Graphs in Machine Learning

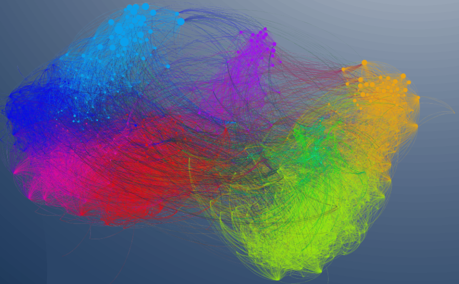
Google PageRank: Introduction

Random Surfer Model

Michal Valko

Inria & ENS Paris-Saclay, MVA

Partially based on material by: Andreas Krause, Branislav Kveton, Michael Kearns



Success story #2 Google PageRank

Objective: **Rank** all web pages (nodes on the graph) by how **many** other pages link to them and how **important** they are.

Success story #2 Google `PageRank`

Objective: **Rank** all web pages (nodes on the graph) by how **many** other pages link to them and how **important** they are.

basic `PageRank` is independent of query and the page content

Success story #2 Google `PageRank`

Objective: **Rank** all web pages (nodes on the graph) by how **many** other pages link to them and how **important** they are.

basic `PageRank` is independent of query and the page content

Internet

Success story #2 Google `PageRank`

Objective: **Rank** all web pages (nodes on the graph) by how **many** other pages link to them and how **important** they are.

basic `PageRank` is independent of query and the page content

Internet → graph

Success story #2 Google PageRank

Objective: **Rank** all web pages (nodes on the graph) by how **many** other pages link to them and how **important** they are.

basic PageRank is independent of query and the page content

Internet → graph → matrix

Success story #2 Google PageRank

Objective: **Rank** all web pages (nodes on the graph) by how **many** other pages link to them and how **important** they are.

basic PageRank is independent of query and the page content

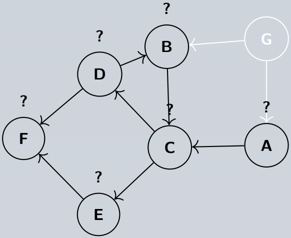
Internet → graph → matrix → stochastic matrix $\mathbf{M}$ $\left(\sum_j \mathbf{M}_{ij} = 1\right)$

Success story #2 Google `PageRank`

Objective: **Rank** all web pages (nodes on the graph) by how **many** other pages link to them and how **important** they are.

basic `PageRank` is independent of query and the page content

Internet $\rightarrow$ graph $\rightarrow$ matrix $\rightarrow$ stochastic matrix $\mathbf{M}$ $\left(\sum_j \mathbf{M}_{ij} = 1\right)$

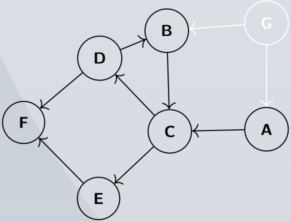


Success story #2 Google PageRank

Objective: **Rank** all web pages (nodes on the graph) by how **many** other pages link to them and how **important** they are.

basic PageRank is independent of query and the page content

Internet → graph → matrix → stochastic matrix $\mathbf{M}$ $\left(\sum_j \mathbf{M}_{ij} = 1\right)$



Success story #2 Googl Rank

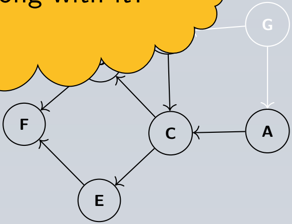
Objective: **Rank** all wel how
many other pages li re.

basic PageRank is inc tent

Internet → graph → matrix 1
$\left(\sum_j \mathbf{M}_{ij} = 1\right)$

Random Surfer Process

What is wrong with it?



Success story #2 Googl Rank

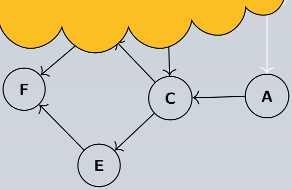
Objective: **Rank** all wel how
many other pages li re.

Random Surfer Process

basic PageRank is in tent

Internet → graph → matrix
$\left(\sum_j \mathbf{M}_{ij} = 1\right)$

What is wrong with it?



dangling pages act like sinks

Success story #2 Google `PageRank`

`http://infolab.stanford.edu/~backrub/google.html`:

PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page.

Success story #2 Google PageRank

`http://infolab.stanford.edu/~backrub/google.html`:

PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page.

- page is **important** if **important** pages link **to** it

Success story #2 Google PageRank

http://infolab.stanford.edu/~backrub/google.html:

PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page.

- page is **important** if **important** pages link **to** it
  - circular definition

Success story #2 Google PageRank

`http://infolab.stanford.edu/~backrub/google.html`:

PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page.

- page is **important** if **important** pages link **to** it
  - circular definition
- importance of a page is distributed **evenly**

Success story #2 Google `PageRank`

`http://infolab.stanford.edu/~backrub/google.html`:

PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page.

- page is **important** if **important** pages link **to** it
  - circular definition
- importance of a page is distributed **evenly**
- probability of being bored is 15%

Michal Valko

michal.valko@inria.fr

Inria & ENS Paris-Saclay, MVA

https://misovalko.github.io/mva-ml-graphs.html