random Agents

- Assume there are M agents where (M >> Number of nodes) that start from random nodes at $T_{0.}$
- In each iteration:
  - The agents hops to a random outgoing link with probability **P**.
  - The agents hop to another random node with probability **1-P.**
- Where would these agents be at $T_{infinity}$?

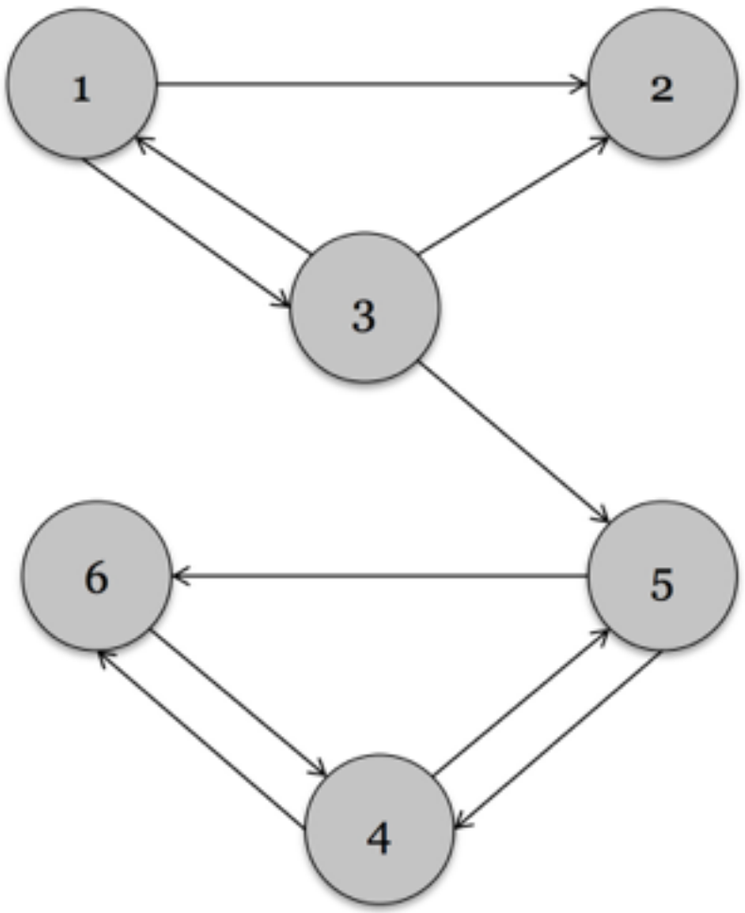
Pagerank

- In a network with n nodes, we assign all nodes the same initial PageRank, set to be $1/n$.
- We choose a number of steps k.
- We then perform a sequence of k *updates* to the PageRank values, using the following rule for each update:

> *Basic PageRank Update Rule:* Each page divides its current PageRank equally across its out-going links, and passes these equal shares to the pages it points to. (If a page has no out-going links, it passes all its current PageRank to itself.) Each page updates its new PageRank to be the sum of the shares it receives.

Pagerank

Example



Iteration 0	Iteration1	Iteration2	Rank at iteration 2
R0(P1)=1/6	R1(P1)=1/18	R2(P1)=1/36	5
R0(P2)=1/6	R1(P2)=5/36	R2(P2)=1/18	4
R0(P3)=1/6	R1(P3)=1/12	R2(P3)=1/36	5
R0(P4)=1/6	R1(P4)=1/4	R2(P4)=17/72	1
R0(P5)=1/6	R1(P5)=5/36	R2(P5)=11/72	3
R0(P6)=1/6	R1(P6)=1/6	R2(P6)=14/72	2

Pagerank

A Question: *Sink node Case*

What happens if there is a node with 0 out-degree in the network?

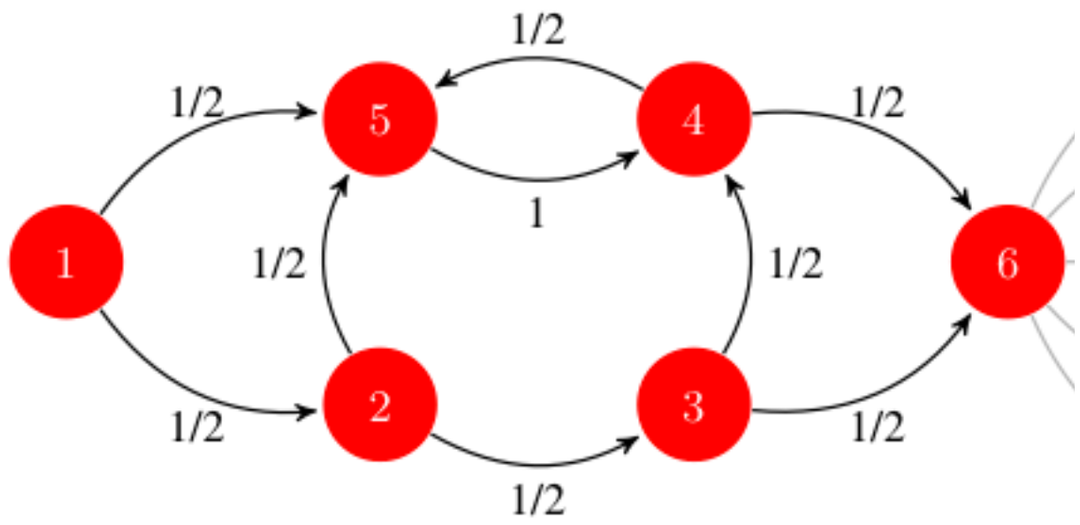
Scaled Pagerank

Agent update rule change as follows:

Scaled PageRank Update Rule: First apply the Basic PageRank Update Rule. Then scale down all PageRank values by a factor of s. This means that the total PageRank in the network has shrunk from 1 to s. We divide the residual $1 - s$ units of PageRank equally over all nodes, giving $(1 - s)/n$ to each.

Random Agent Approach

- Assign 10 *agents* to each node.
- In each iteration move the agents randomly based on the probabilities in the state diagram below.
- This gives the same result as the original page rank.



Agent with Strategies

- Agents might have different strategies.
- How does a web user behave?
  - Different demographic characteristics.
  - Different user experience.
  - etc.

Tricking The System

- A significant proportion of the web traffic come through web search engines.
- Higher ranks ~ Higher revenue for websites.
- Search Engine Optimization (SEO)
    - Websites actively optimize their sites to be listed in higher ranks.

References

- https://cse.google.com/cse/publicurl?cx=002720237717066476899:v2wv26idk7m : Google Dataset Search Engine

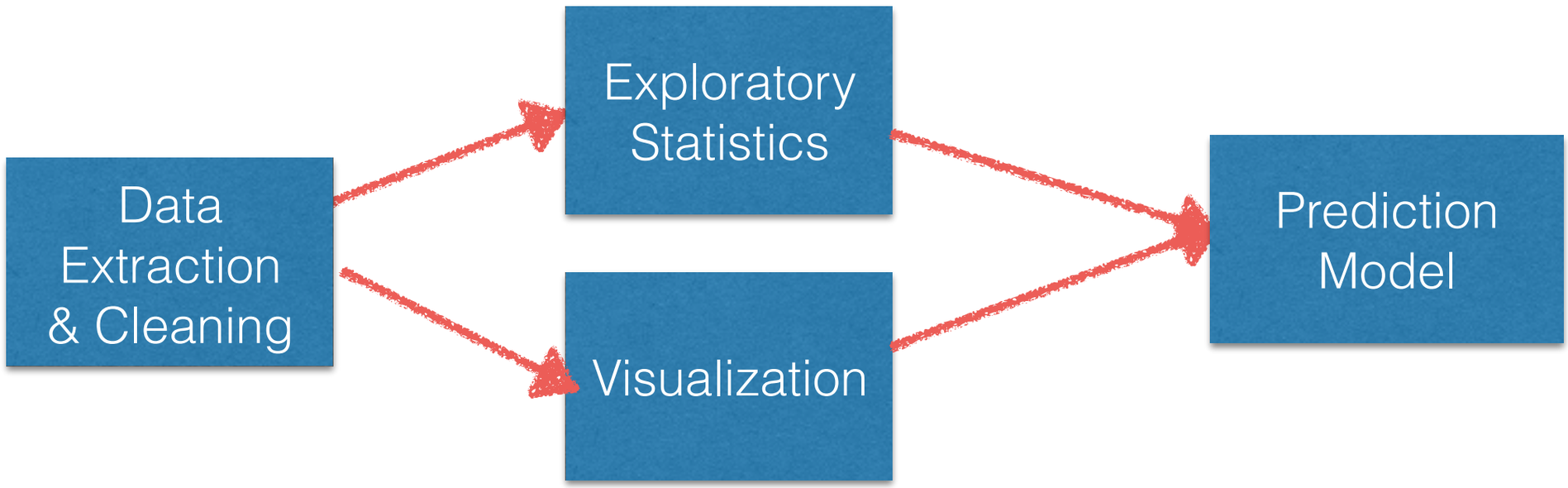
- http://visual.ly/exploration-google-pagerank-algorithm?view=true Page Rank Visualization

Automation of Data Analytics Tasks

Automation

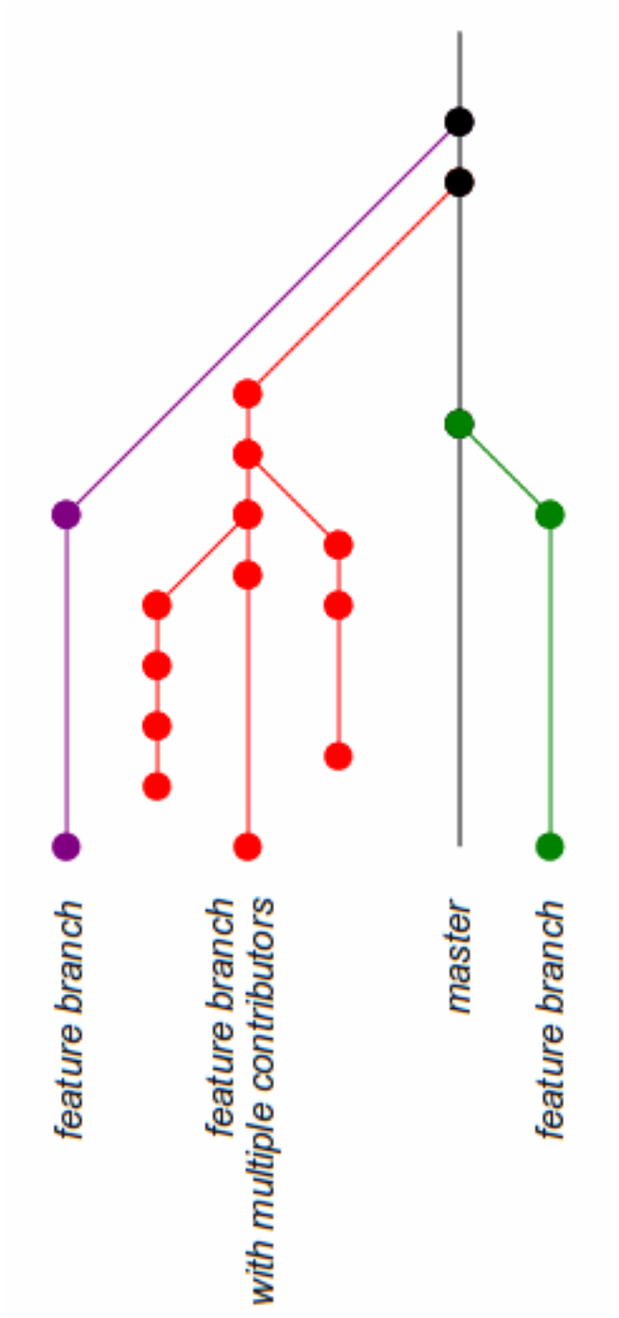
- Automation means one command/click execution of everything in your analysis.

- You should be able to run portions of your code independently. For example you do not need to re-extract the code everytime.



Version Control

- Use a version control system to store your code and metadata for every non-trivial project:
    - I recommend ***git*** if you do not have a favourite.
    - Dropbox **is not** version control.
- Store each of your results in unique folders with the model parameters.
- Data storage is more tricky. A few suggestions:
    - Create 3 folders (temp, output, raw)
    - Never change your raw data if possible. Tracking the changes might be hard.
    - Temp is your intermediate results.
    - For the output store at least the model parameters for each run and the output with timestamps.
- Poorly implemented code has some technical debt: (no test modules, no documentation, poor design.) You will pay for the debt in the future.



Review of the Final Topics

Final Topics

13+1 questions

25 point: R related basic questions - 5 questions
20 point: Regression - 2 questions
30 point: Classification - 3 questions
10 point: Experiment design - 1 question
15 point: Network analytics - 2 questions
2 point: Bonus question