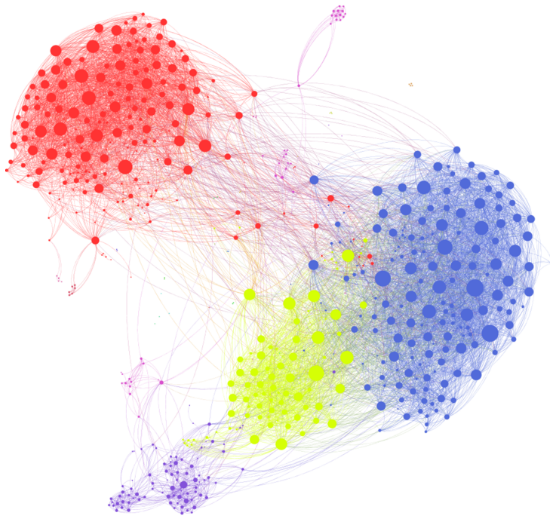


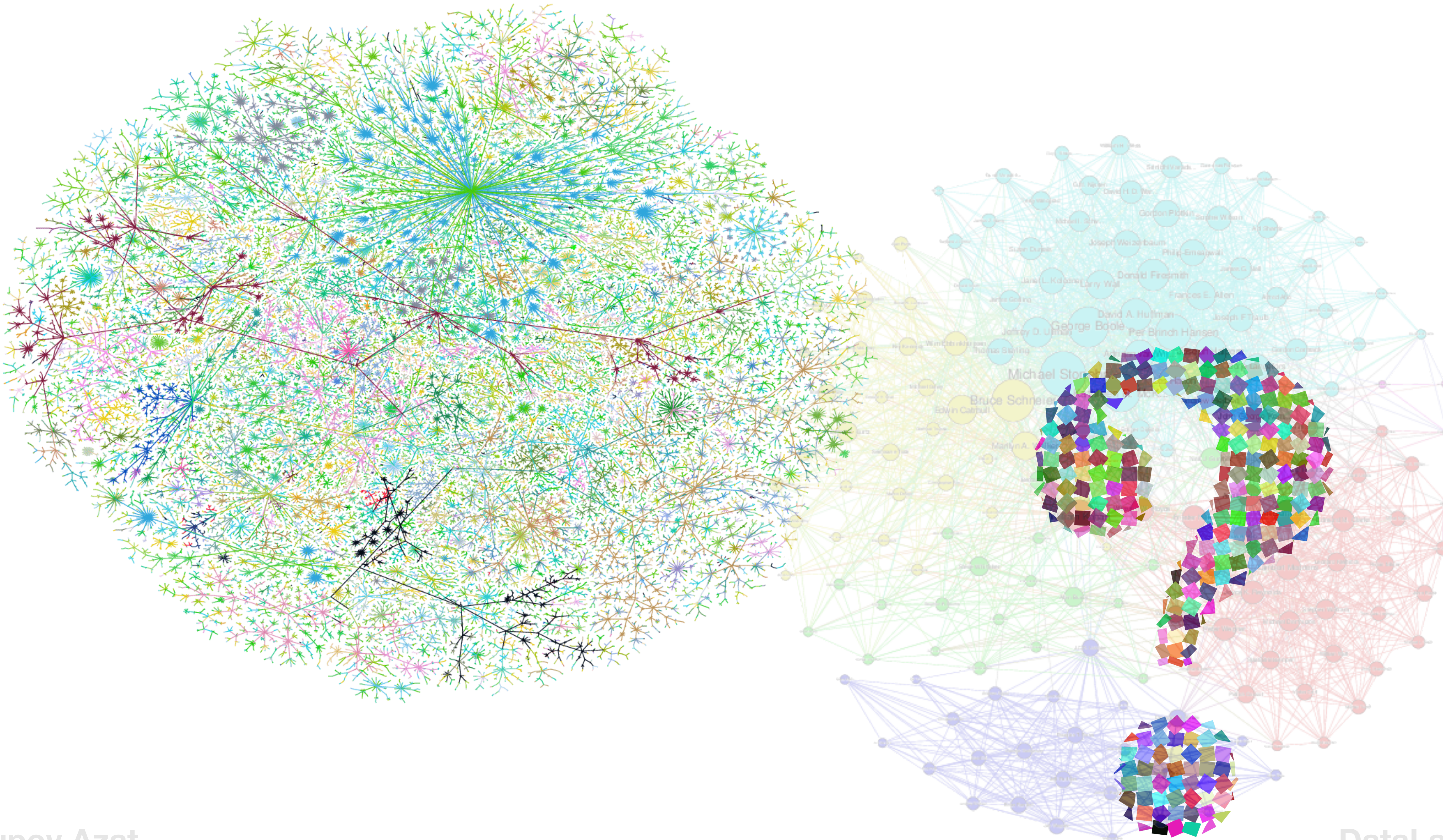
Data Mining

Lecturer: Якупов Азат Шавкатович

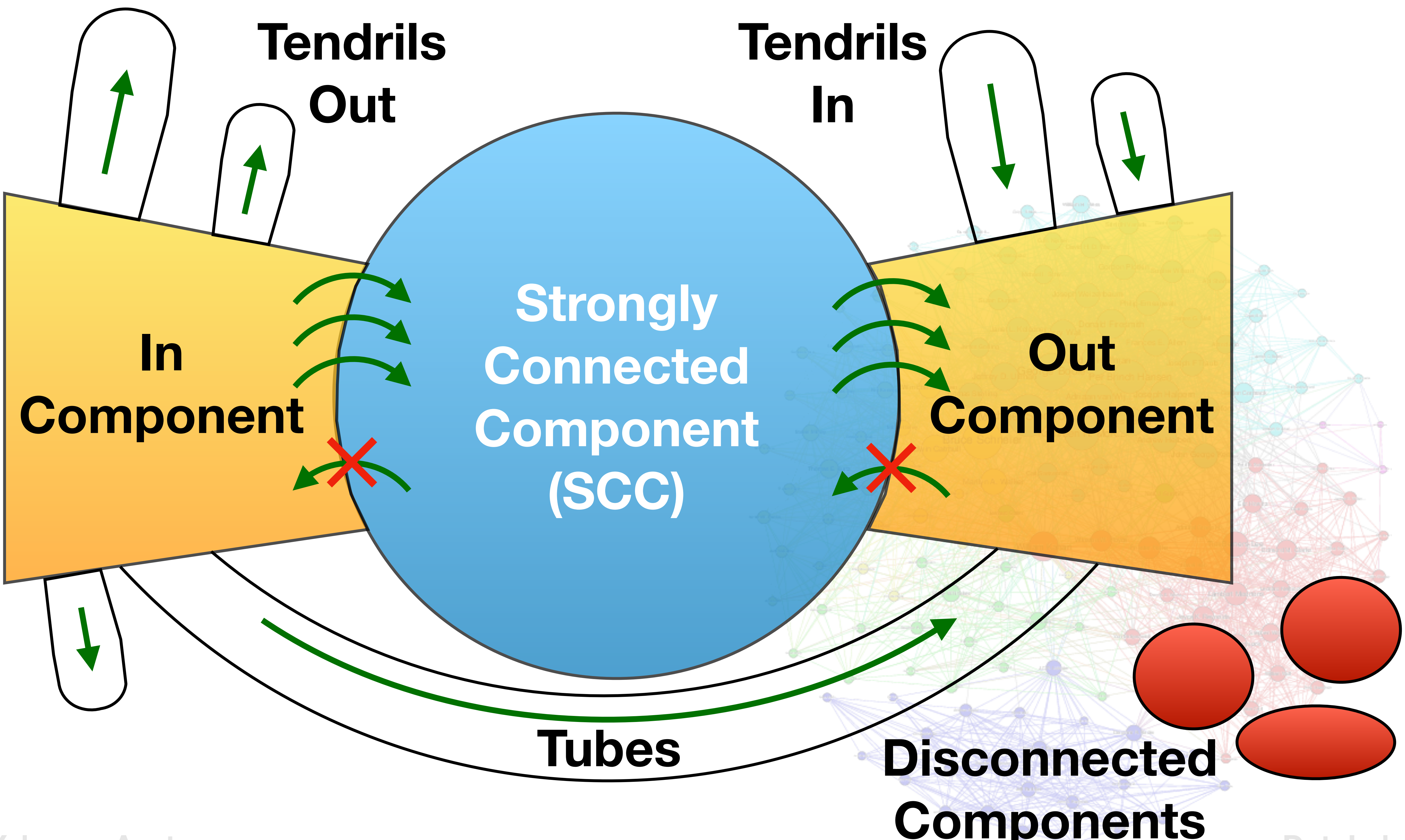
https://ru.linkedin.com/in/ayakupov

https://datalaboratory.one

Introduction



Structure of the Web



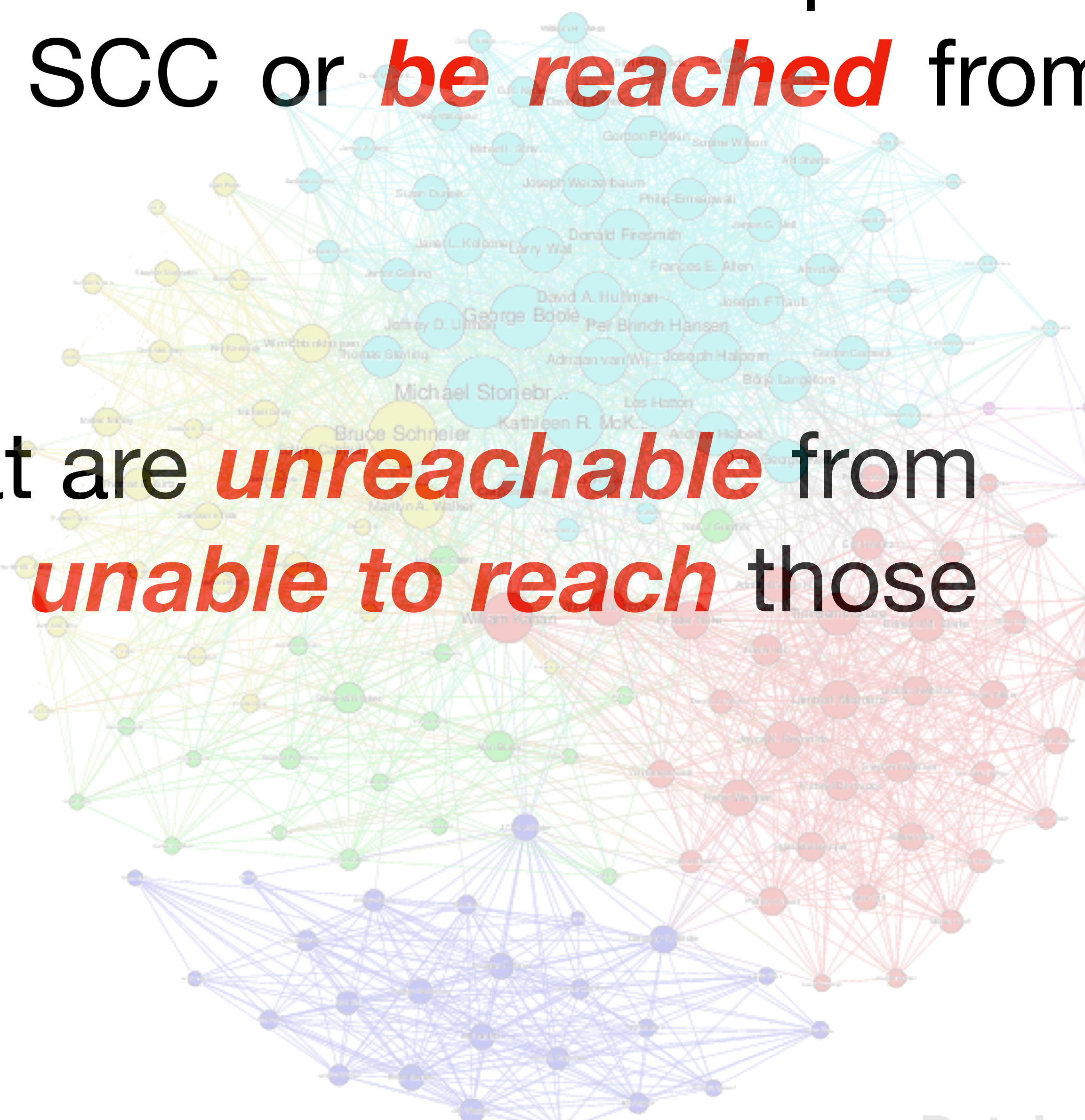
Structure of the Web

- The ***in-component***, consisting of pages that could reach the **SCC** by following links, but were ***not reachable*** from the **SCC**
- The ***out-component***, consisting of pages ***reachable*** from the SCC but unable to reach the SCC
- ***Tendrils***. Some tendrils consist of pages ***reachable*** from the in-component but ***not able to reach*** the in-component. The other tendrils ***can reach*** the out-component, but are ***not reachable*** from the out-component

Structure of the Web

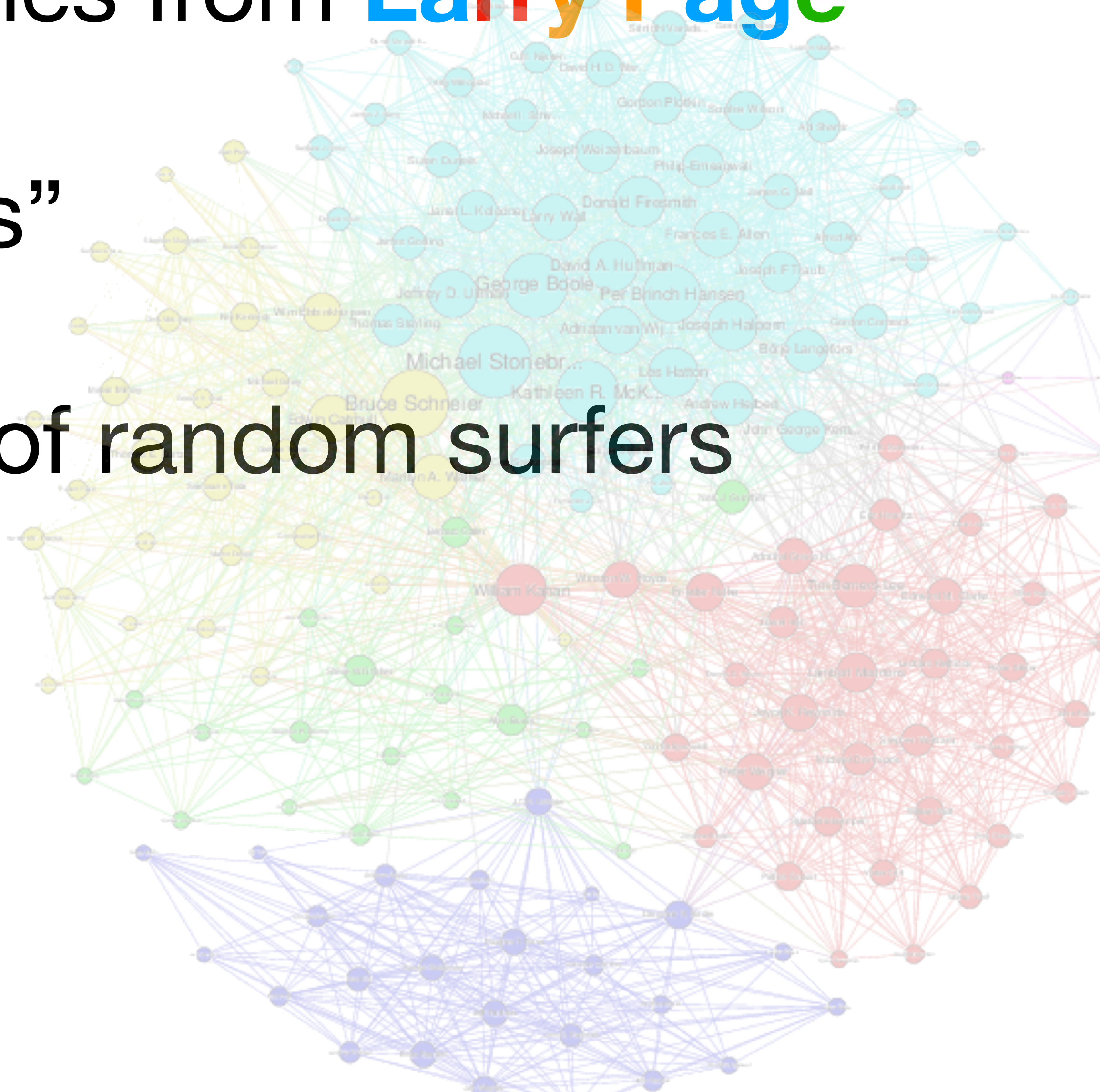
- ***Tubes***, which are pages ***reachable*** from the in-component and ***able to reach*** the out-component, but ***unable to reach*** the SCC or ***be reached*** from the SCC

- ***Isolated components*** that are ***unreachable*** from the large components and ***unable to reach*** those components

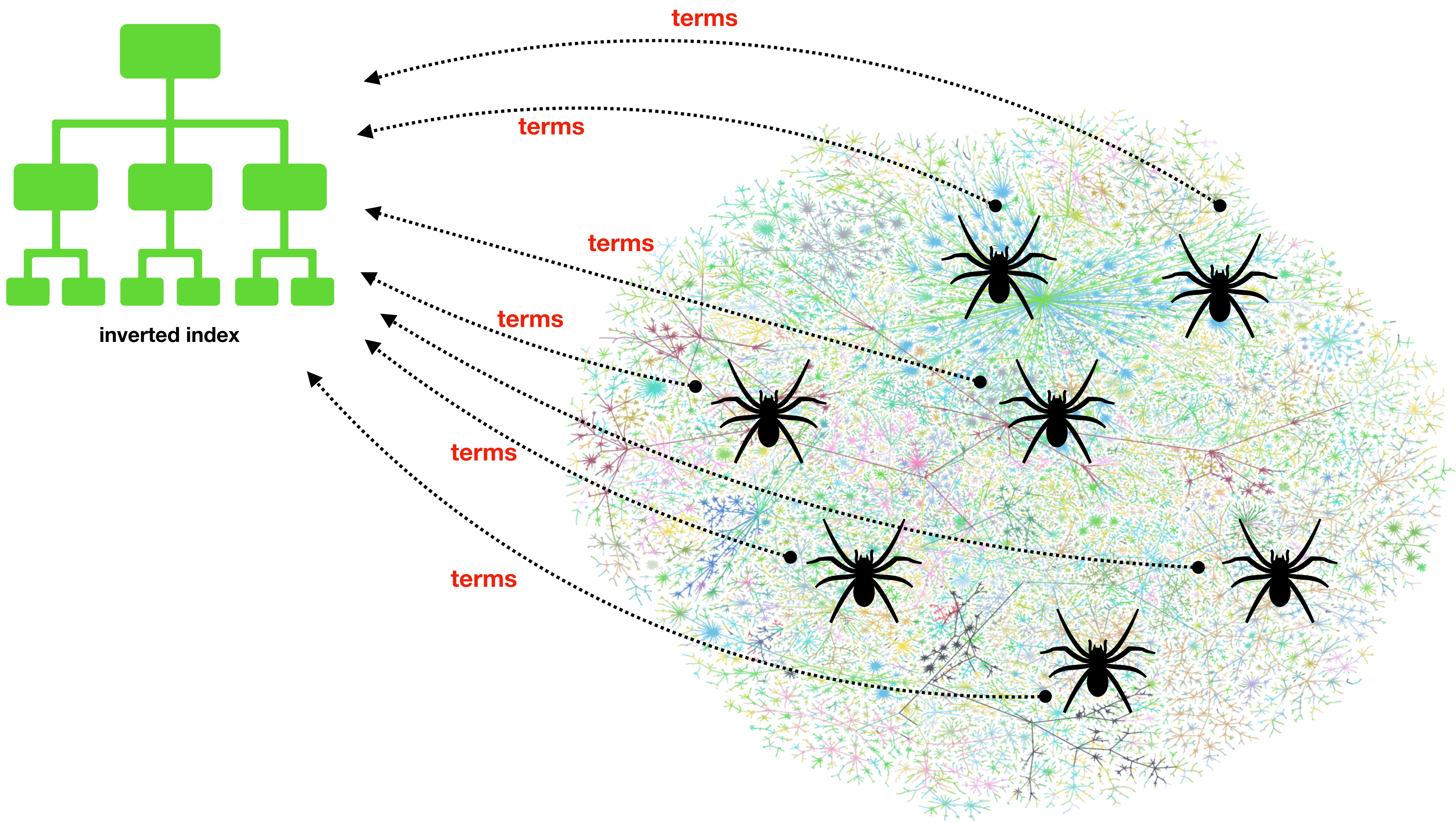


Page Rank Algorithm

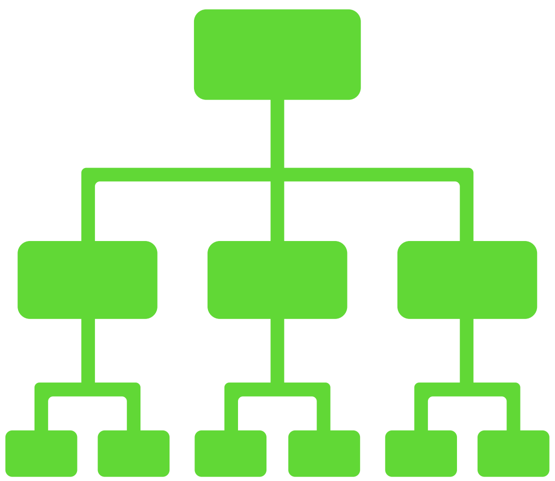
- The term PageRank comes from Larry Page
- Idea of “Random Surfers”
- Technique of “taxation” of random surfers



Early Search Engines



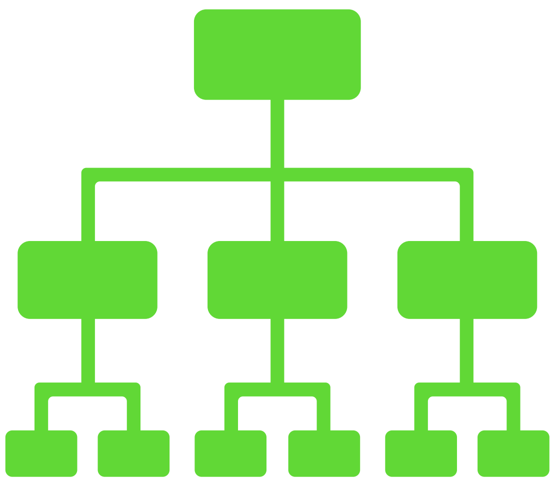
Early Search Engines



inverted index

page_id	term
www.mail.ru	mail
www.mail.ru	news
www.mail.ru	currencies
https://datalaboratory.one	data structures
https://datalaboratory.one	data mining
…	…
https://thedatalab.com	data structures

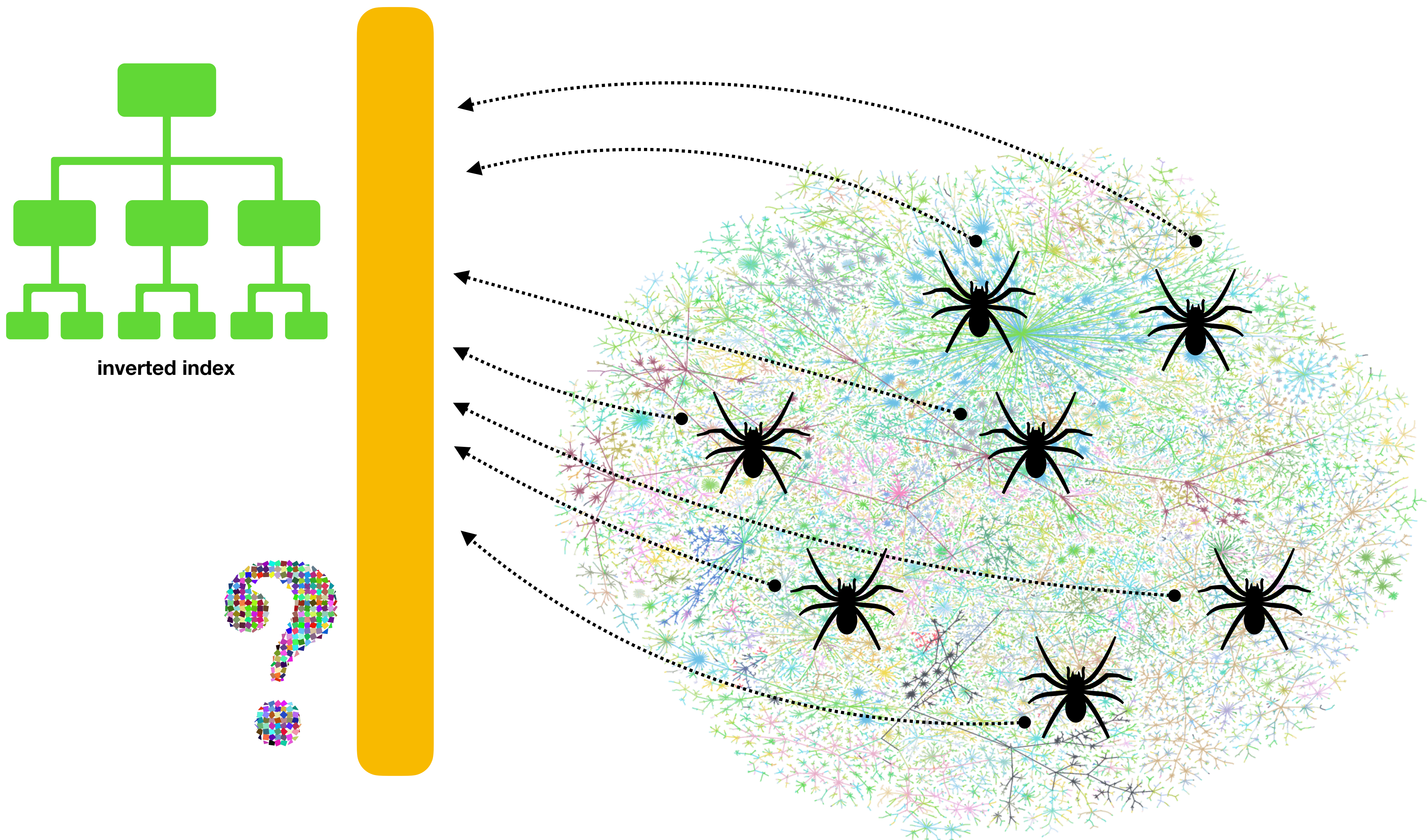
Early Search Engines



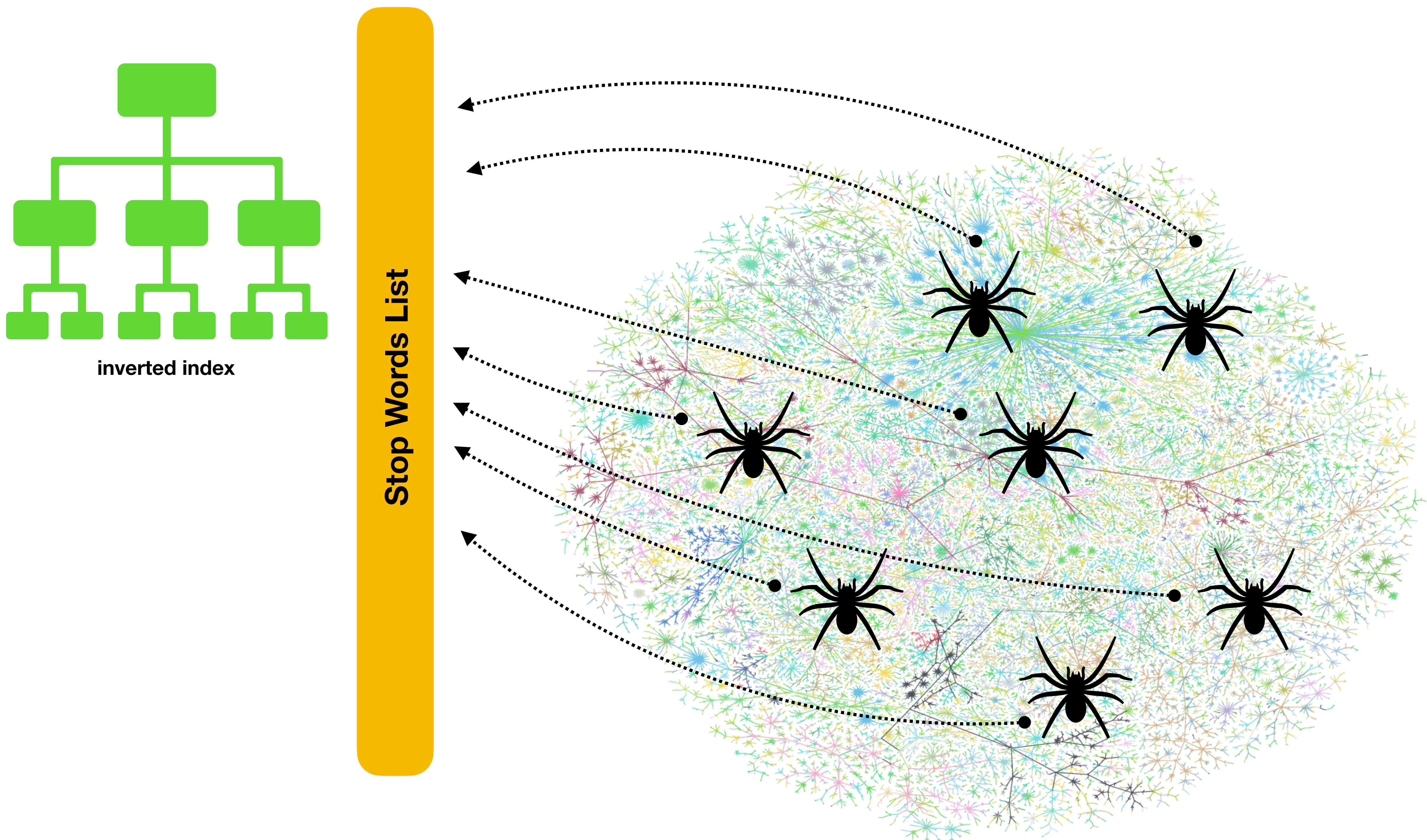
inverted index

term	document
mail	1
news	1
currencies	1
data structures	2 10
data mining	2
...	...
...	...

Early Search Engines



Early Search Engines



Early Search Engines

How to “hack” to make a SEO

Early Search Engines

The Data Laboratory

Hi there!

We are highly "SCI-IT-motivated" students from Kazan Fe

We are here to understand a real world by different aspec

> "Data is the new oil"
>
> *Clive Humby*

BECOMING A CLINICAL LABORATORY PROFESSIONAL

What is a medical laboratory science professional?

Medical laboratory science professionals, often called medical laboratorians, are vital healthcare detectives, uncovering and providing laboratory information from laboratory analyses that assist physicians in patient diagnosis and treatment, as well as in disease monitoring or prevention (maintenance of health). We use sophisticated biomedical instrumentation and technology, computers, and methods requiring manual dexterity to perform laboratory testing on blood and body fluids. Laboratory testing encompasses such disciplines as clinical chemistry, hematology, immunology, immunohematology, microbiology, and molecular biology. Medical laboratory science professionals generate accurate laboratory data that are needed to aid in detecting cancer, heart attacks, diabetes, infectious mononucleosis, and identification of bacteria or viruses that cause infections, as well as in detecting drugs of abuse. In addition, we monitor testing quality and consult with other members of the healthcare team.

FINANCIAL TIMES

HOME WORLD US COMPANIES TECH MARKETS GRAPHICS OPINION WORK & CAREERS LIFE & ARTS HOW TO SPEND IT

MARKETS › COMMODITIES › OIL

Get a fresh start.

Oil + Add to myFT

MARCH 24 2020

Chevron Corp

Chevron announces spending cuts and halts buyback programme

US oil group says capex will fall by $4bn, with Permian shale operations hardest hit

Early Search Engines

The Data Laboratory

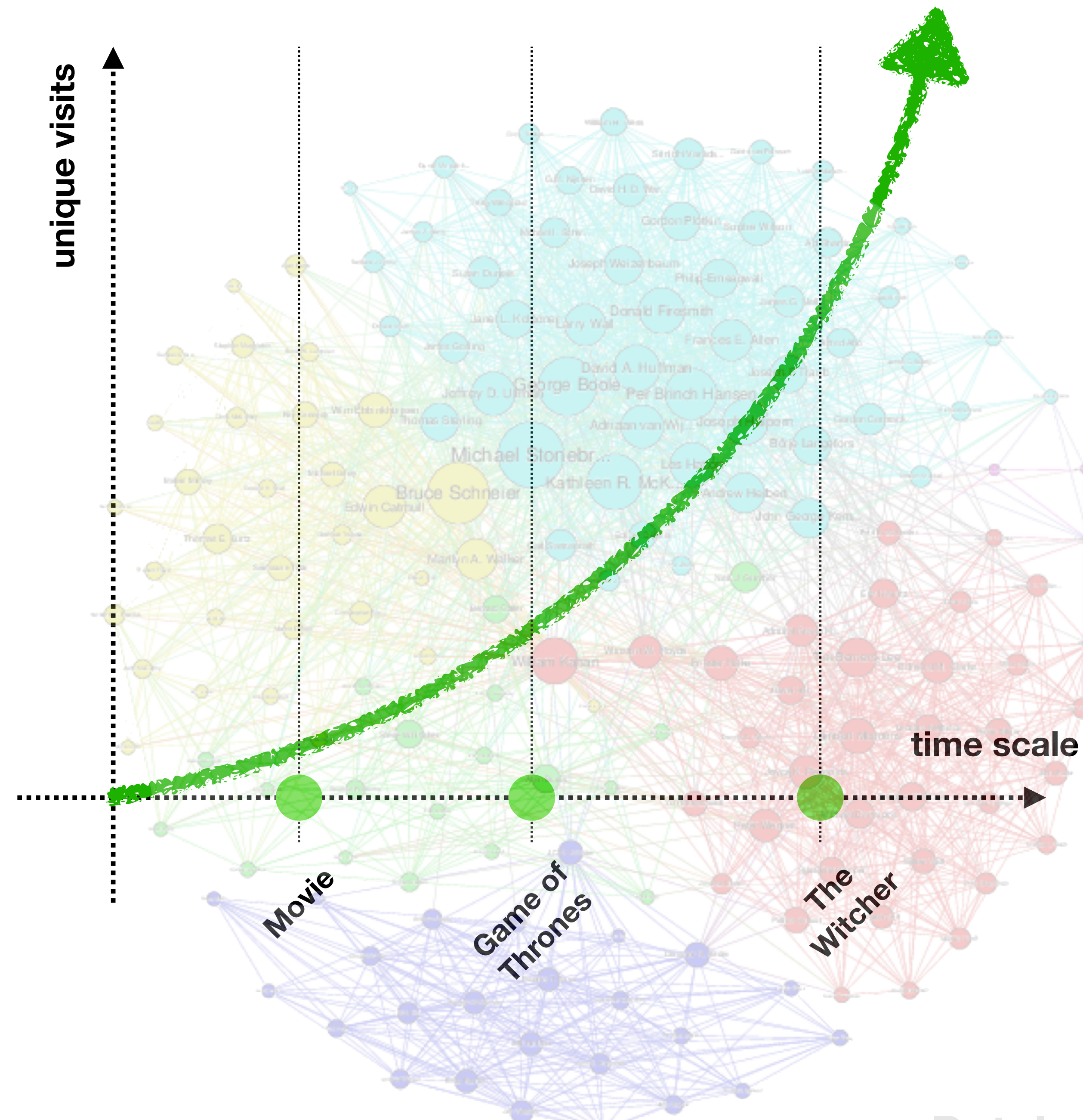
Hi there!

We are highly “SCI-IT-motivated” students from Kazan Fe

We are here to understand a real world by different aspec

> **“Data is the new oil”**
>
> *Clive Humby*

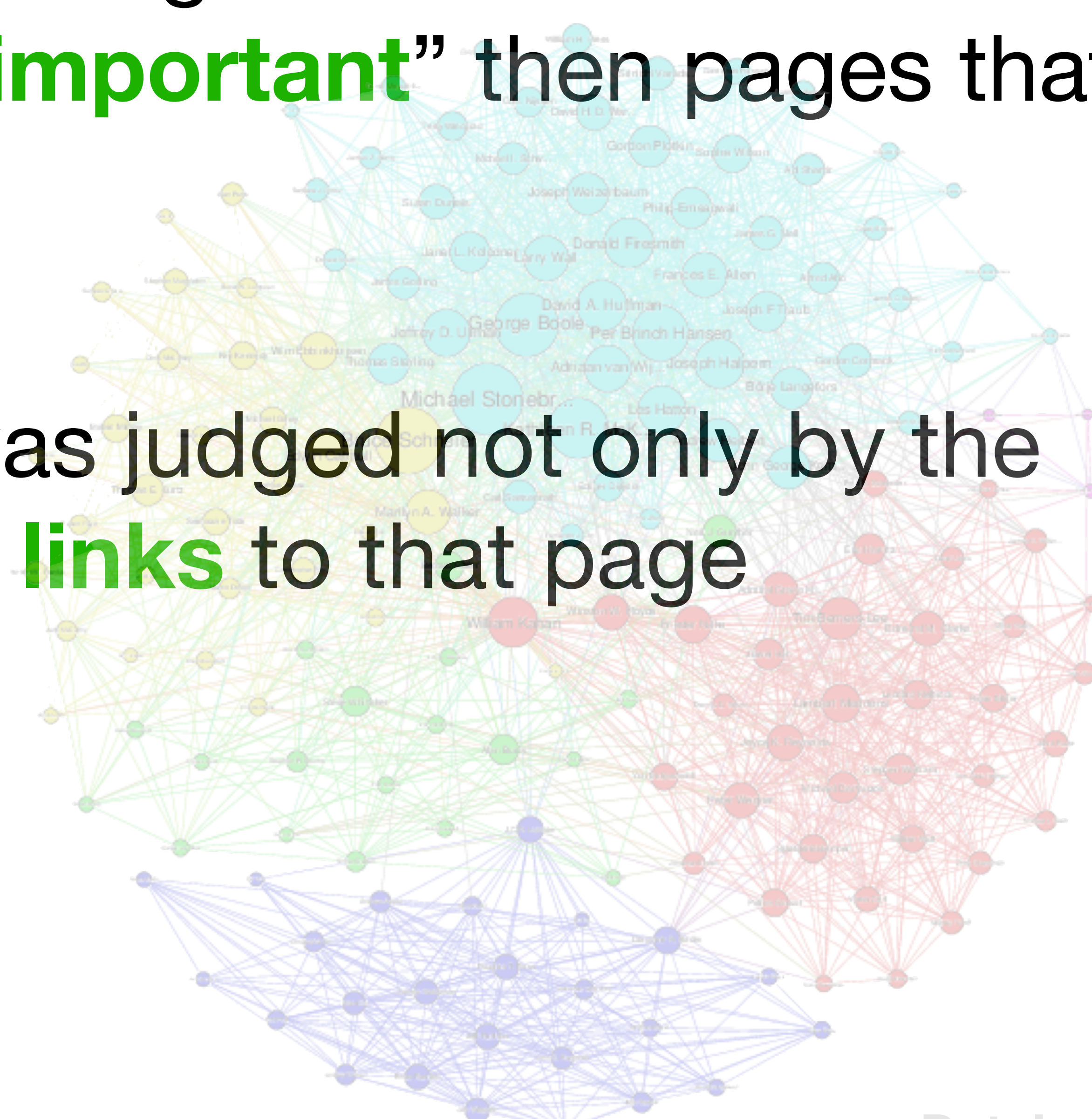
Term SPAM



Google innovations

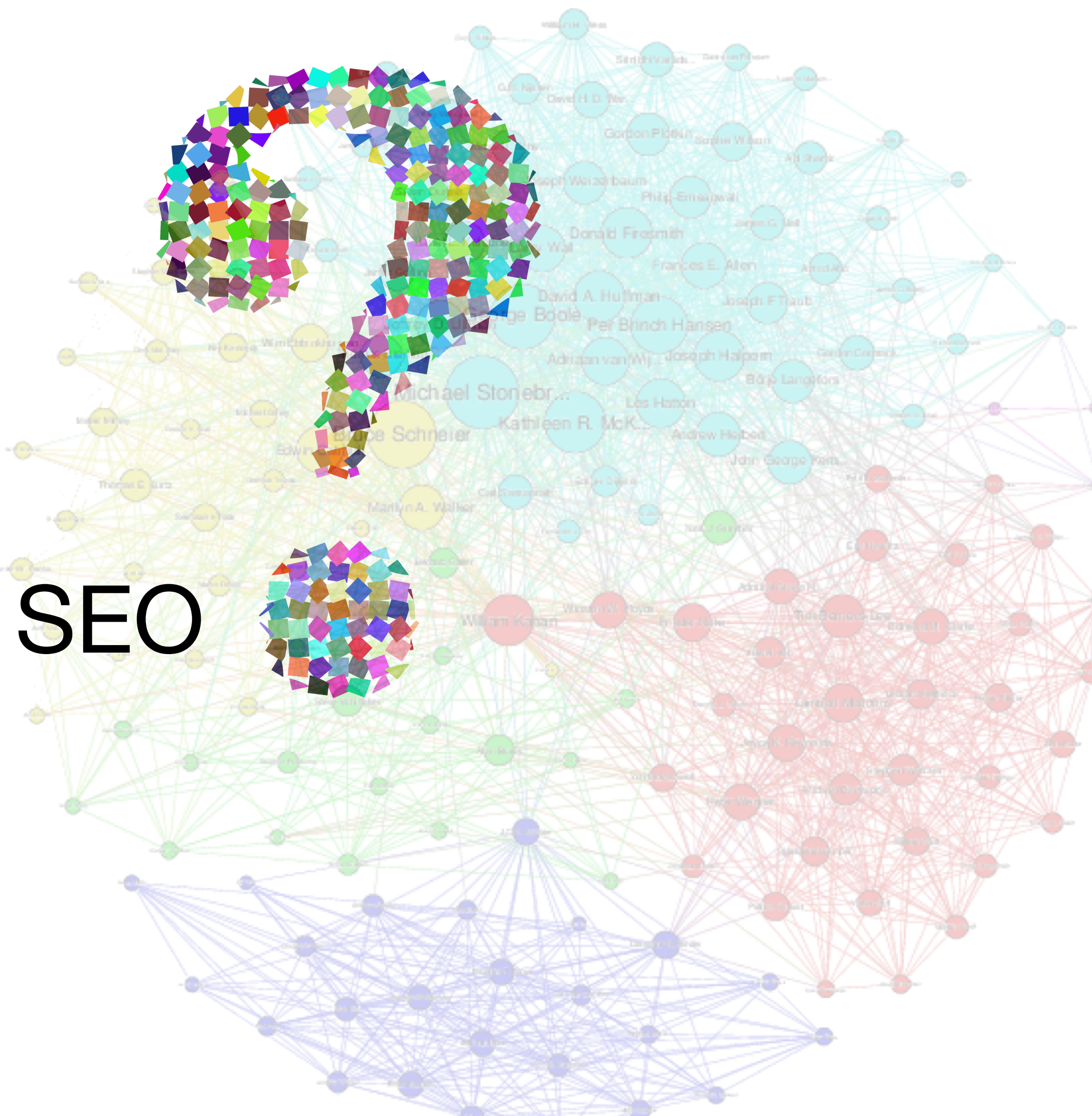
- Web pages would have a large number of **surfers** were considered more "**important**" then pages that would rarely be visited

- The content of a page was judged not only by the terms, but by **in/out the links** to that page

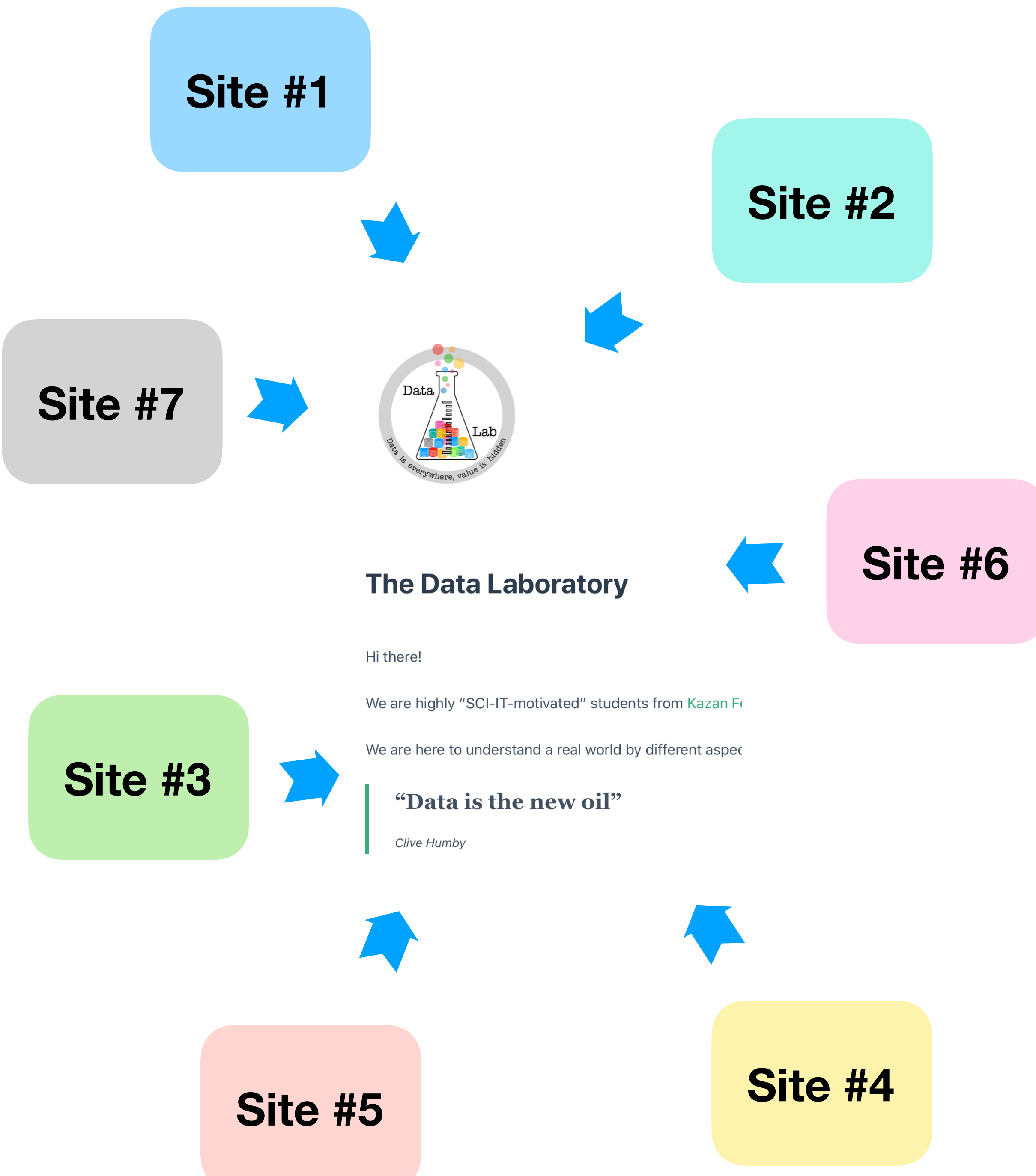


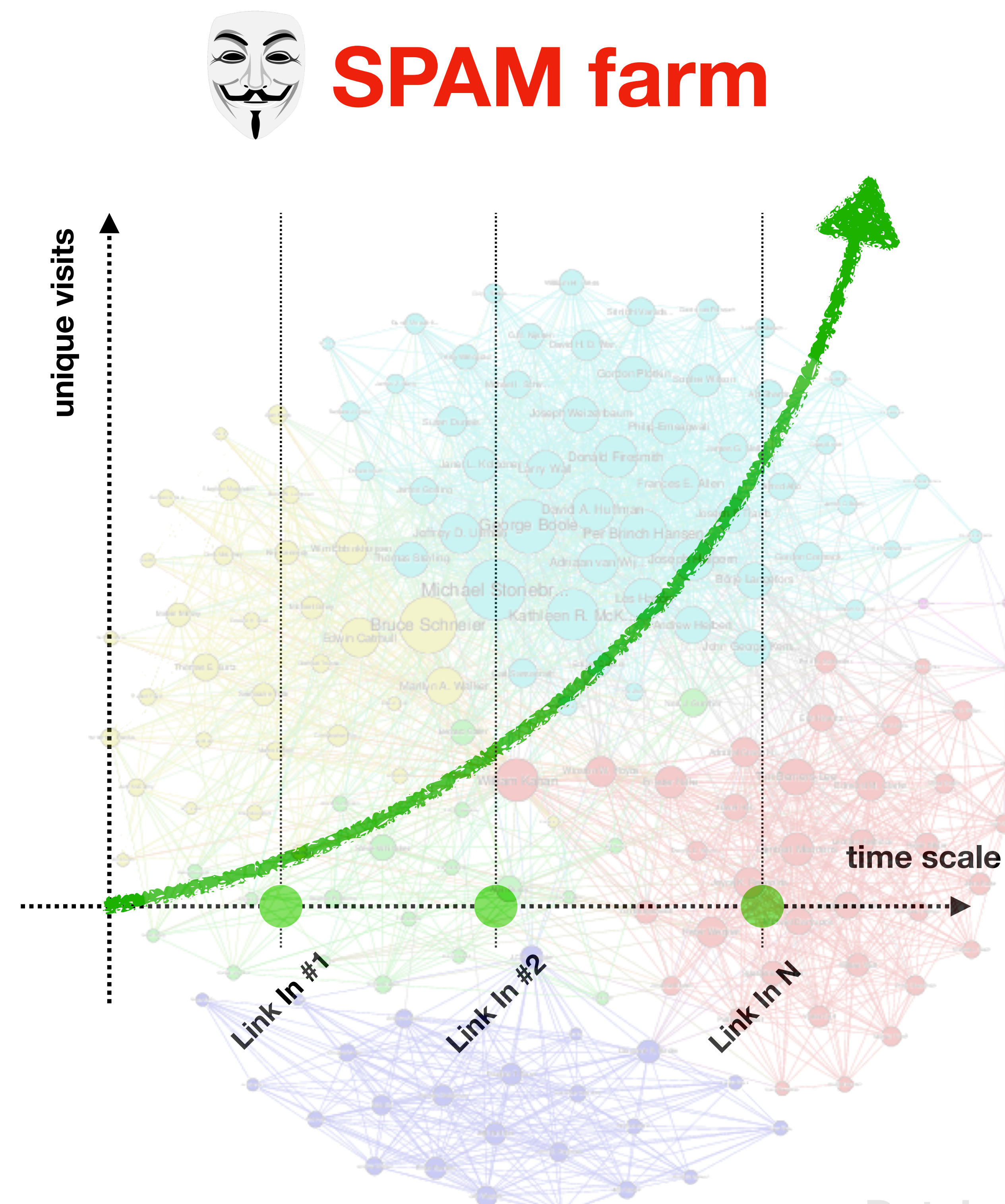
Google innovations

How to “hack” to make a SEO



Google innovations

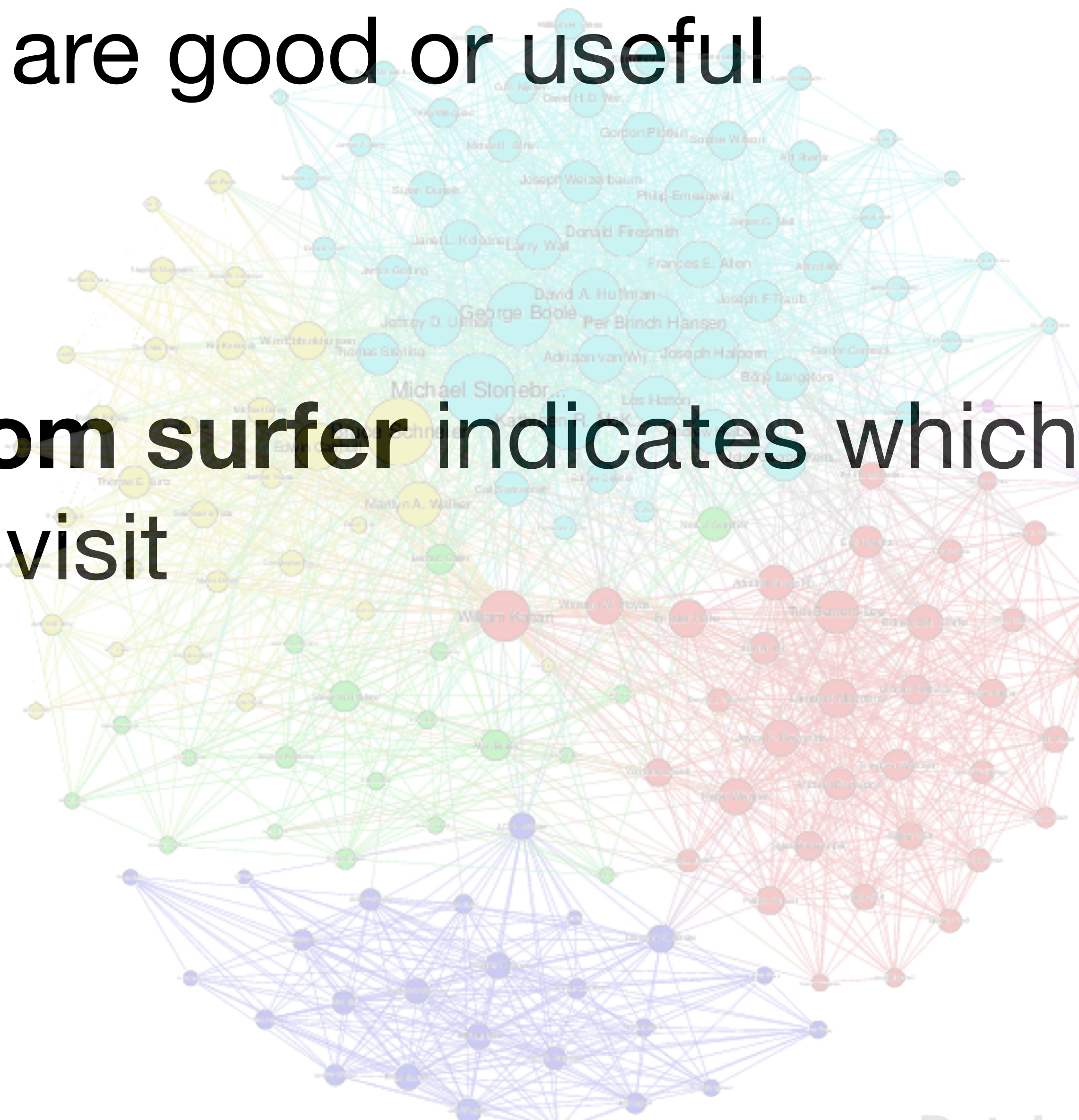




Motivations

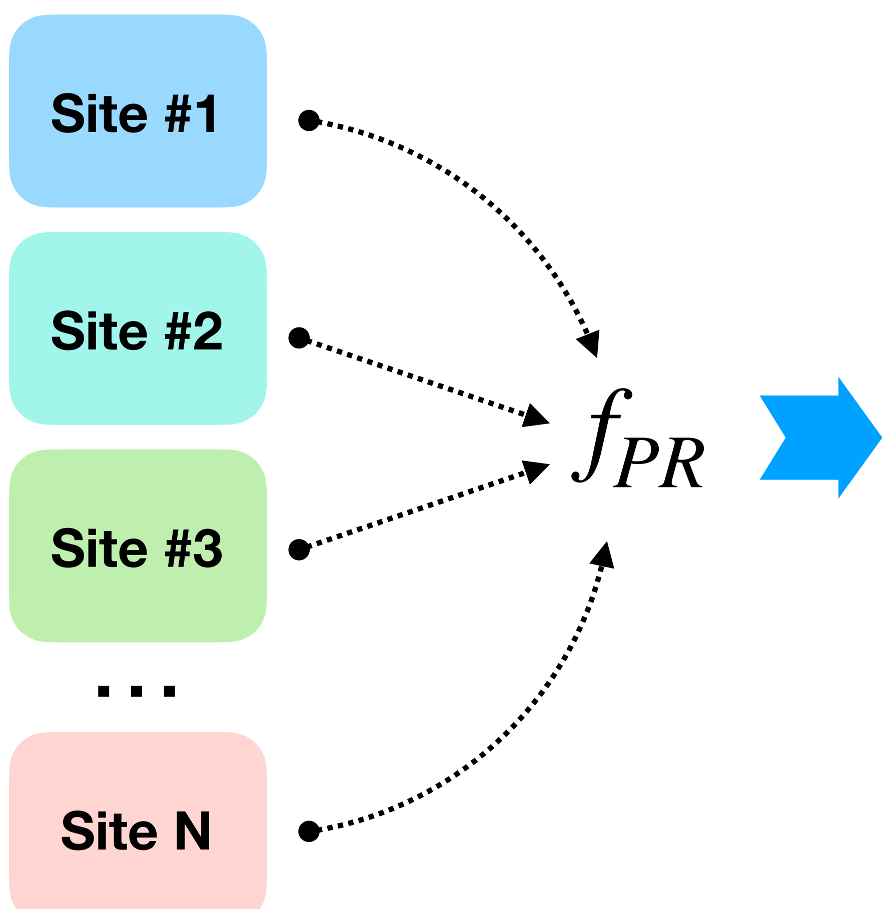
- Users of the Web “vote with their feet”. They share links to pages they think are good or useful

- The behaviour of a **random surfer** indicates which pages users are likely to visit



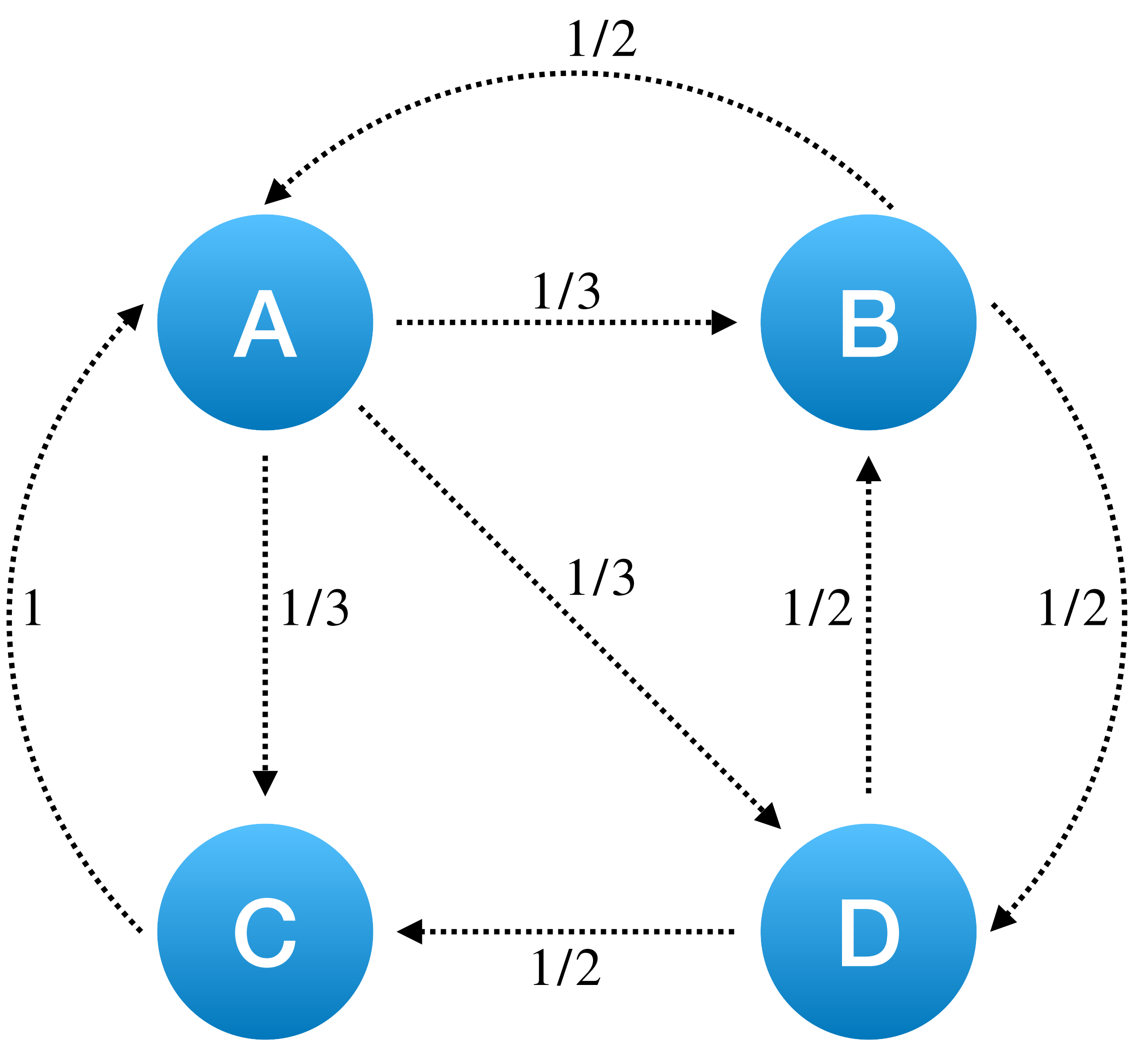
Definition of PageRank

- PageRank is the function f_{PR} which assigns a real number to each page in the Web



Web Page	f_{PR}
Site #1	0.001
Site #2	0.00045
Site #3	0.0000045
...	...
Site N	0.00000000…9

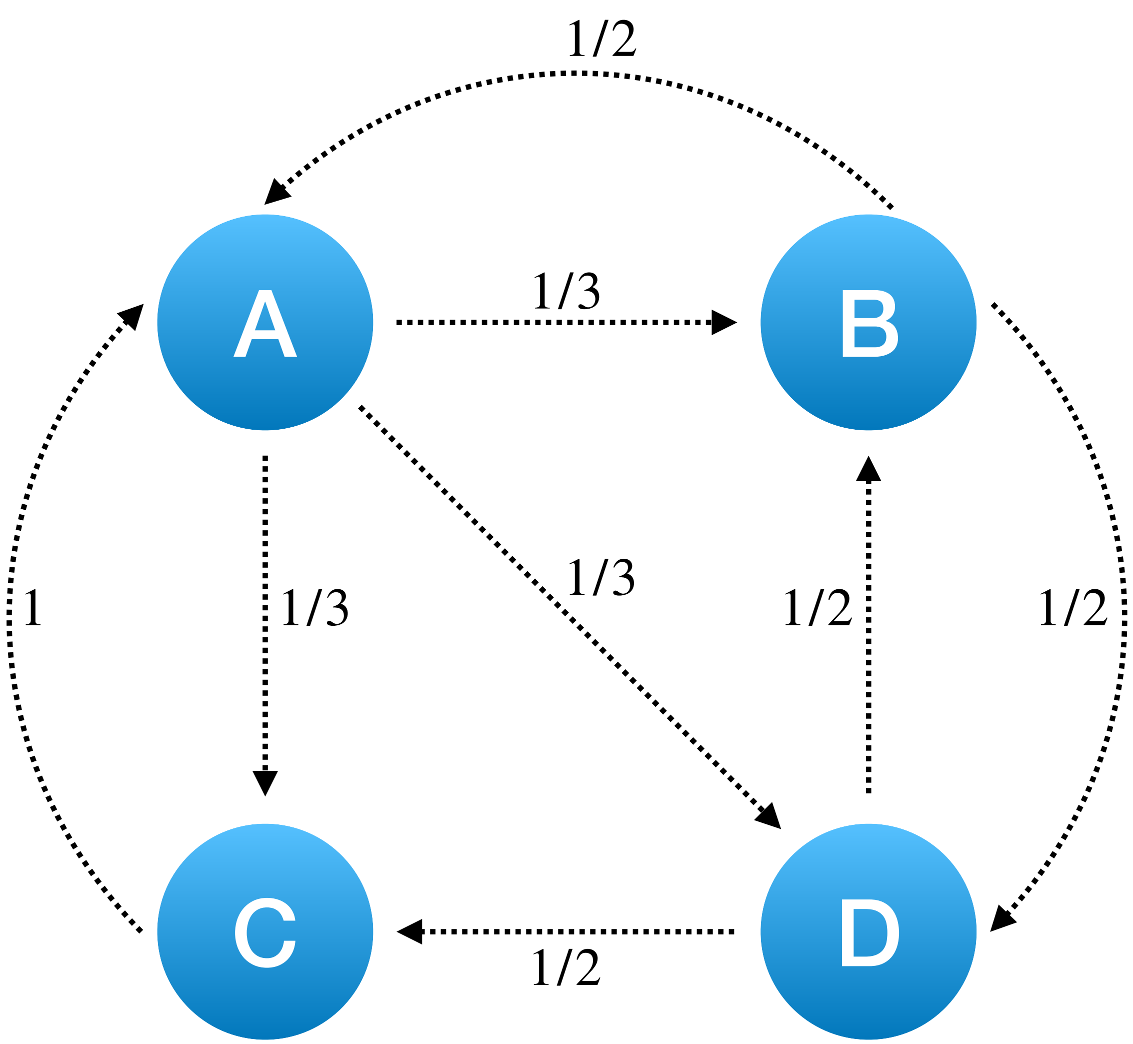
PageRank Sample



transition matrix

$$M = \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \end{array} \begin{array}{c} \\ A \\ B \\ C \\ D \end{array}$$

PageRank Sample

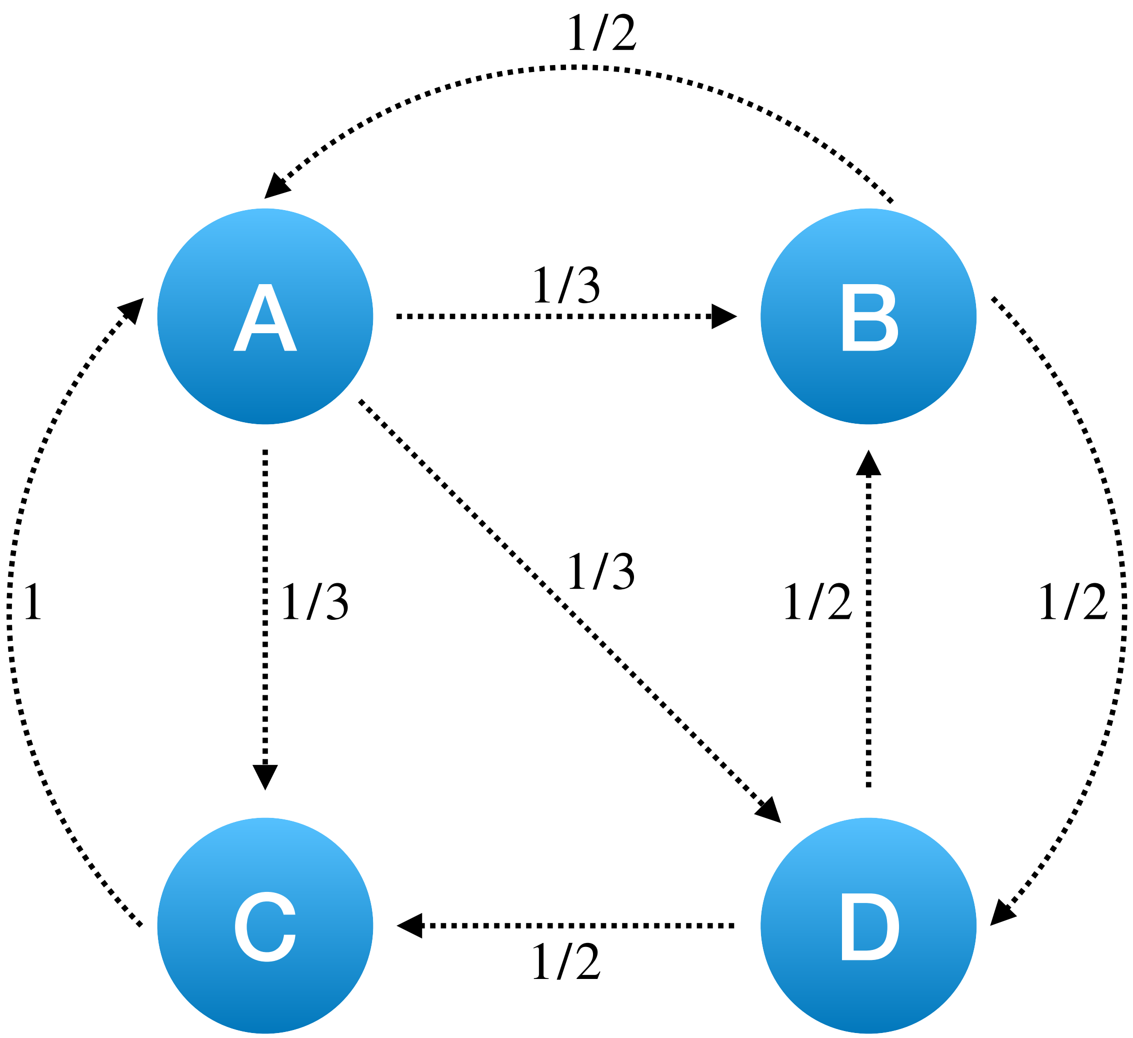


transition matrix

$$M = \begin{array}{cc} \begin{array}{cccc} A & B & C & D \end{array} & \\ \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} & \begin{array}{c} A \\ B \\ C \\ D \end{array} \end{array}$$

$$v_0 = \begin{array}{cc} \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} & \begin{array}{c} A \\ B \\ C \\ D \end{array} \end{array}$$

PageRank



v_0 → $M \cdot v_0$

1st surf

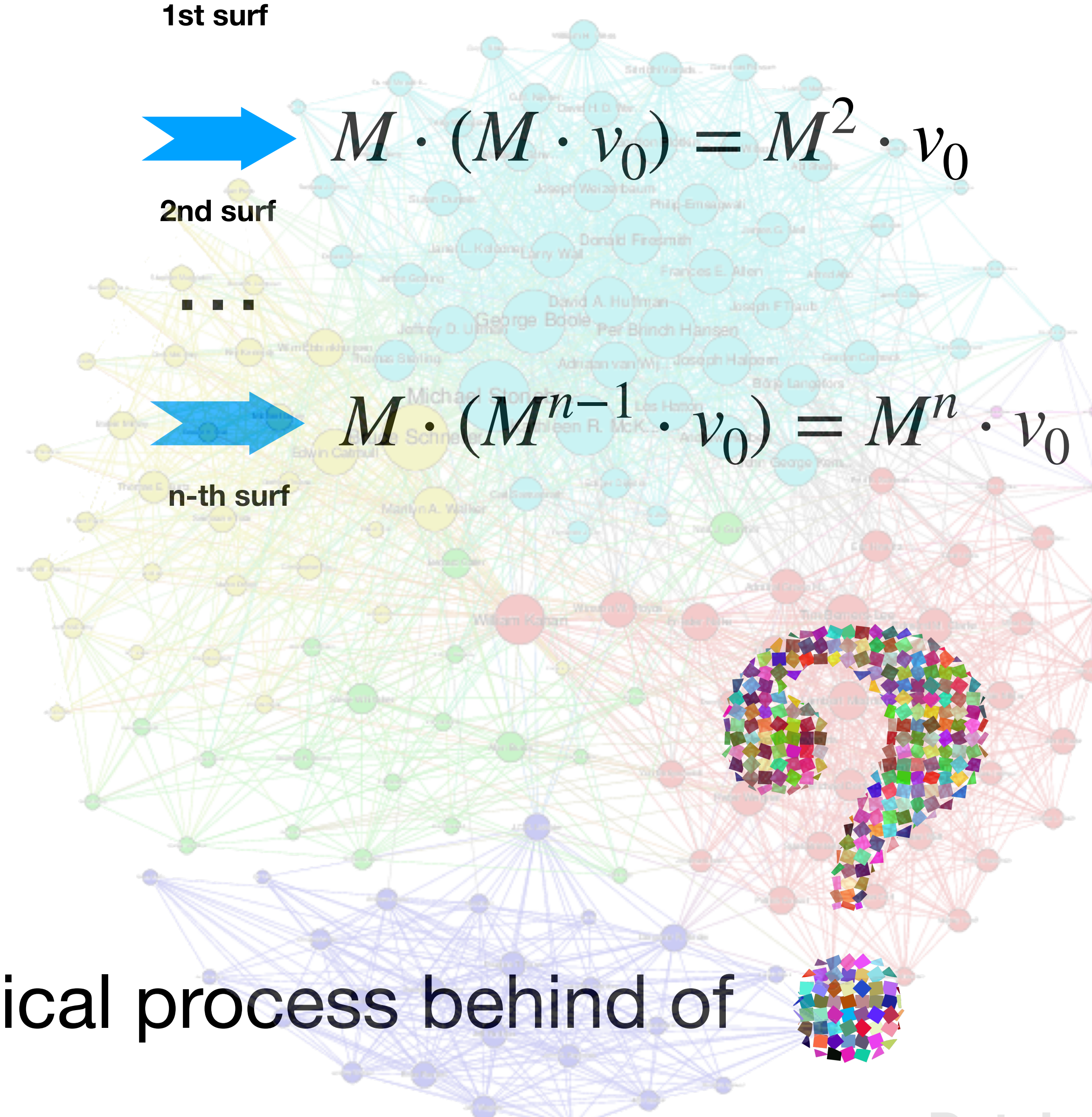
→ $M \cdot (M \cdot v_0) = M^2 \cdot v_0$

2nd surf

...

→ $M \cdot (M^{n-1} \cdot v_0) = M^n \cdot v_0$

n-th surf



What is a mathematical process behind of ?

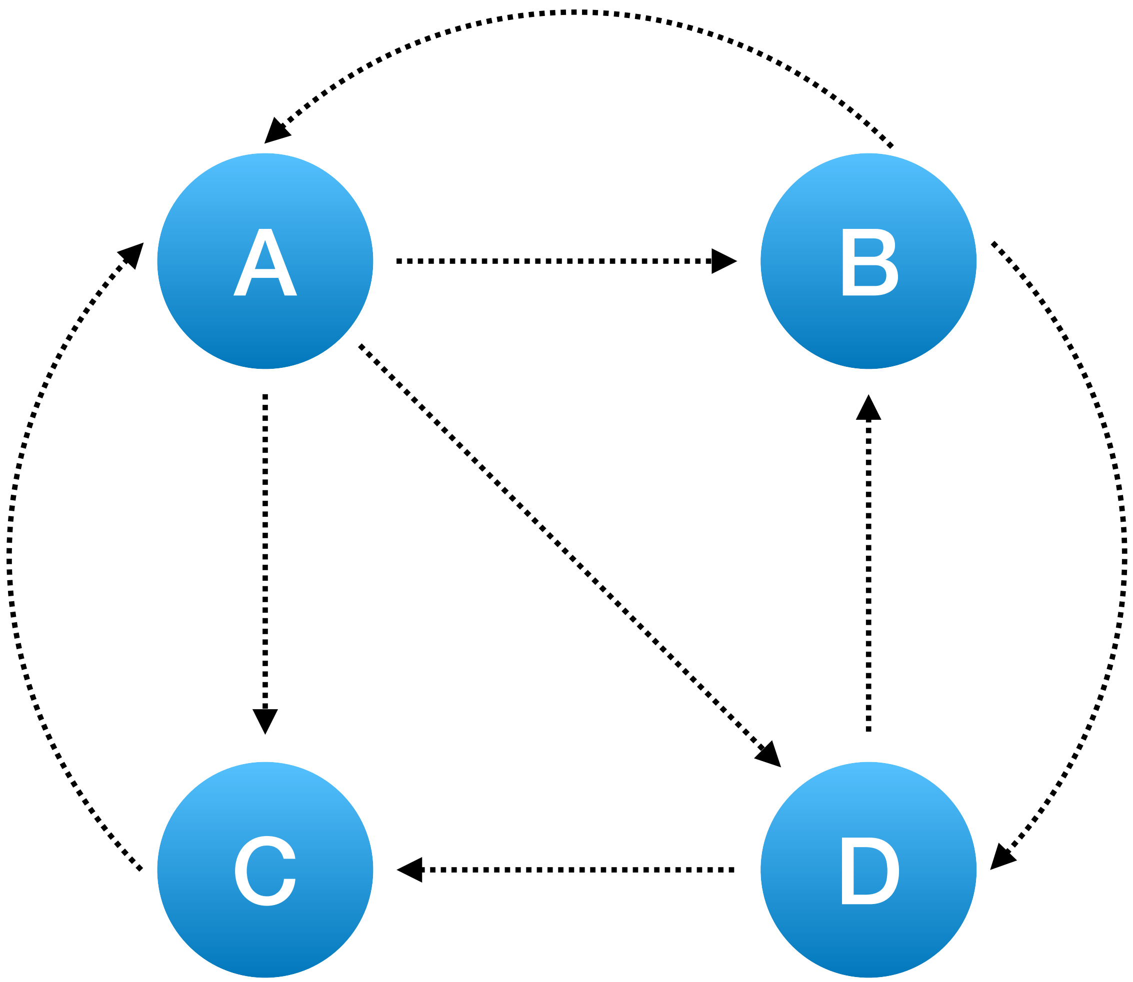
PageRank Sample

Russian mathematician **Andrey Markov**

> “Markov process is a process for which one can make predictions for its future based solely on its present state just as well as one could knowing the process's full history. In other words, conditional on the present state of the system, its future and past states are independent.”

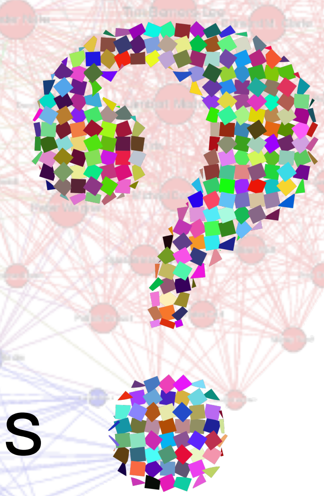
Wikipedia

Main conditions

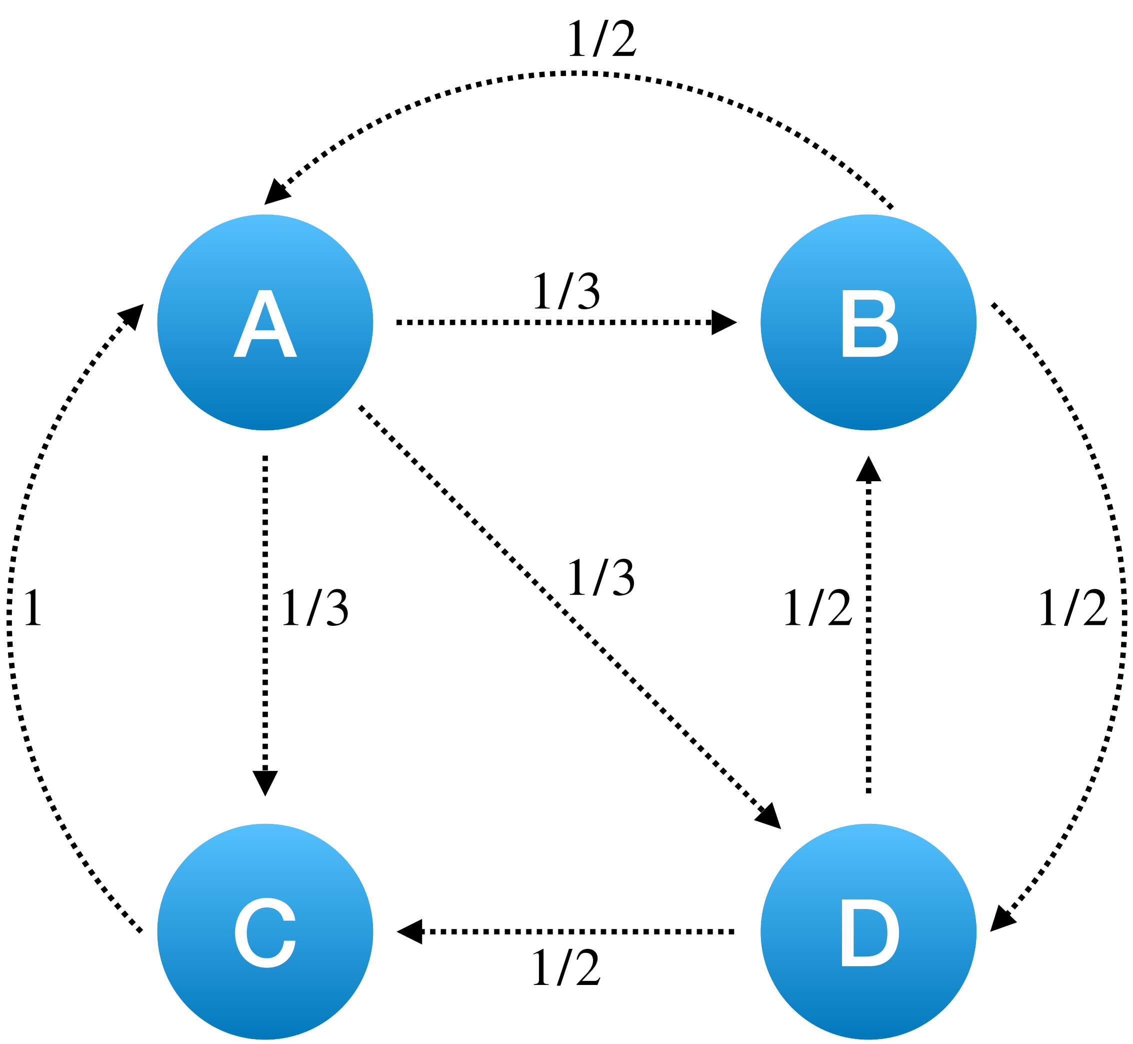


- The graph is **strongly connected**; that is, it's possible to get from any node to any other node
- There are no **dead ends**. nodes that have no link out

Does our graph satisfy both conditions

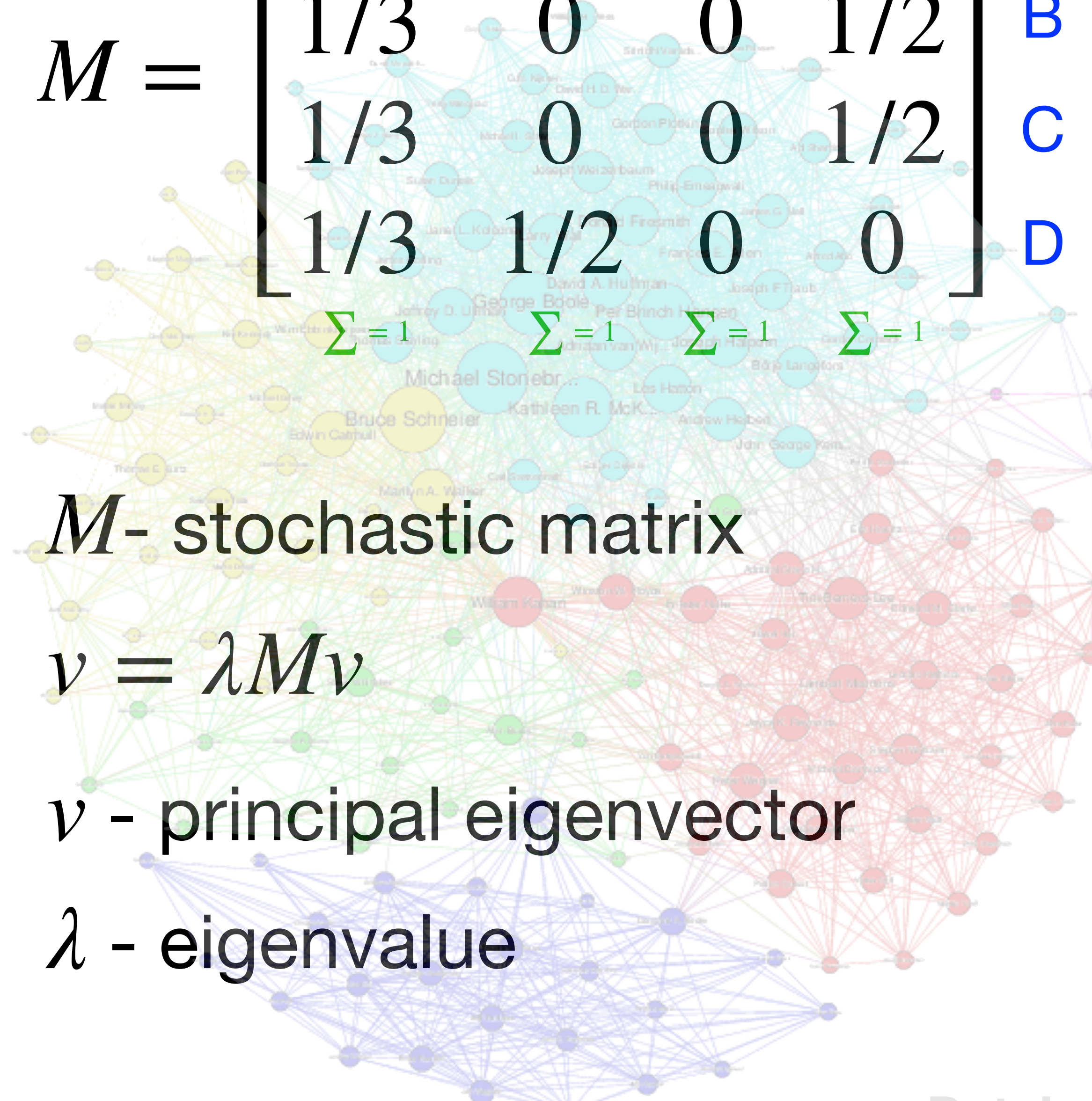


PageRank



transition matrix

$$M = \begin{array}{cc} \begin{array}{cccc} A & B & C & D \end{array} & \\ \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} & \begin{array}{c} A \\ B \\ C \\ D \end{array} \\ \begin{array}{cccc} \sum = 1 & \sum = 1 & \sum = 1 & \sum = 1 \end{array} & \end{array}$$



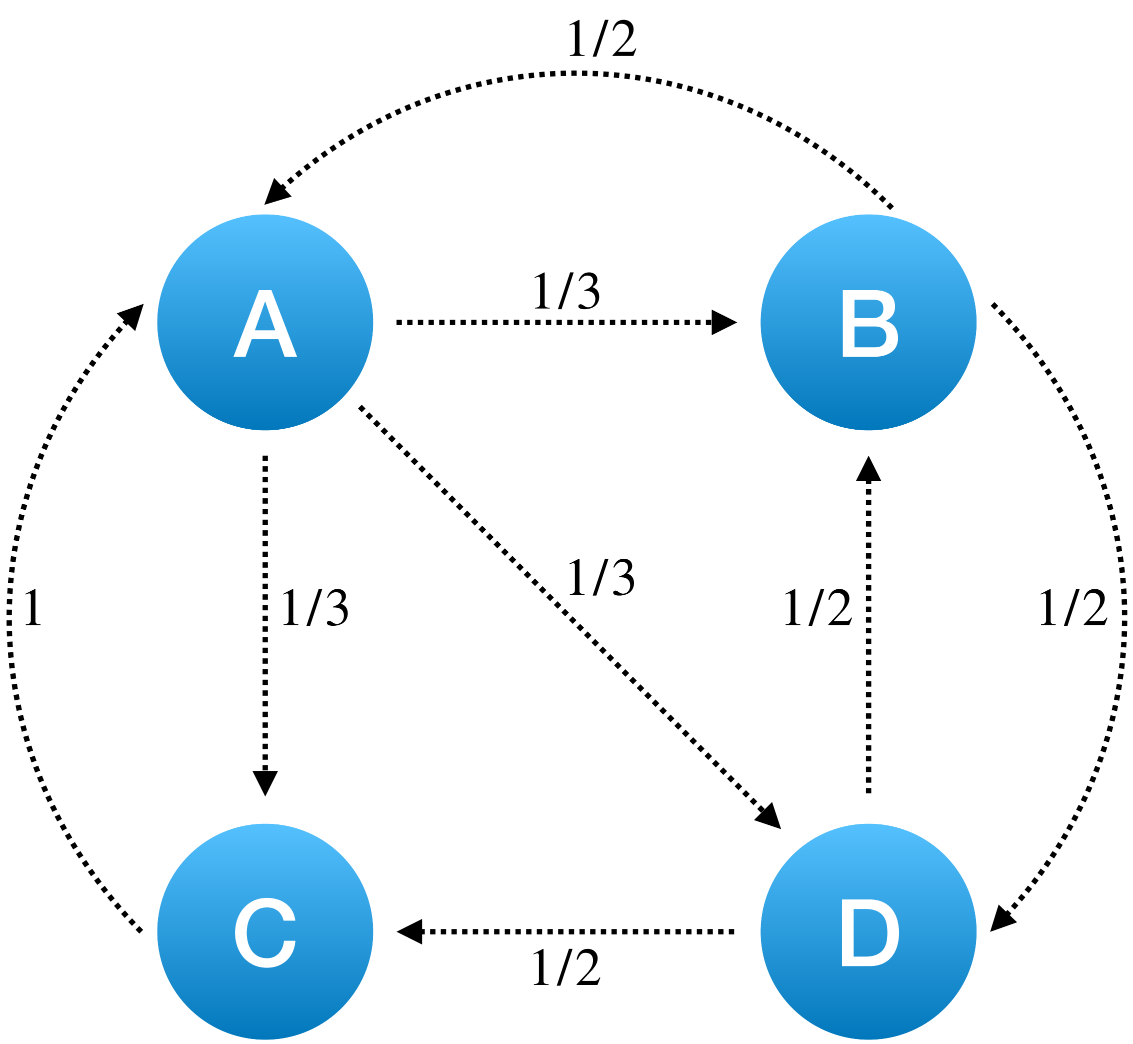
M- stochastic matrix

$v = \lambda M v$

v - principal eigenvector

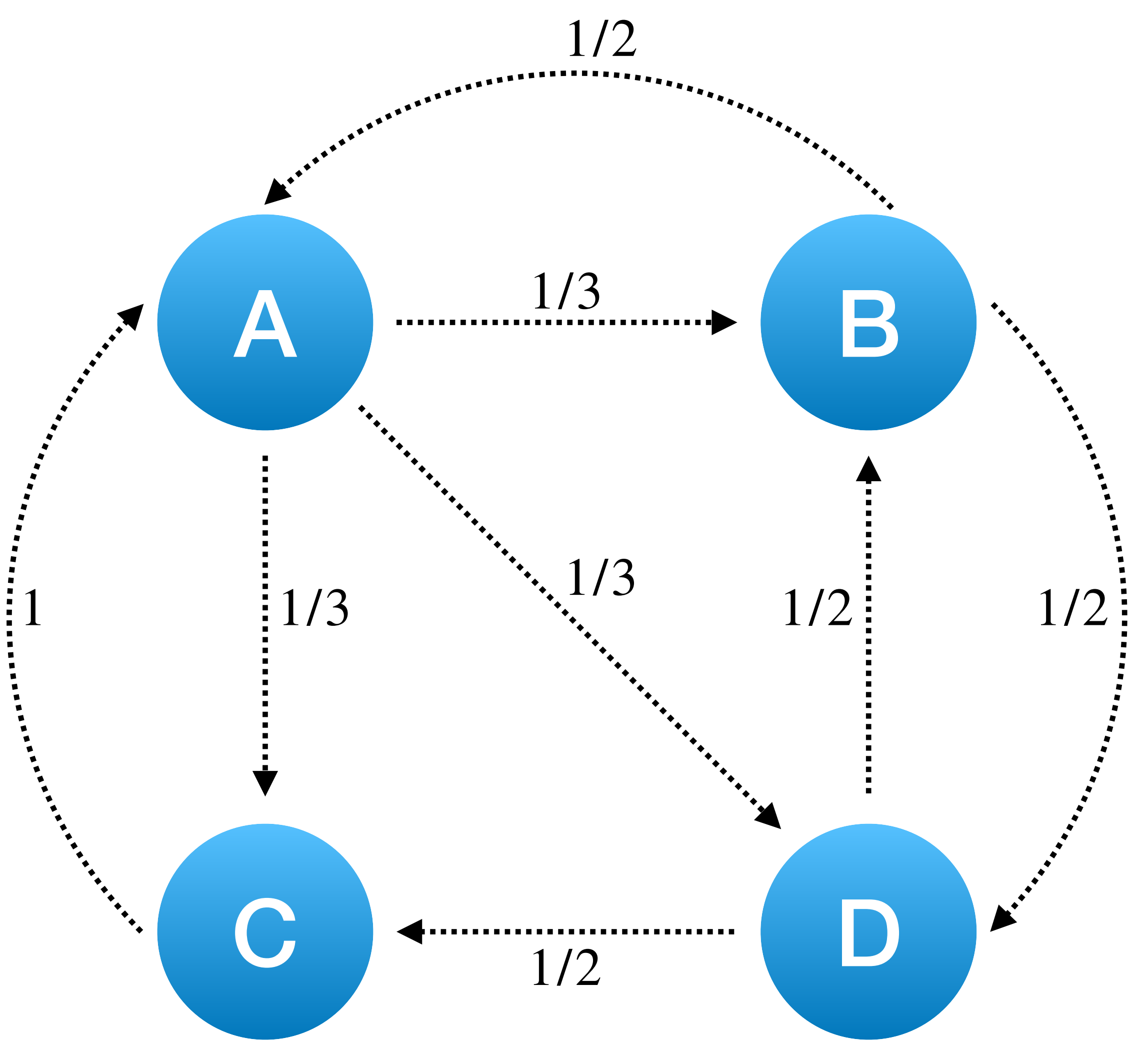
λ - eigenvalue

Let’s surf



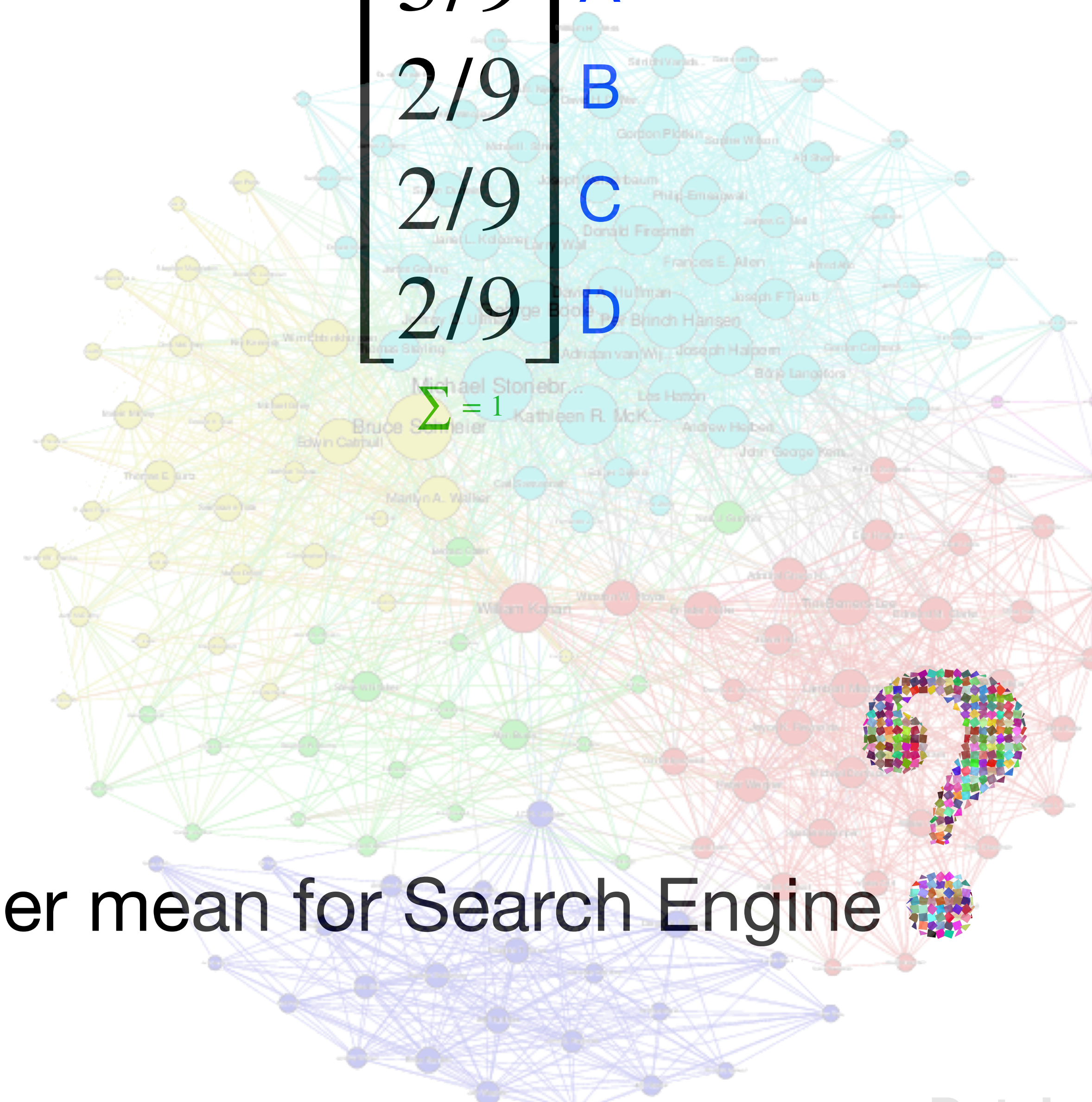
$$v_i = \begin{bmatrix}1/4\\1/4\\1/4\\1/4\end{bmatrix}, \begin{bmatrix}9/24\\5/24\\5/24\\5/24\end{bmatrix}, \begin{bmatrix}15/48\\11/48\\11/48\\11/48\end{bmatrix}, \begin{bmatrix}11/32\\7/32\\7/32\\7/32\end{bmatrix}, \ldots, \begin{bmatrix}3/9\\2/9\\2/9\\2/9\end{bmatrix}$$

Let’s surf



$$\begin{array}{cl} \begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix} & \begin{matrix} A \\ B \\ C \\ D \end{matrix} \end{array}$$

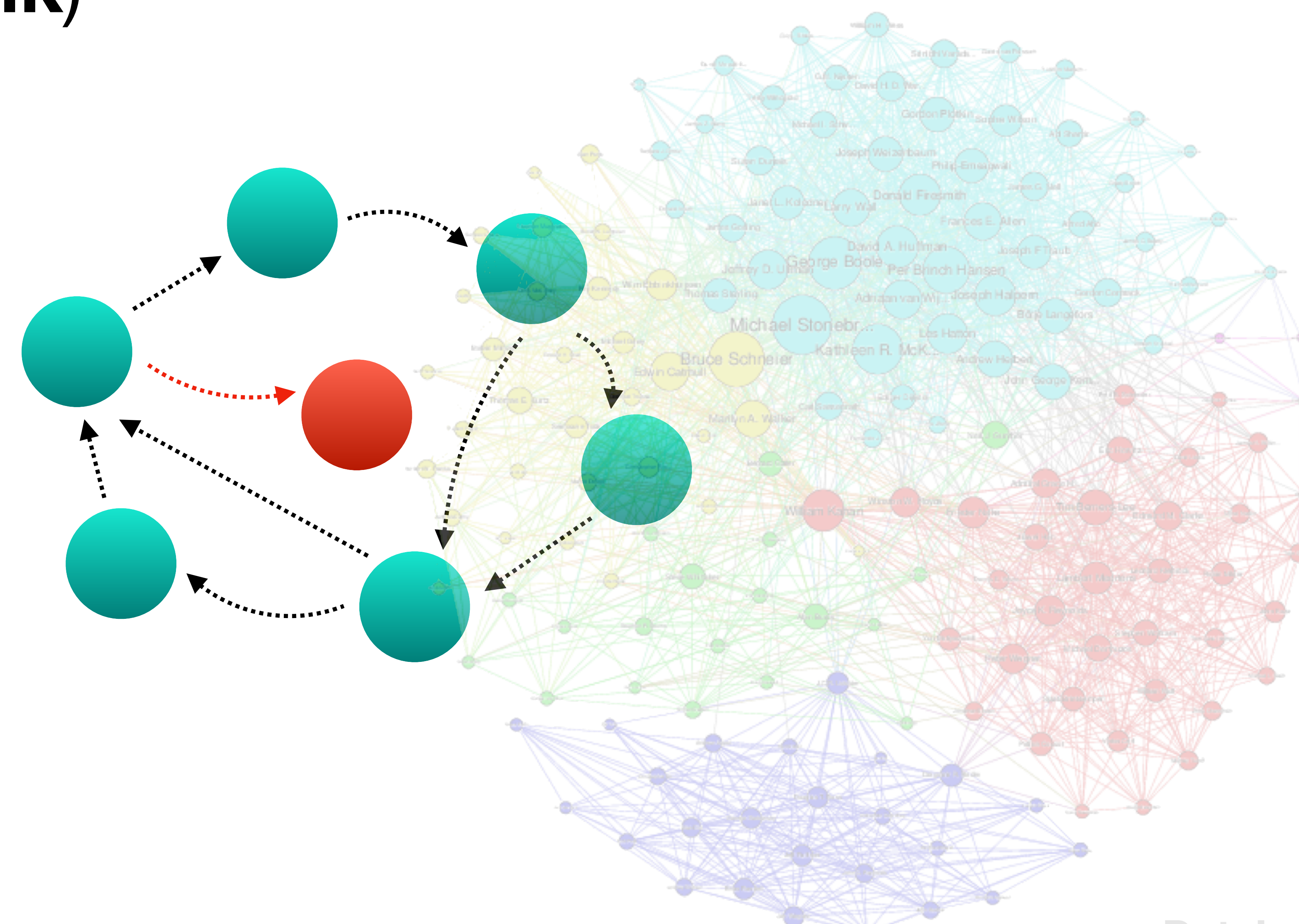
$\sum = 1$



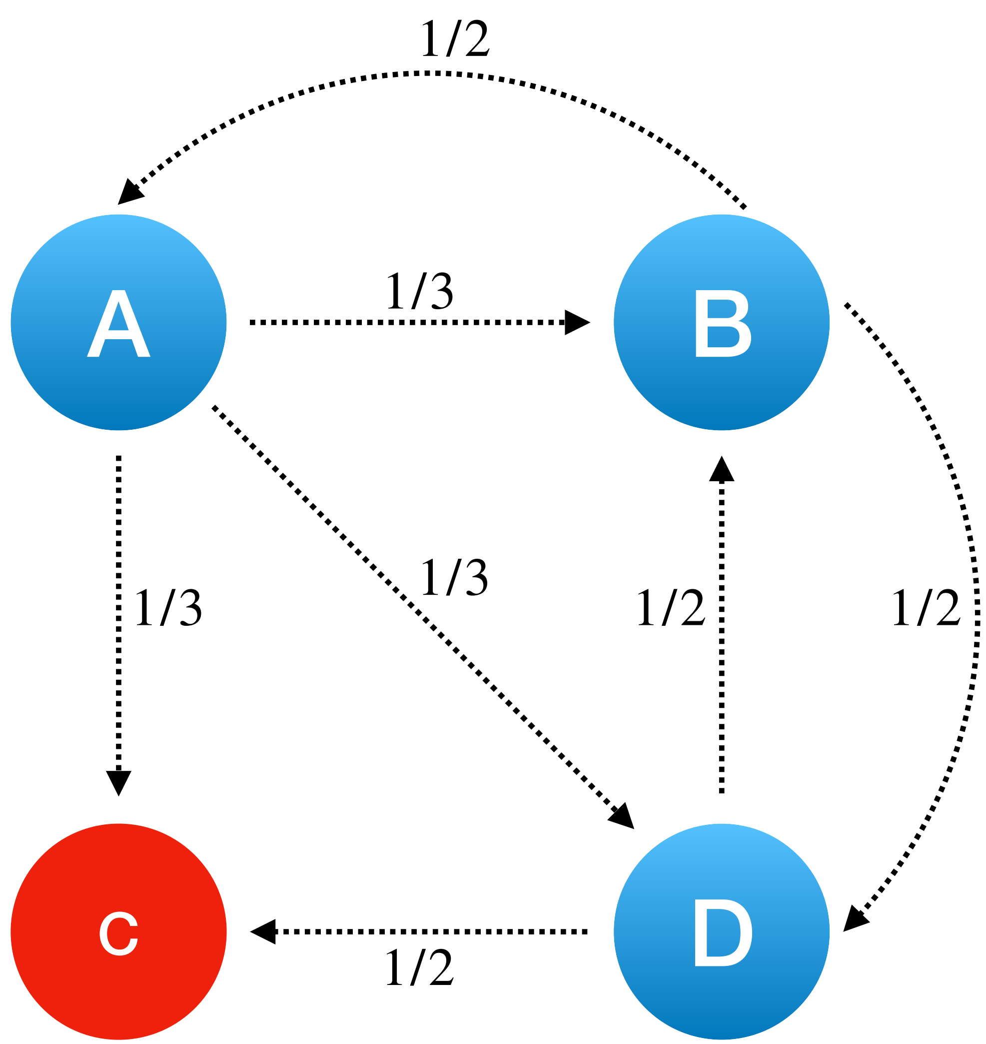
What does particular number mean for Search Engine ?

Dead Ends

- Dead-ends occur when pages have no out-links. (~ **dangling link**)



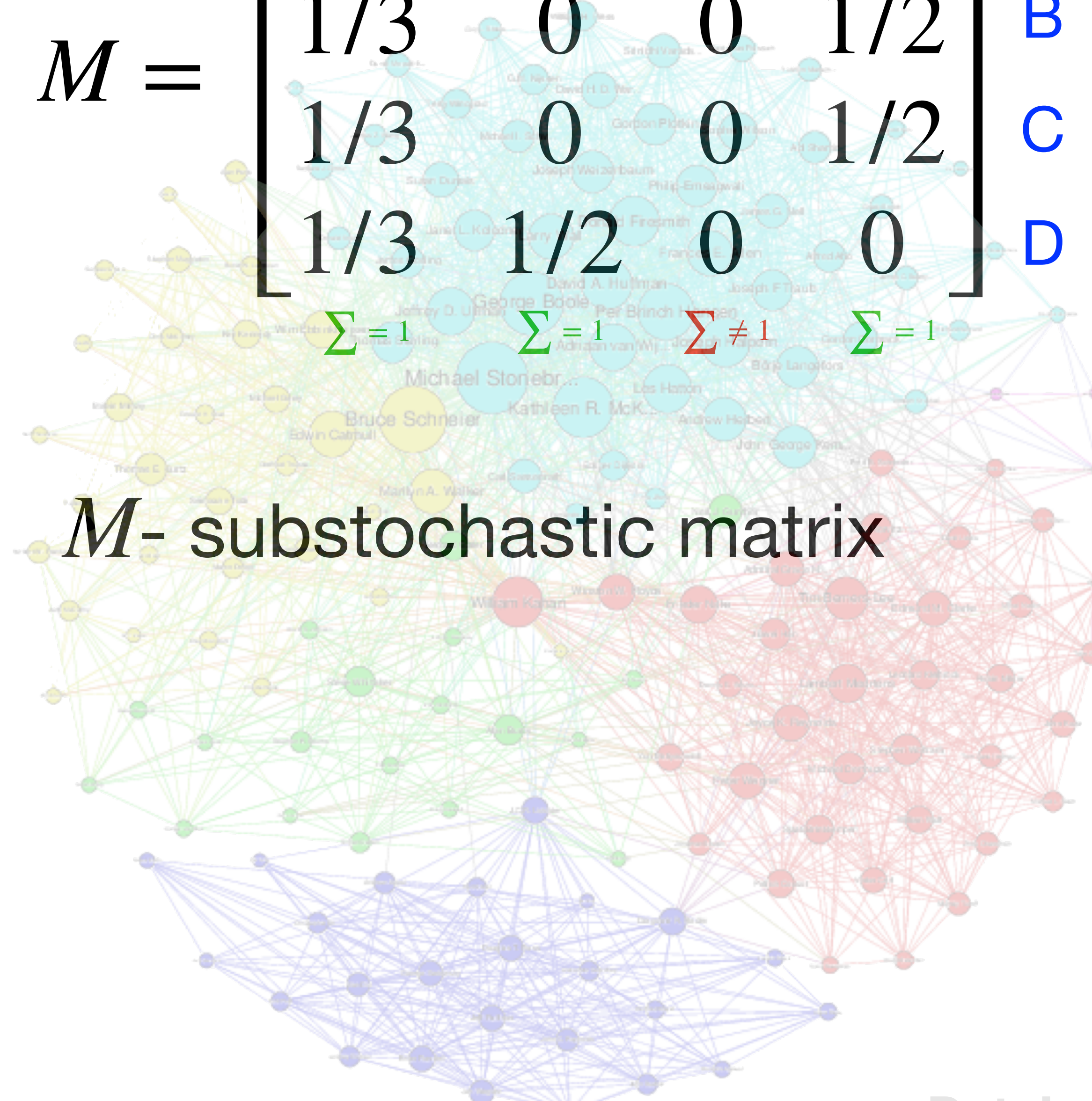
Dead Ends



$$M = \begin{array}{c} \\ \\ \end{array}\begin{array}{cccc} A & B & C & D \end{array}$$

$$M = \begin{array}{cc} & \begin{array}{cccc} \quad A \quad & \quad B \quad & \quad C \quad & \quad D \quad \end{array} \\ & \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \begin{array}{c} A \\ B \\ C \\ D \end{array} \\ & \begin{array}{cccc} \sum = 1 & \sum = 1 & \sum \neq 1 & \sum = 1 \end{array} \end{array}$$

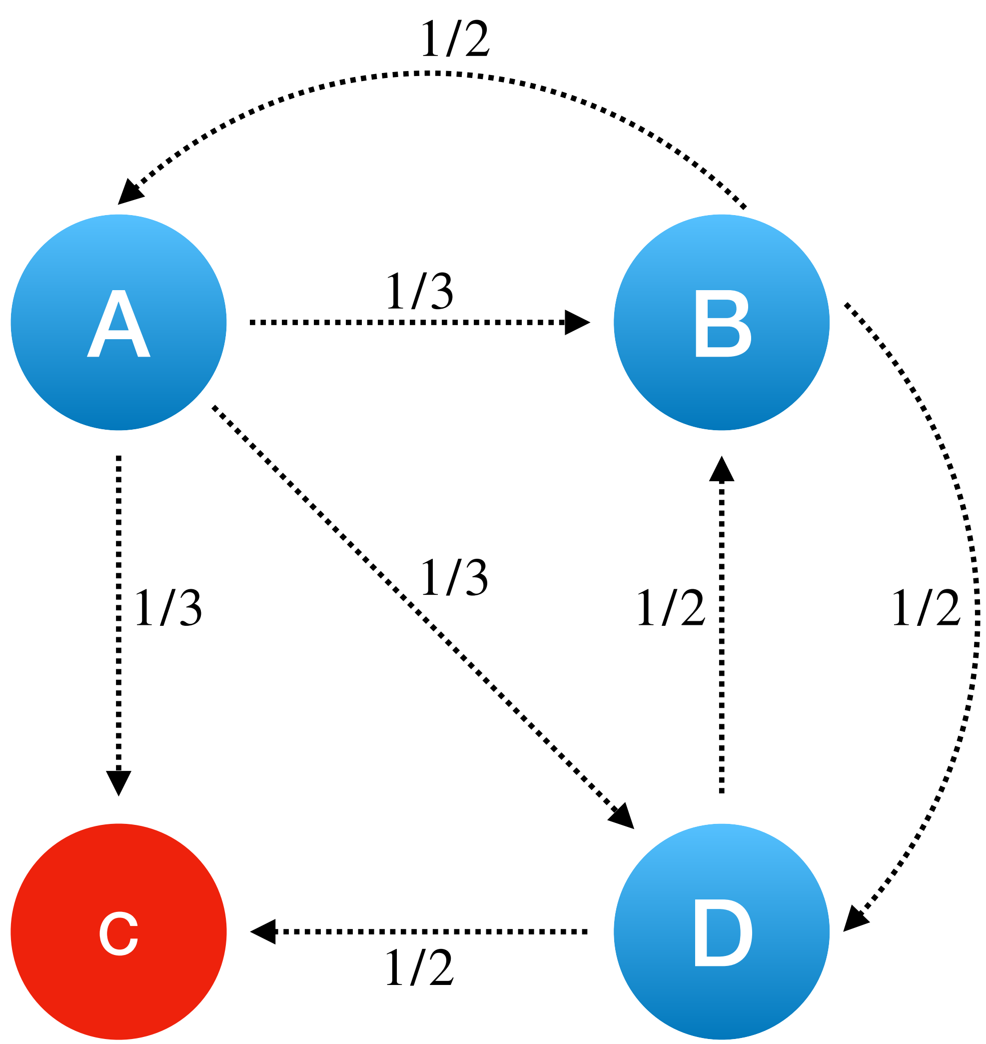
M- substochastic matrix



Yakupov Azat

DataLab

Dead Ends

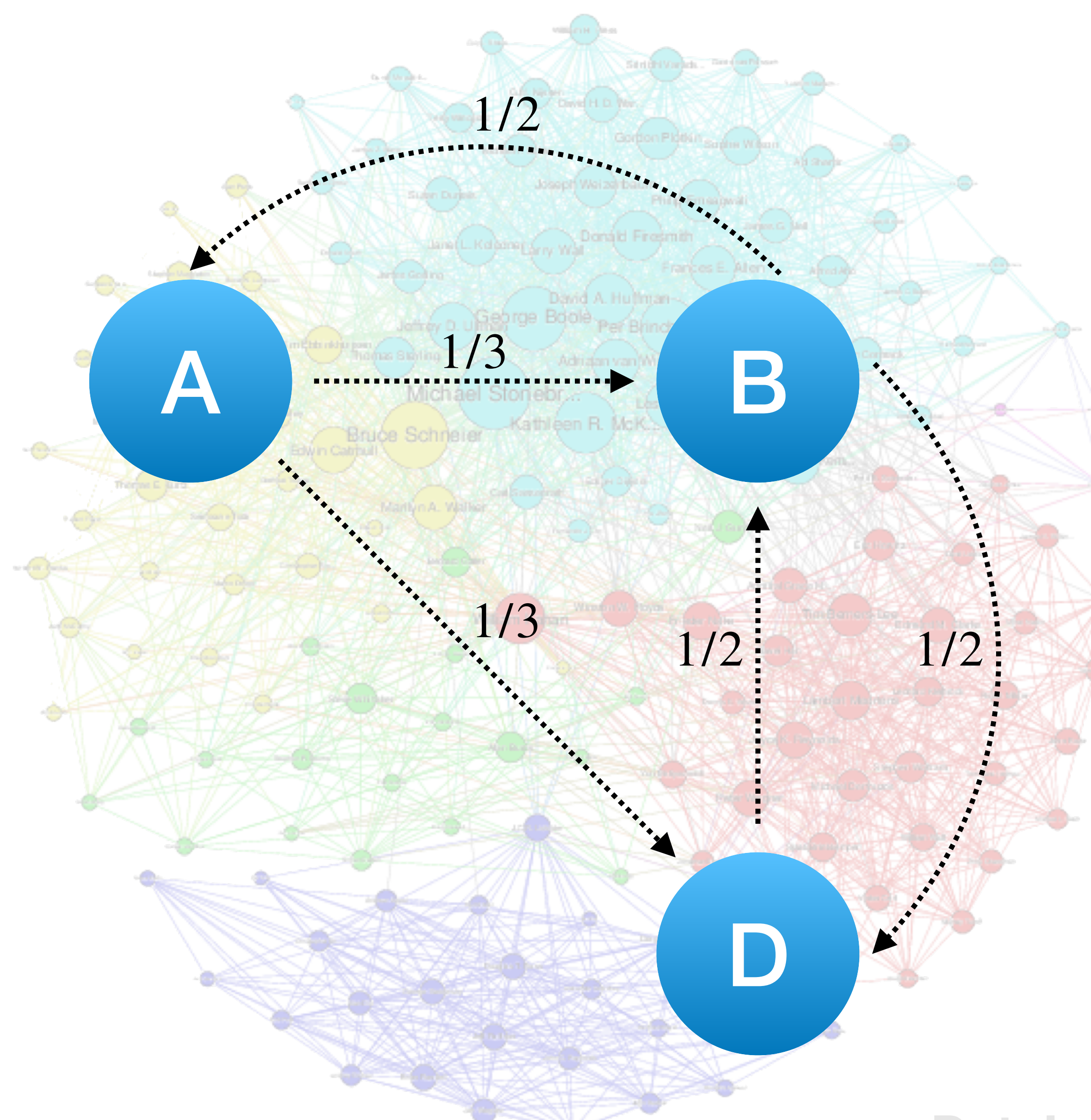


$$v_i = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 3/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}, \begin{bmatrix} 5/48 \\ 7/48 \\ 7/48 \\ 7/48 \end{bmatrix},$$

$$\begin{bmatrix} 21/288 \\ 31/288 \\ 31/288 \\ 31/288 \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

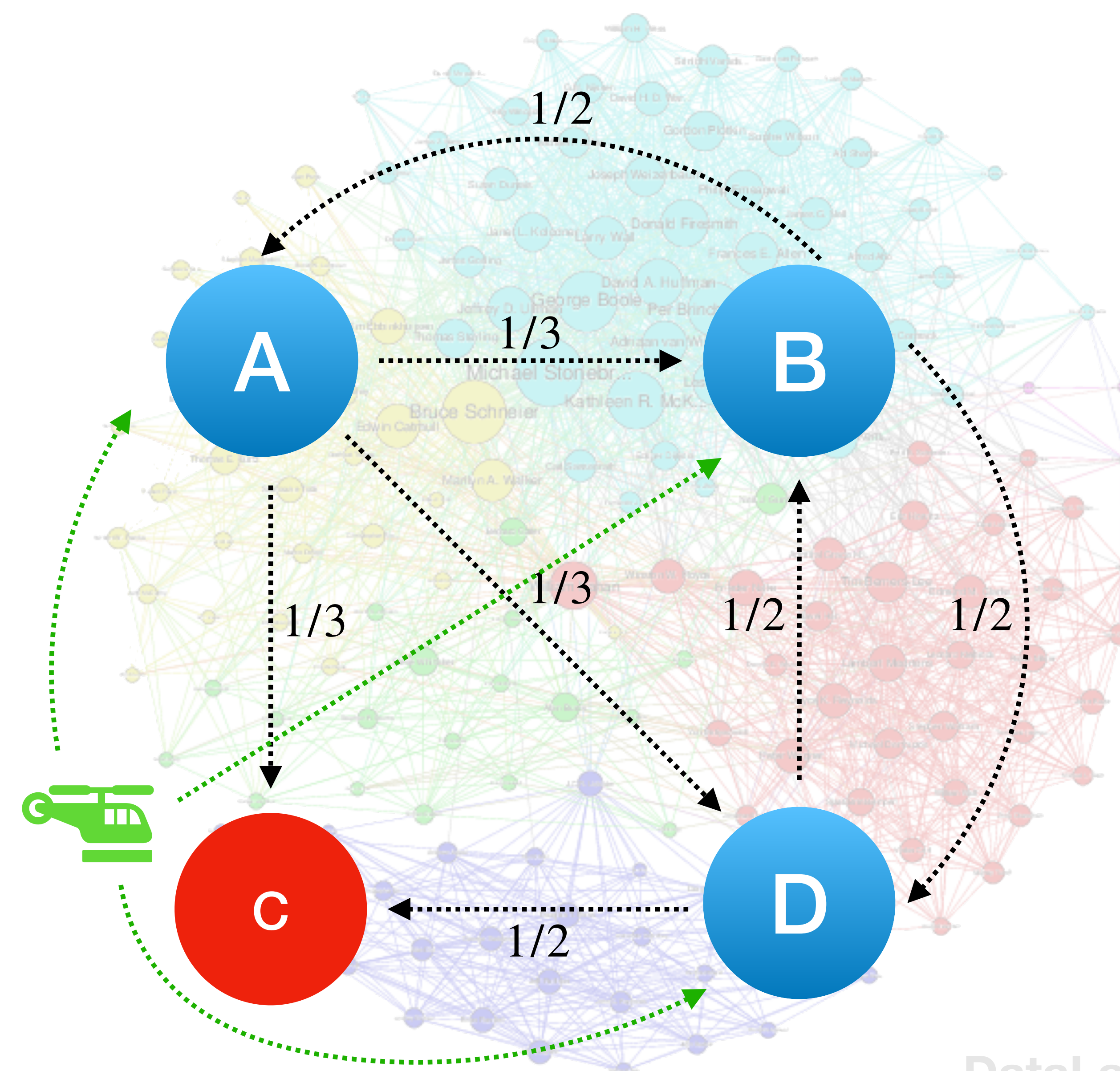
Dead Ends

- Drop the dead ends from the graph and also drop their incoming links

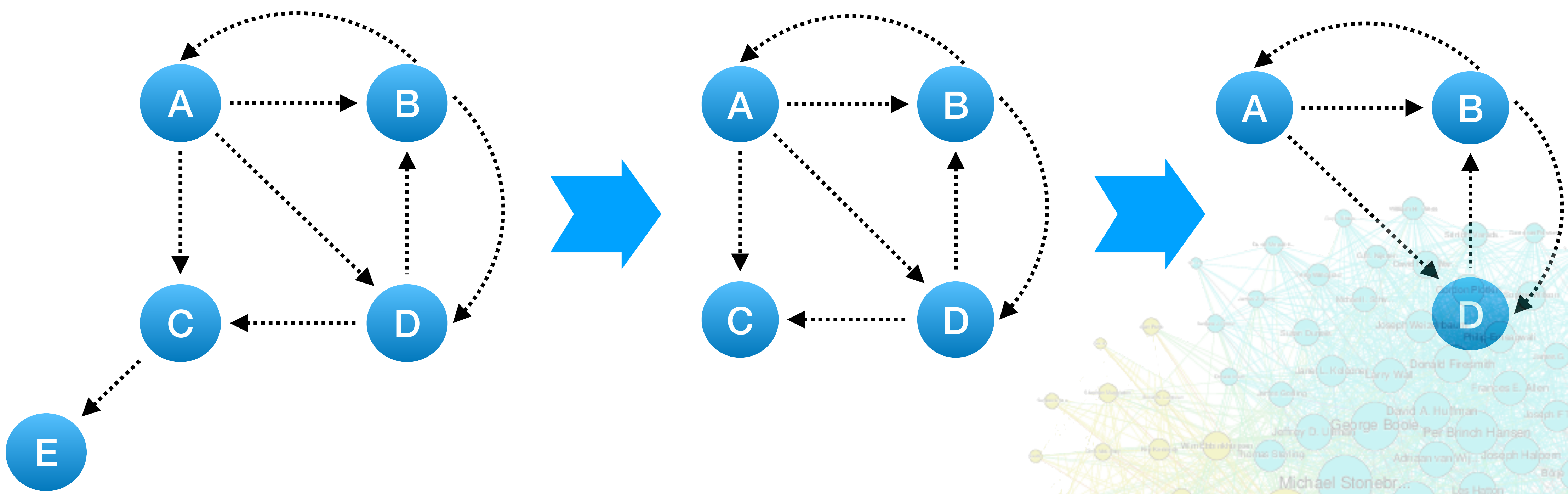


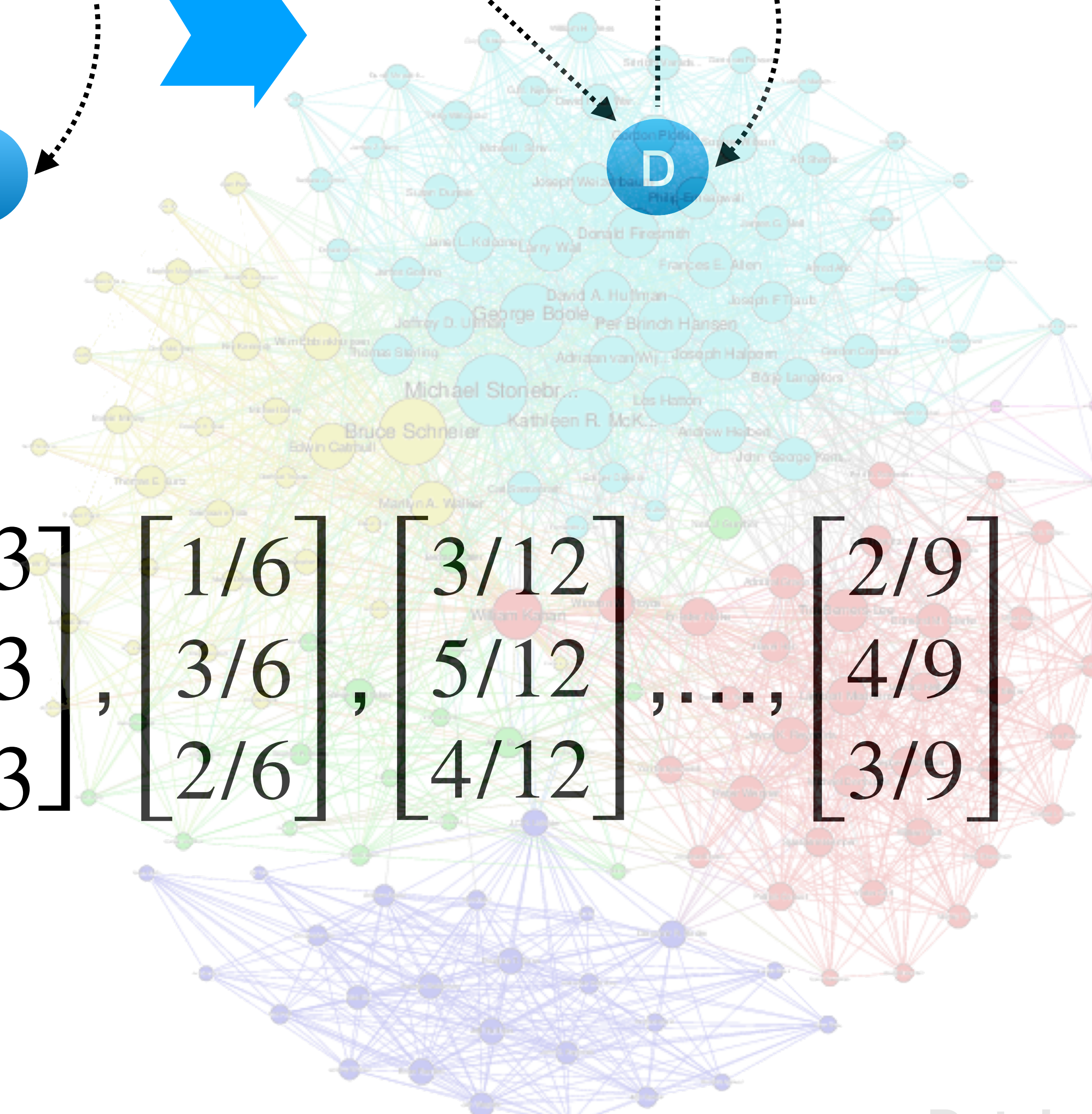
Dead Ends

- "Taxation" method



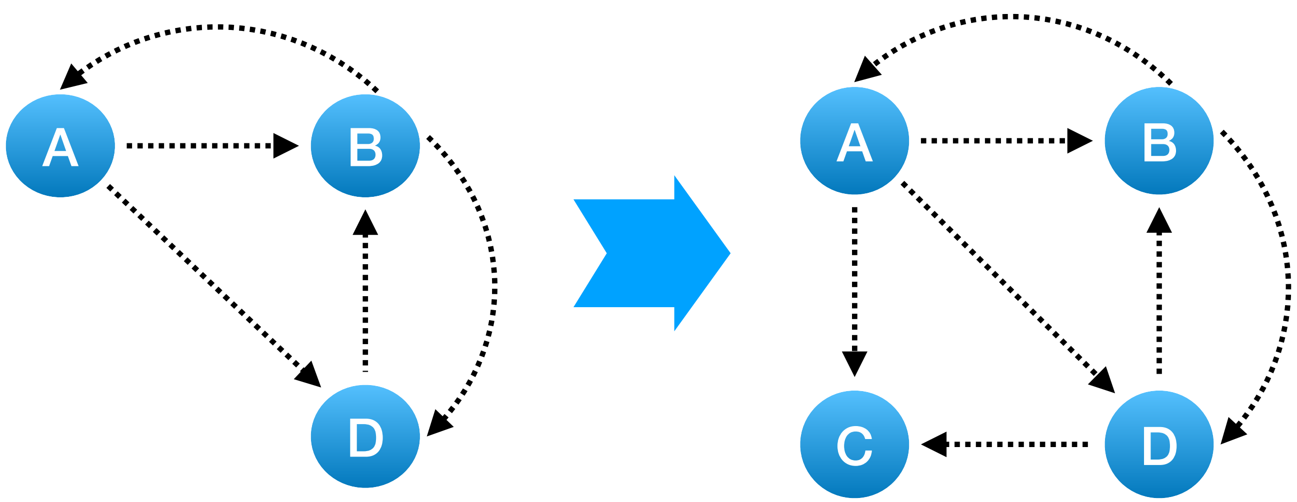
Dead Ends



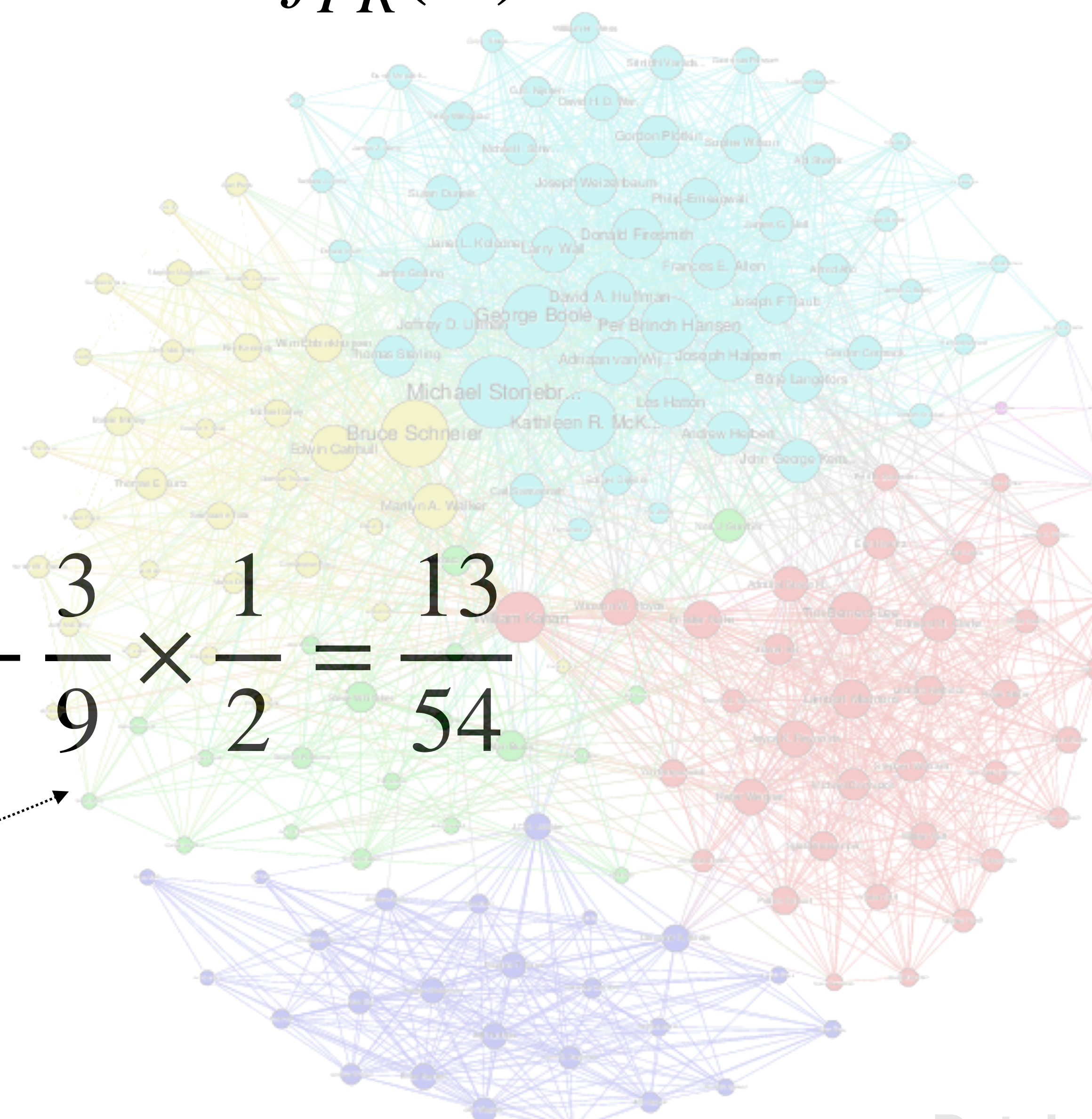


$$M = \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix} \quad v_i = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \begin{bmatrix} 1/6 \\ 3/6 \\ 2/6 \end{bmatrix}, \begin{bmatrix} 3/12 \\ 5/12 \\ 4/12 \end{bmatrix}, \ldots, \begin{bmatrix} 2/9 \\ 4/9 \\ 3/9 \end{bmatrix}$$

Dead Ends



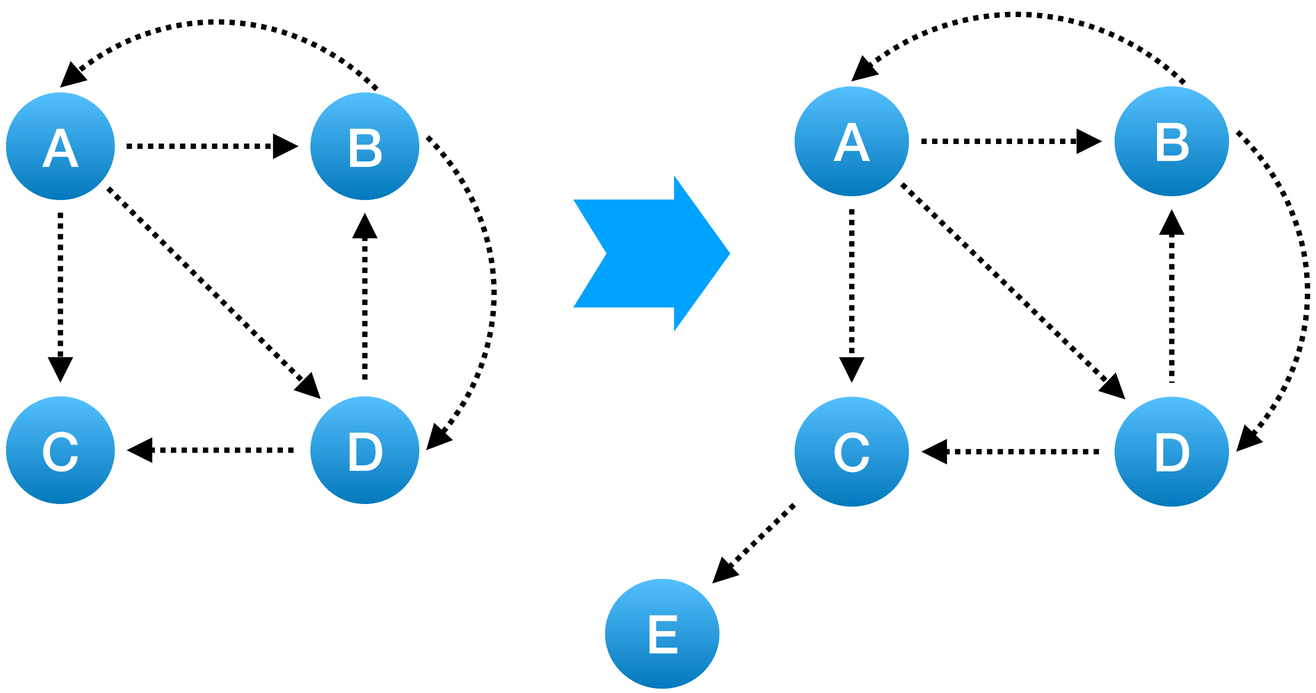
$$f_{PR}(C) = ?$$



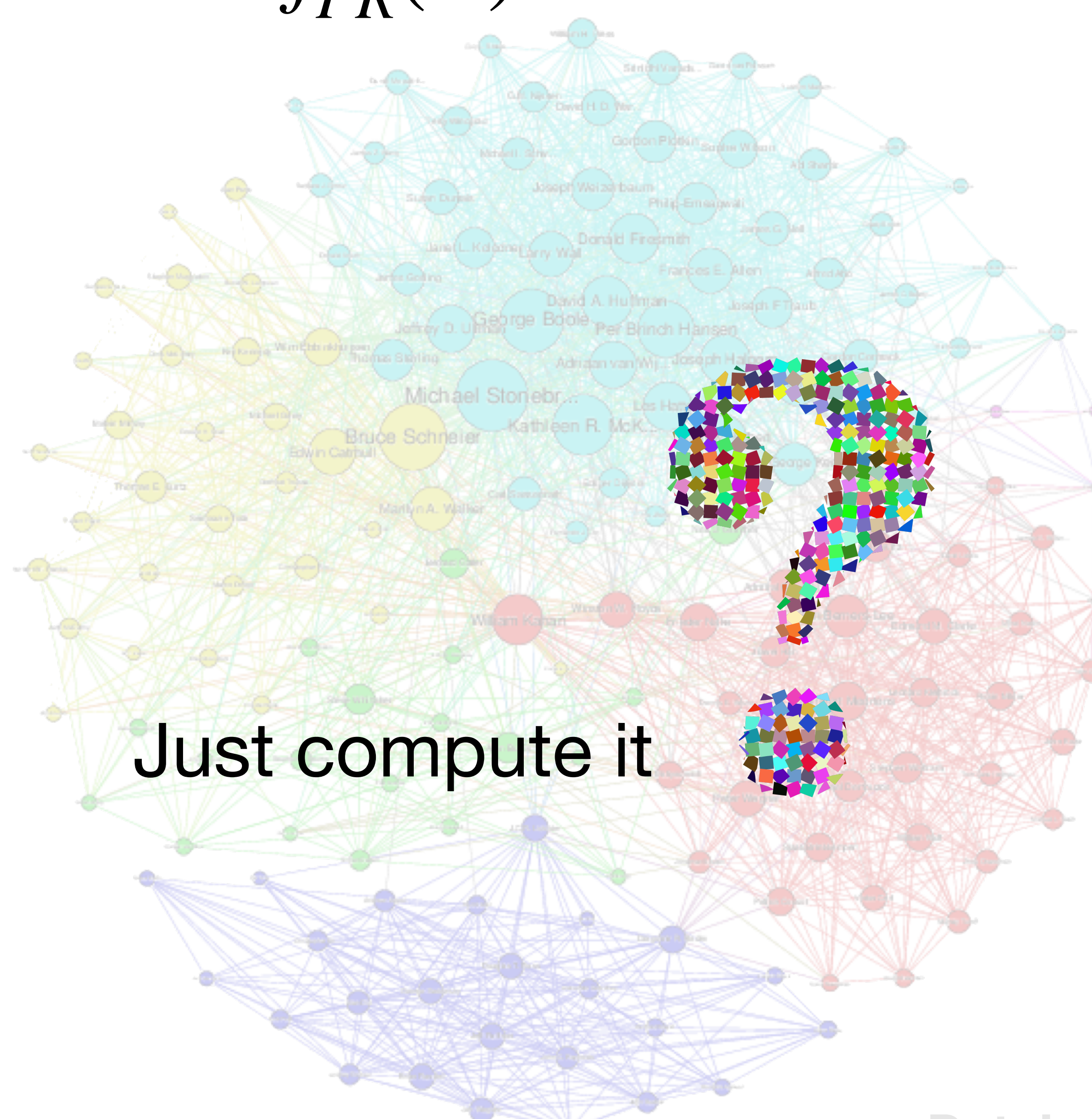
$$\begin{array}{c} \color{blue}{A} \\ \color{blue}{B} \\ \color{blue}{D} \end{array} \begin{bmatrix} 2/9 \\ 4/9 \\ 3/9 \end{bmatrix}$$

$$f_{PR}(C) = \frac{2}{9} \times \frac{1}{3} + \frac{3}{9} \times \frac{1}{2} = \frac{13}{54}$$

Dead Ends

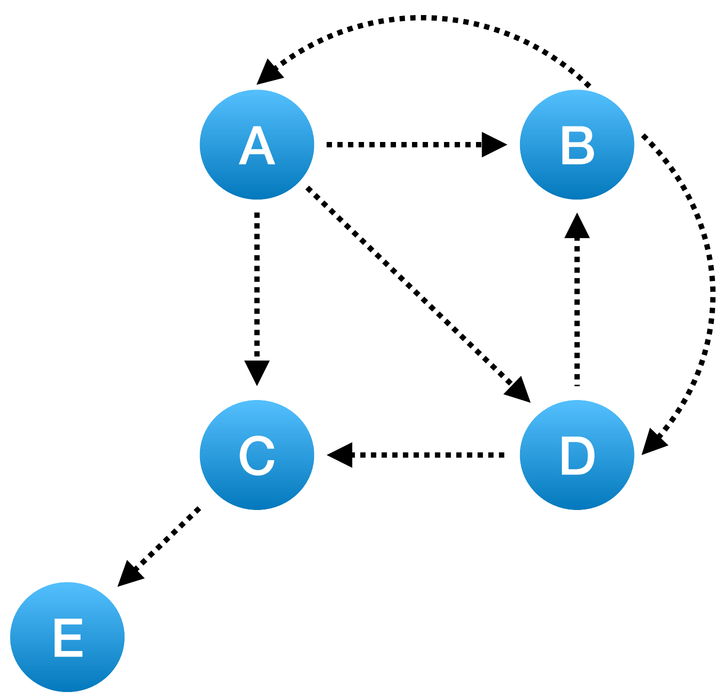


$$f_{PR}(E) = ?$$

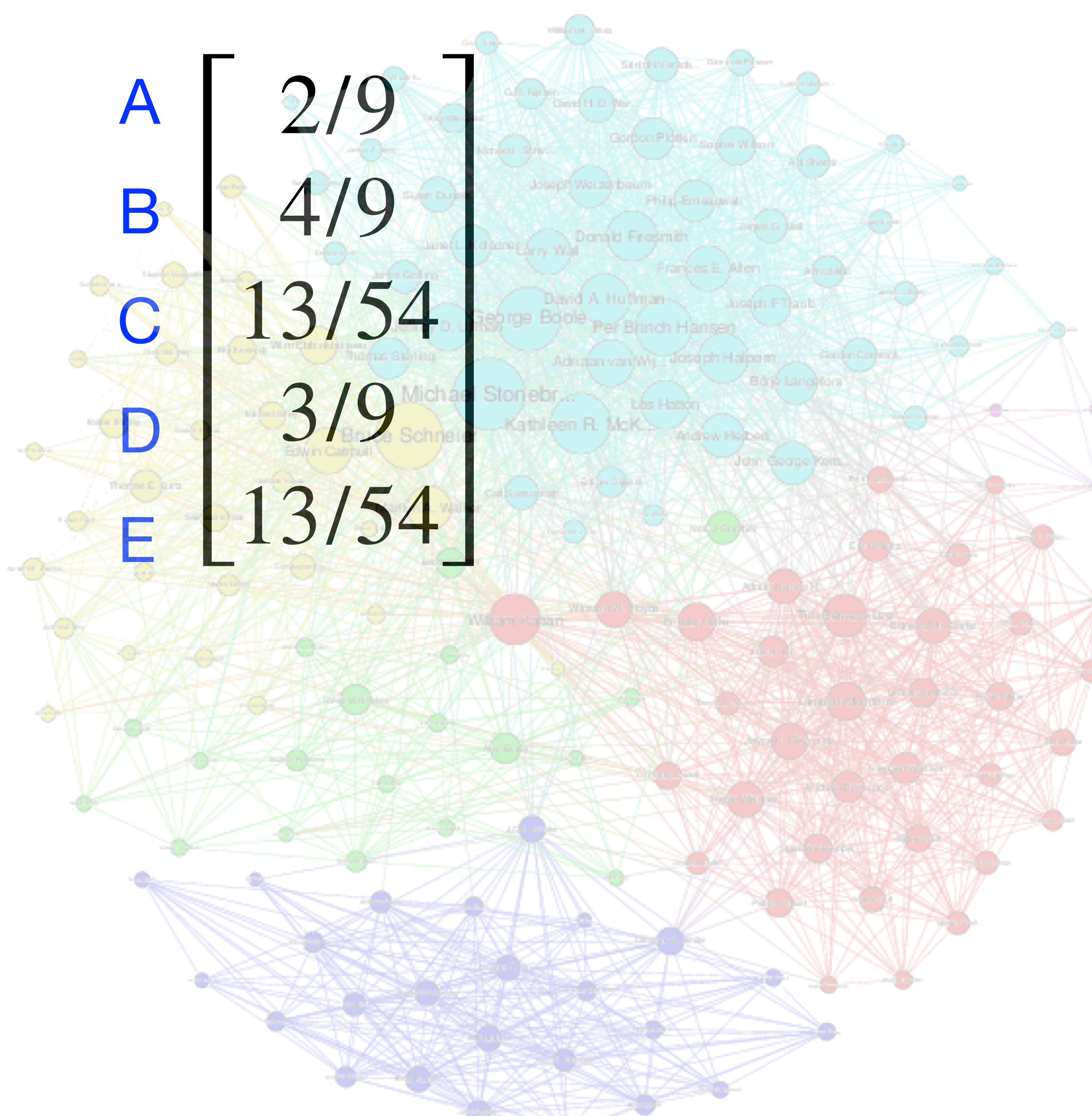


$$\begin{array}{c} \color{blue}{A} \\ \color{blue}{B} \\ \color{blue}{C} \\ \color{blue}{D} \end{array} \begin{bmatrix} 2/9 \\ 4/9 \\ 13/54 \\ 3/9 \end{bmatrix}$$

Dead Ends

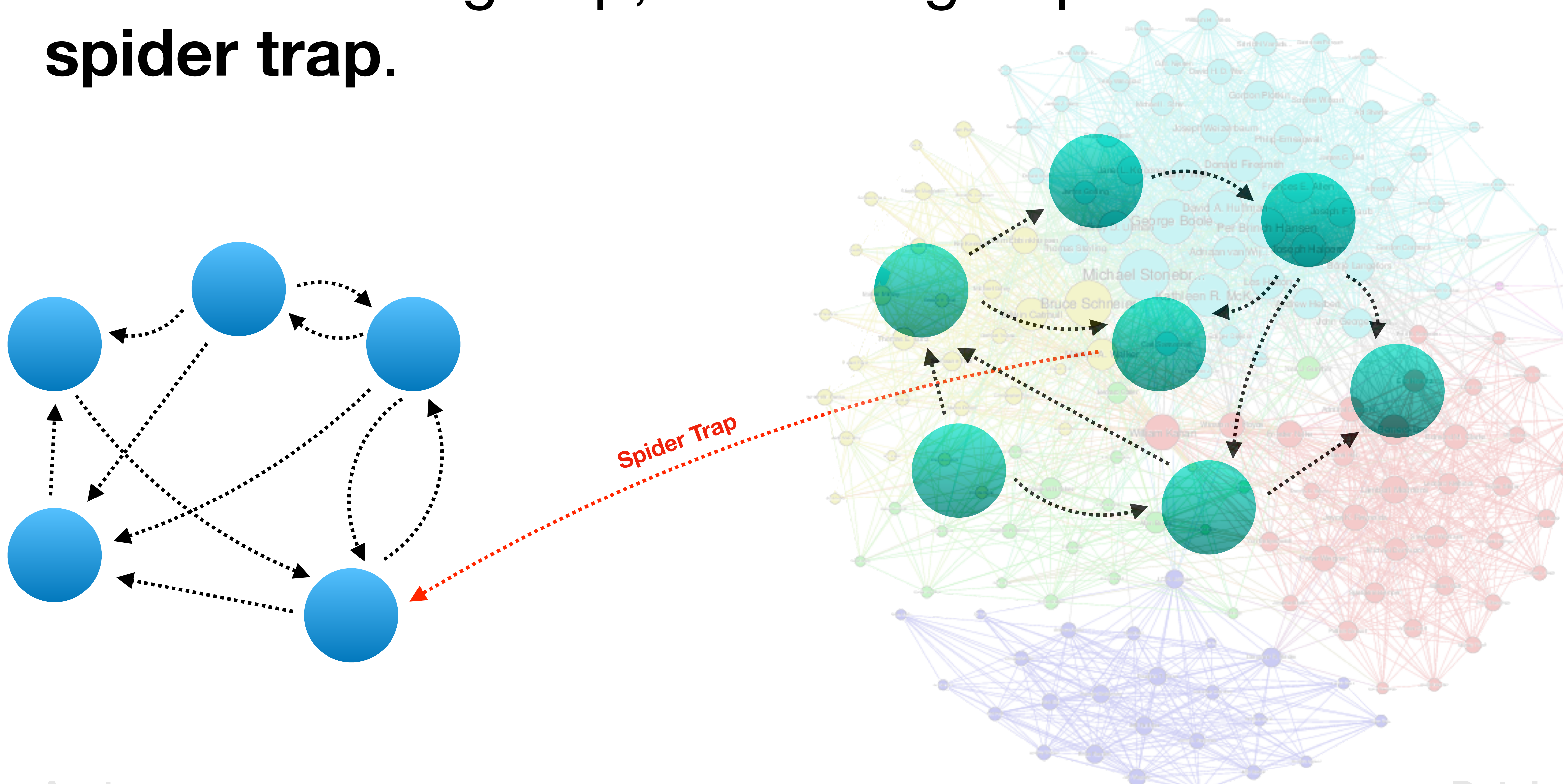


$$\begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} \begin{bmatrix} 2/9 \\ 4/9 \\ 13/54 \\ 3/9 \\ 13/54 \end{bmatrix}$$

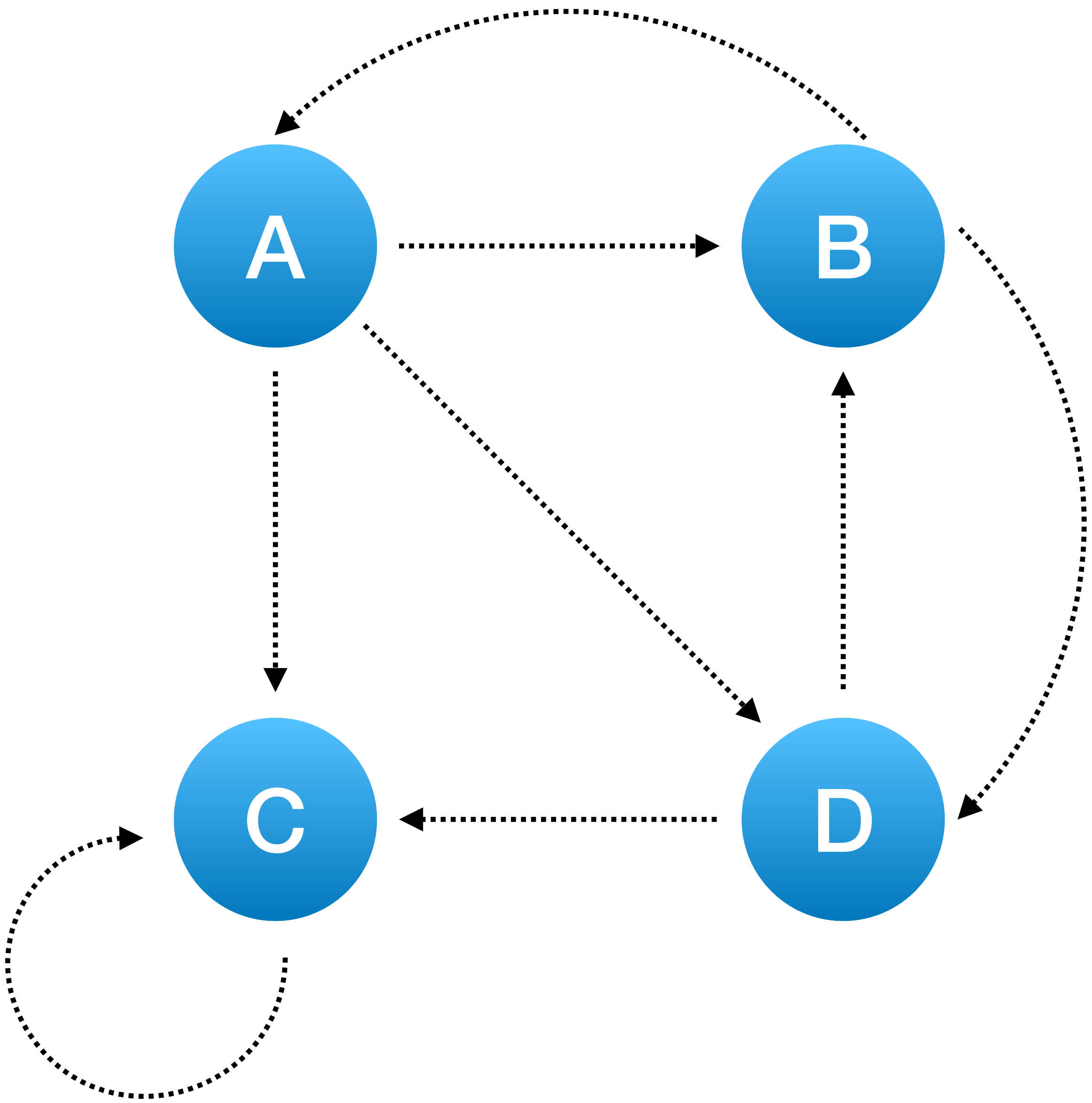


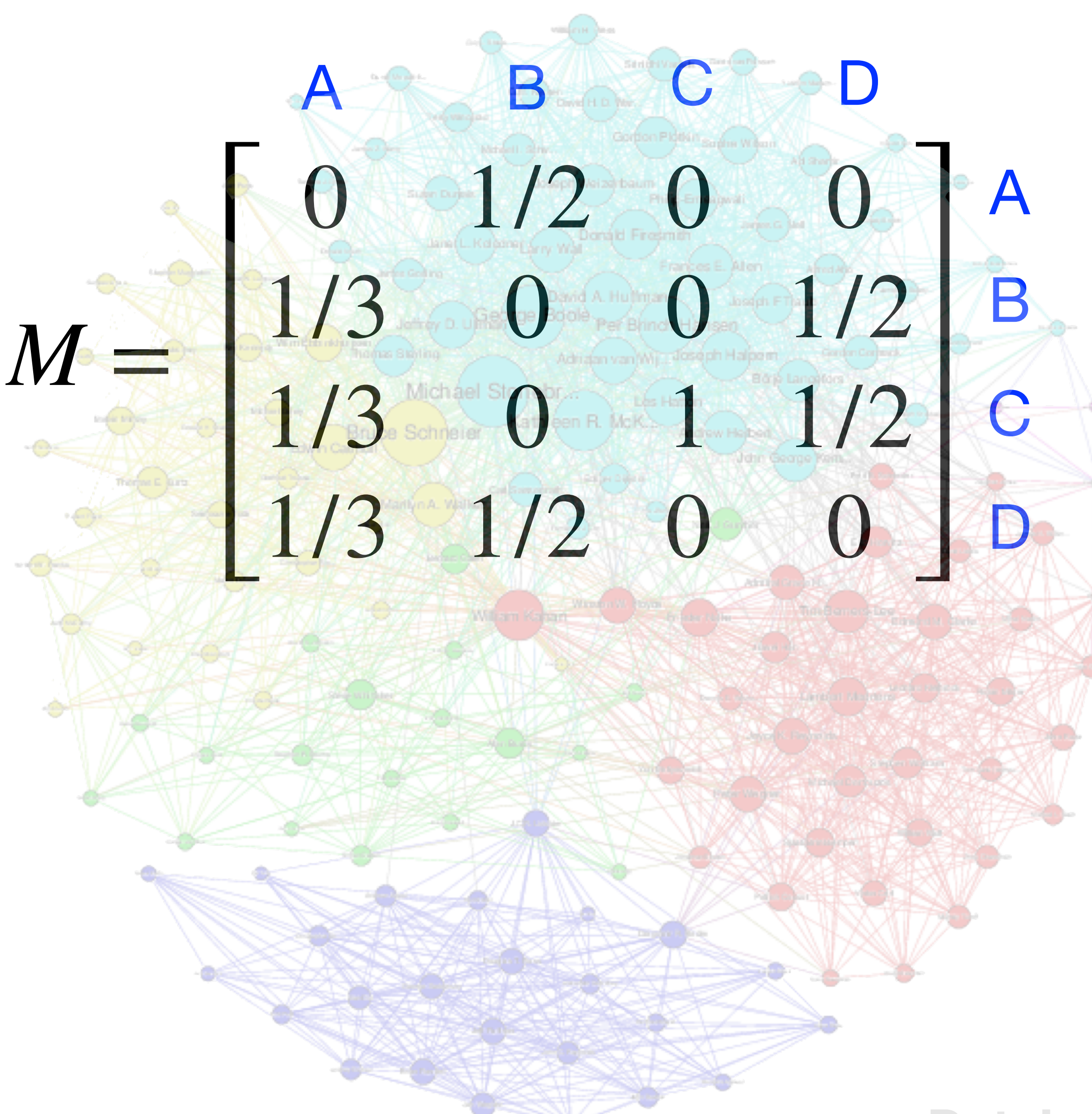
Spider Trap

- If there are no links from within a group of pages to outside of the group, then the group is considered a **spider trap**.



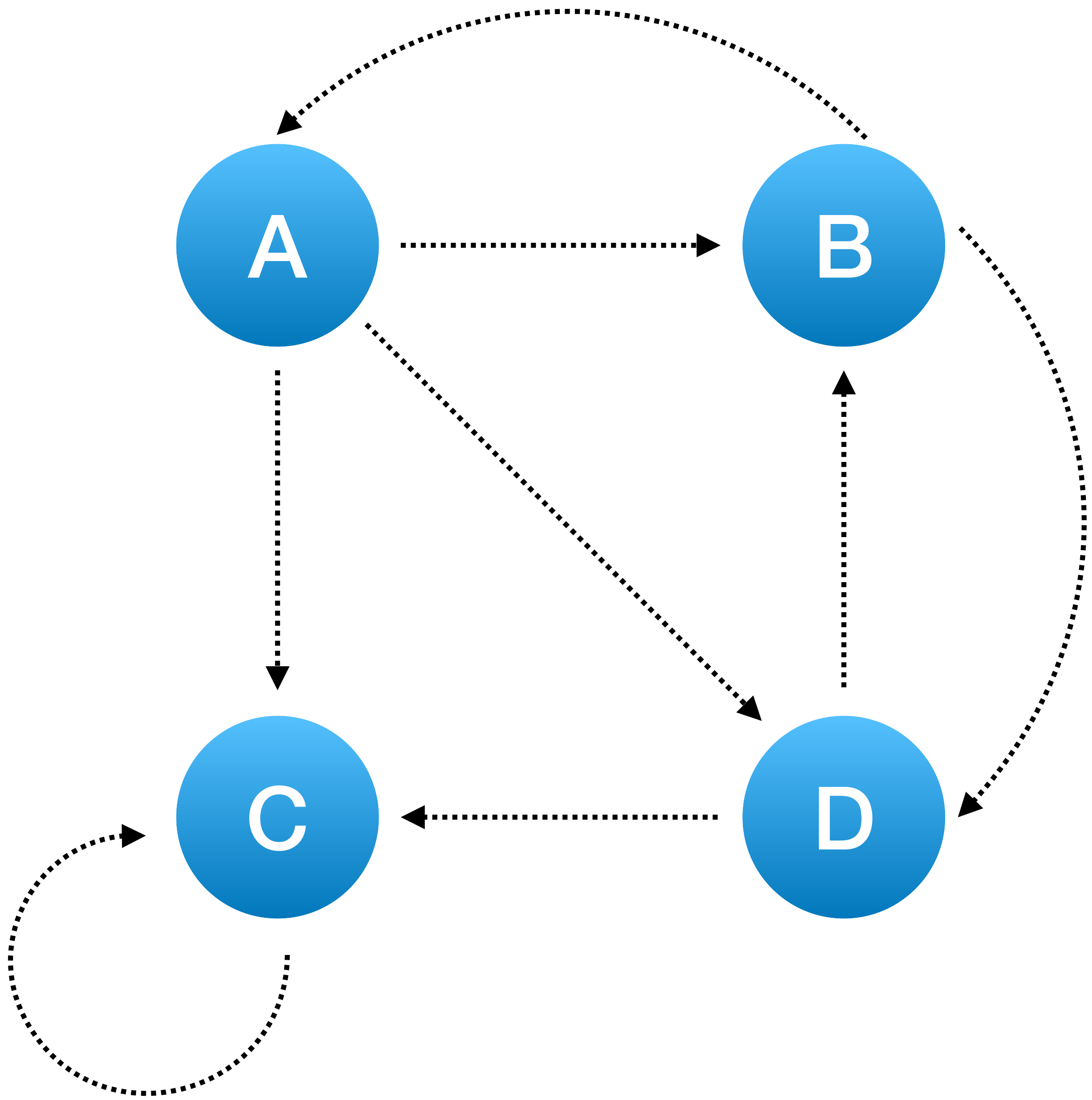
Spider Trap





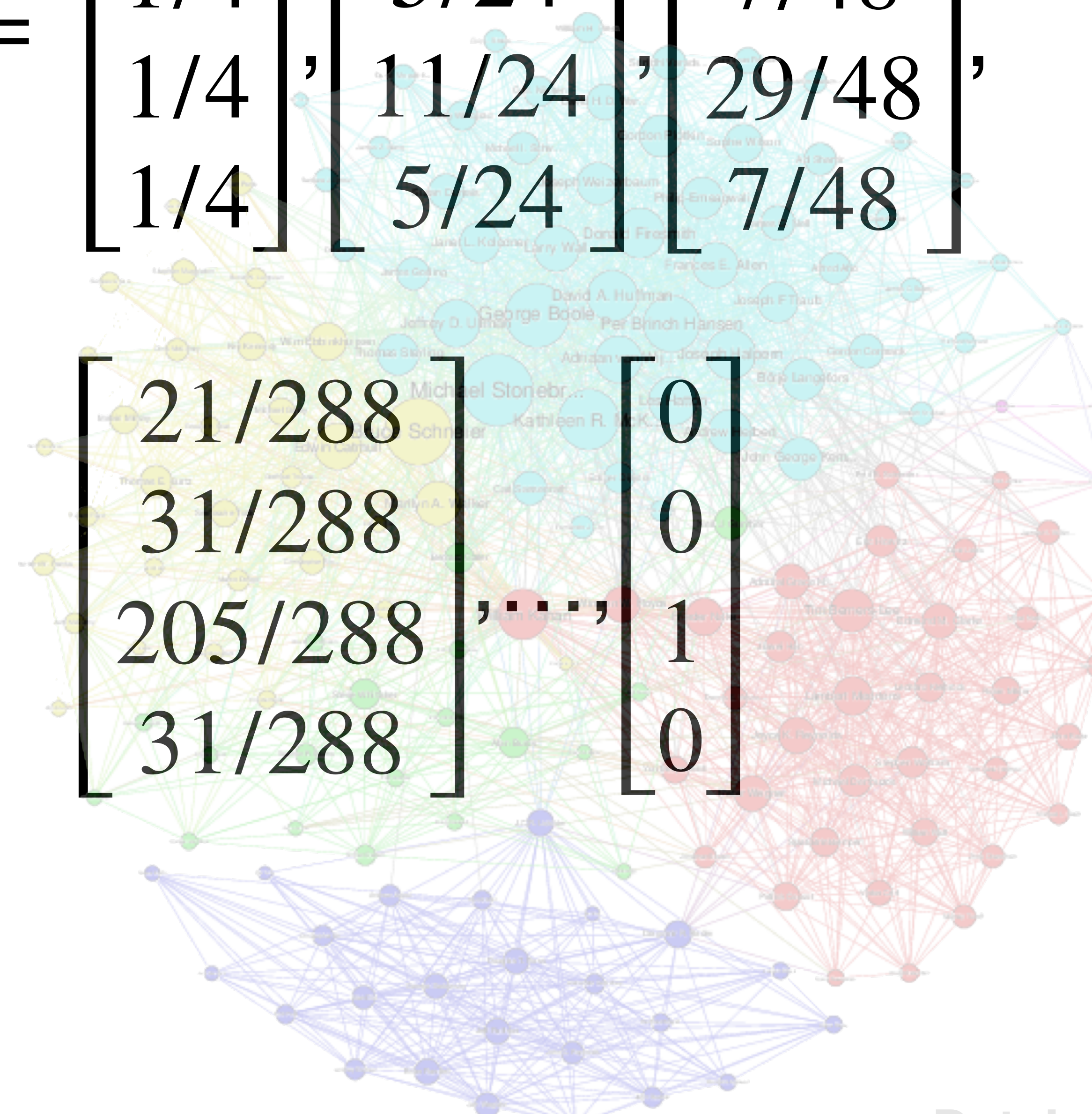
$$M = \begin{array}{c} \\ \\ \end{array}\begin{array}{cc} \begin{array}{cccc} \quad A & \quad B & \quad C & \quad D \end{array} & \\ \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} & \begin{array}{c} A \\ B \\ C \\ D \end{array} \end{array}$$

Spider Trap

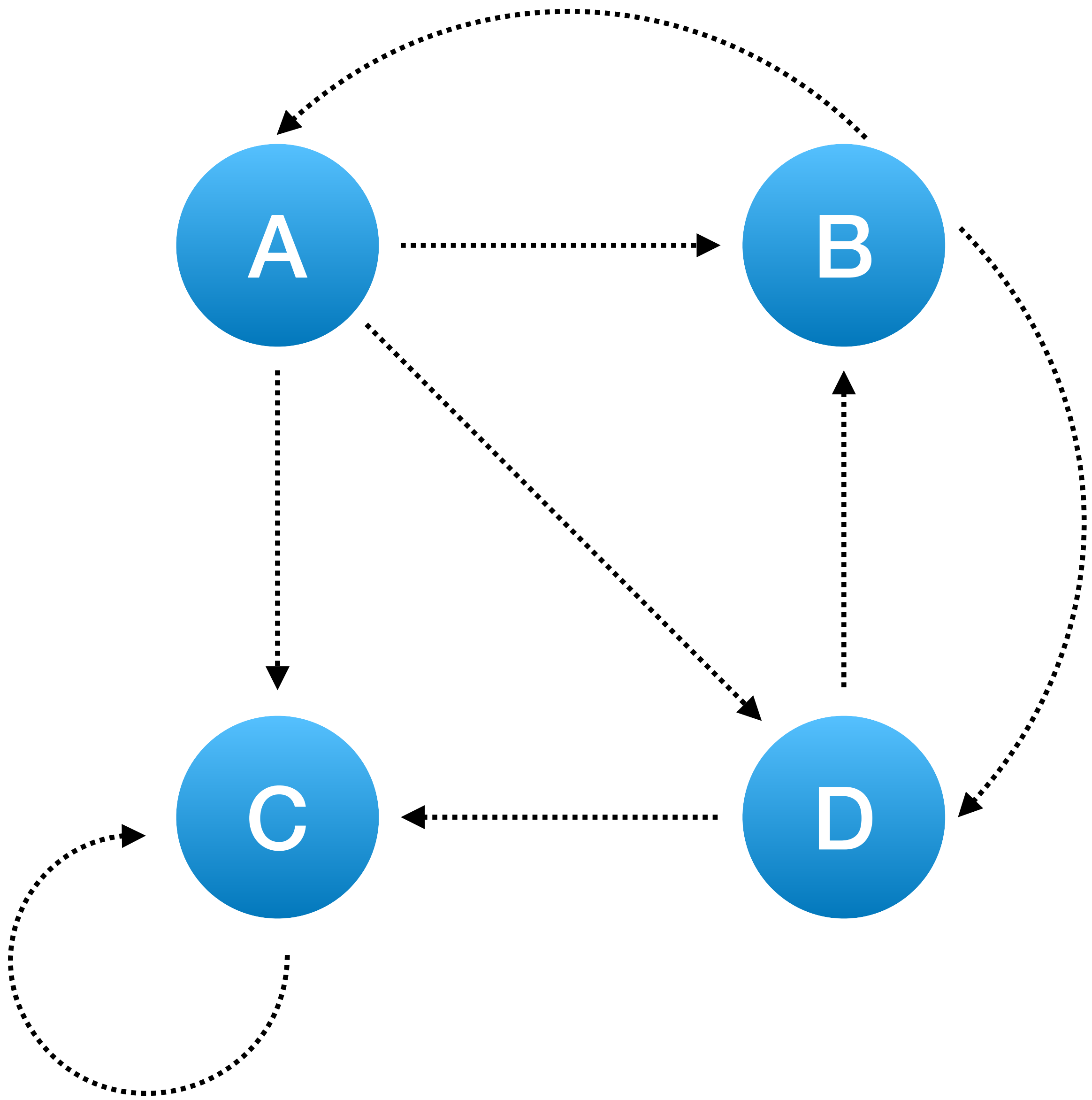


$$v_i = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 3/24 \\ 5/24 \\ 11/24 \\ 5/24 \end{bmatrix}, \begin{bmatrix} 5/48 \\ 7/48 \\ 29/48 \\ 7/48 \end{bmatrix},$$

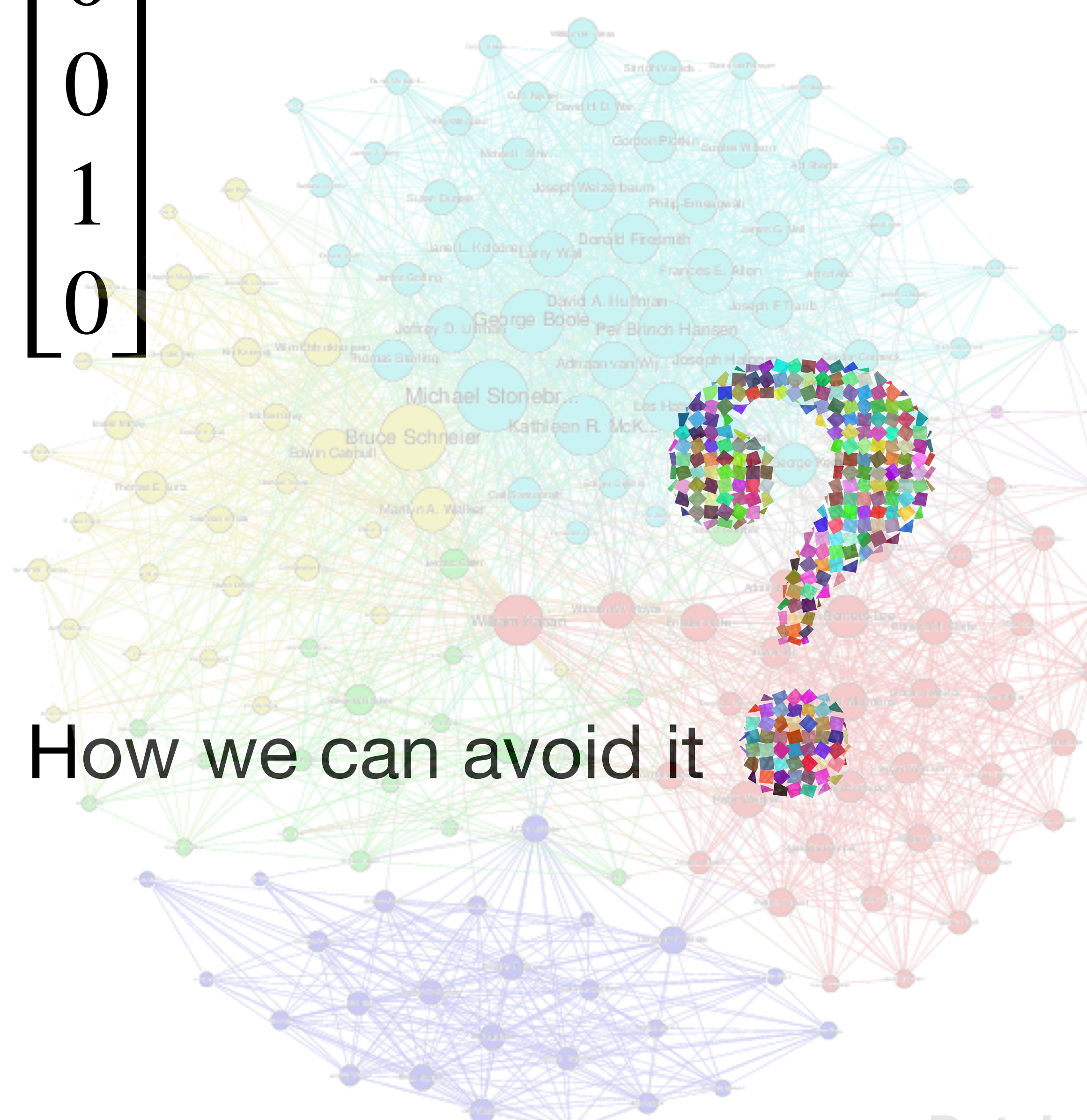
$$\begin{bmatrix} 21/288 \\ 31/288 \\ 205/288 \\ 31/288 \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$



Spider Trap

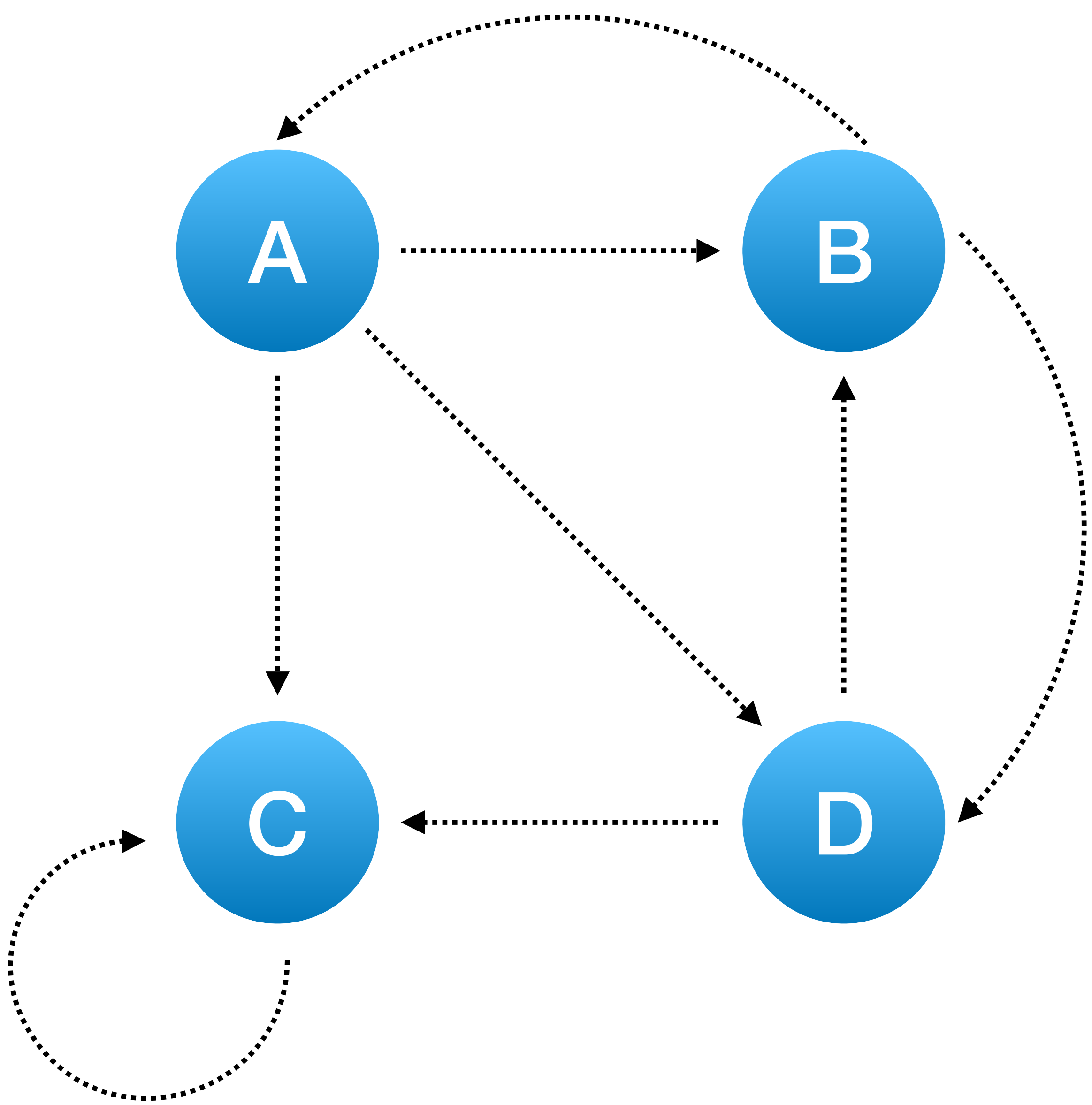


$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$



How we can avoid it

PageRank Teleporting

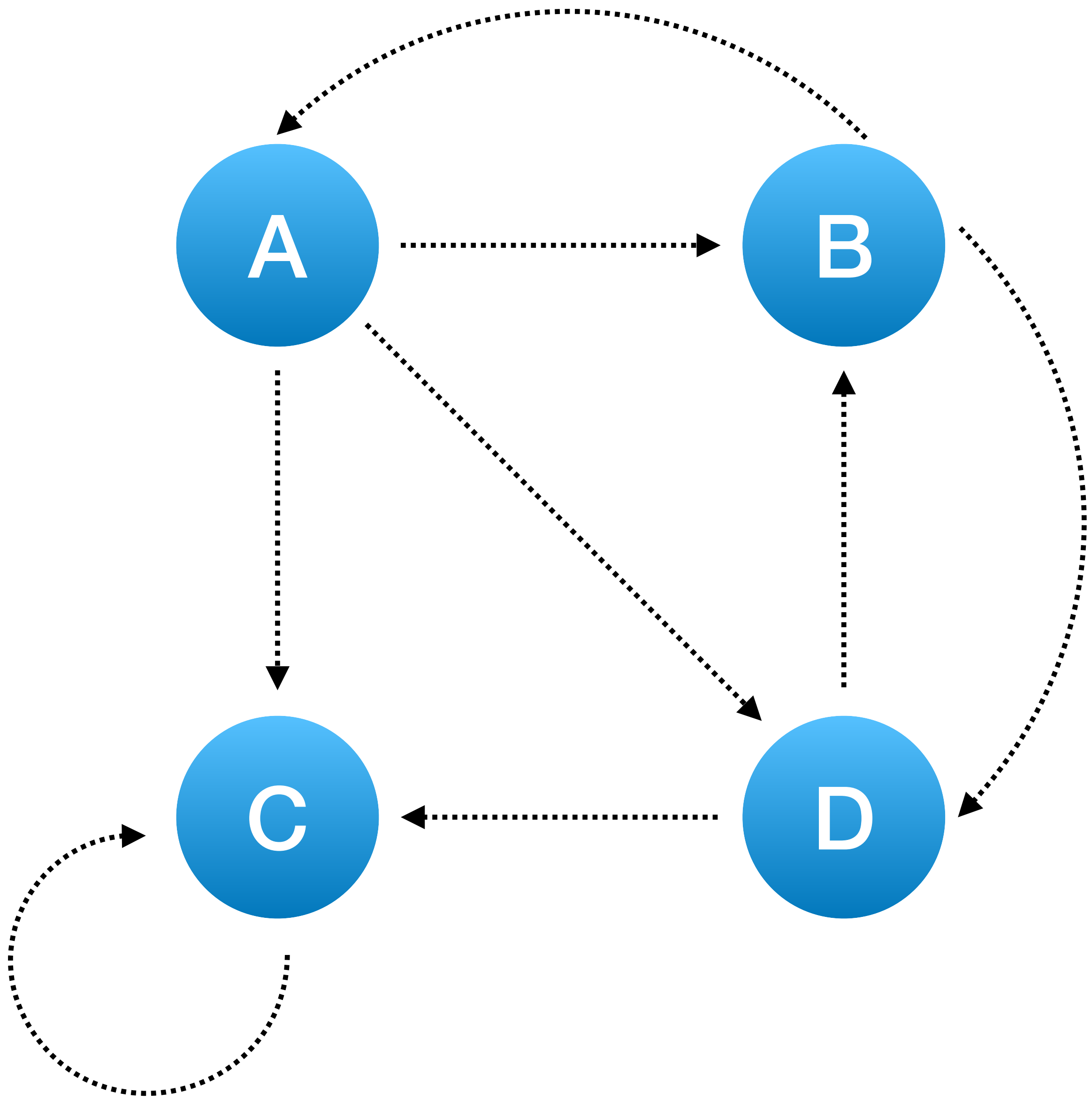


$$v' = \beta M v + \frac{(1-\beta)}{n} e$$

- $\beta \in [0,1]$

  $\beta = 0.85$
- e is a vector of all 1's
- n is a number of nodes
- M is a transition matrix
- v is a PR vector of iteration

Spider Trap

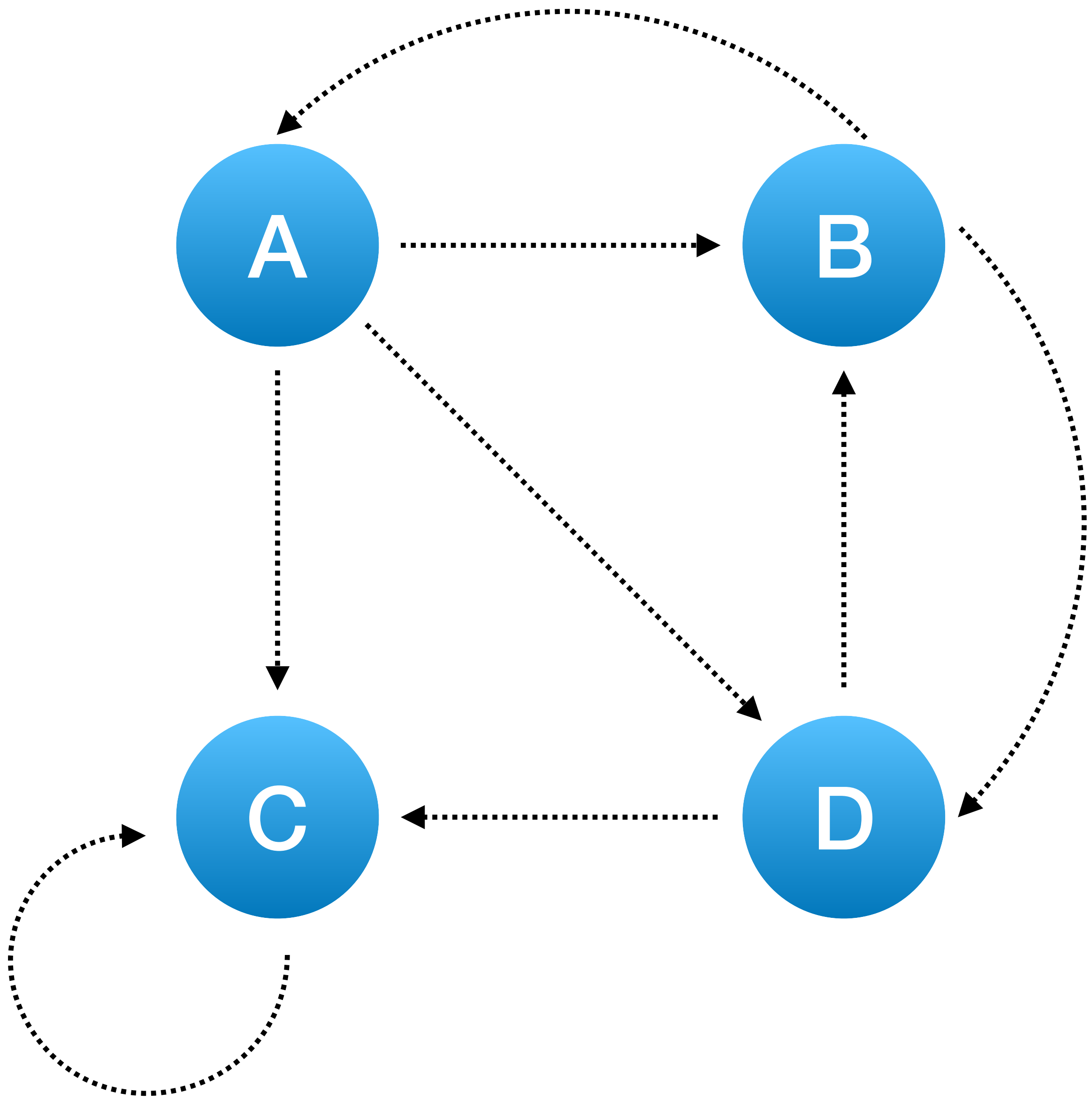


$$\beta = 0.8$$

$$v' = \begin{bmatrix} 0 & 2/5 & 0 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 4/5 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} \cdot v + \begin{bmatrix} 1/20 \\ 1/20 \\ 1/20 \\ 1/20 \end{bmatrix}$$

$$v' = \beta M v + \frac{(1-\beta)}{n} e$$

Spider Trap



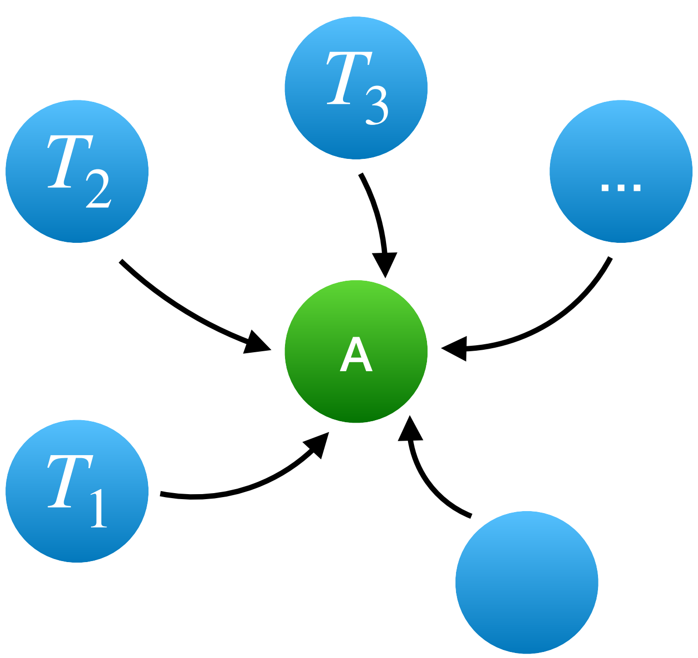
$$v_i = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 9/60 \\ 13/60 \\ 25/60 \\ 13/60 \end{bmatrix}, \begin{bmatrix} 41/300 \\ 53/300 \\ 153/300 \\ 53/300 \end{bmatrix}$$

$$\begin{bmatrix} 543/4500 \\ 707/4500 \\ 2543/4500 \\ 707/4500 \end{bmatrix}, \ldots, \begin{array}{cc} \begin{bmatrix} 15/148 \\ 19/148 \\ 95/148 \\ 19/148 \end{bmatrix} & \begin{matrix} A \\ B \\ C \\ D \end{matrix} \end{array}$$

$f_{PR}(C) > 50\ \%$ - the reason is in a spider trap

PageRank Algorithm

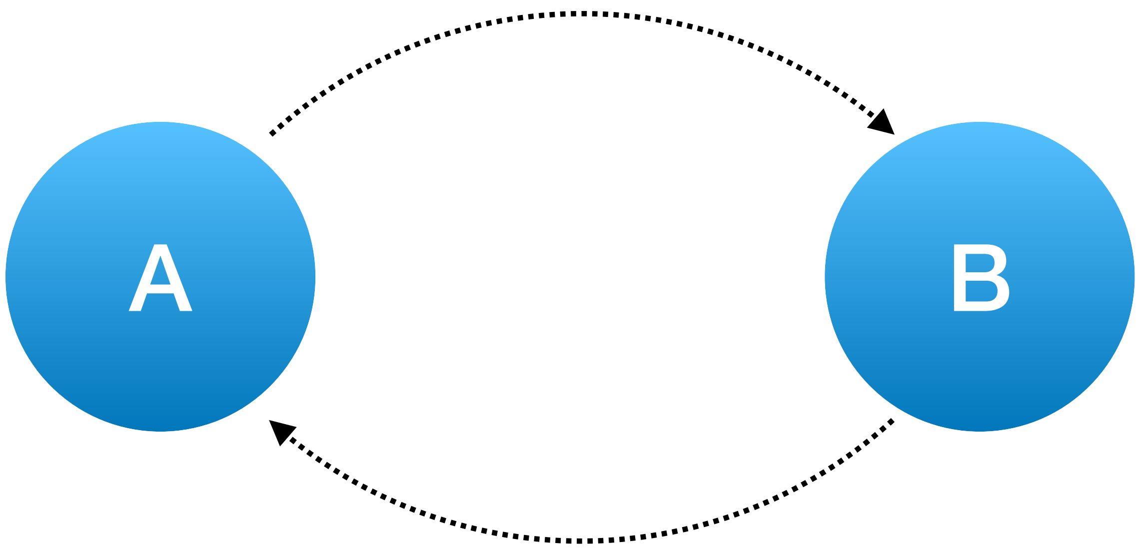
$$f_{PR}(A) = (1-d) + d \cdot \left(\frac{f_{PR}(T_1)}{C(T_1)} + \ldots + \frac{f_{PR}(T_n)}{C(T_n)} \right)$$

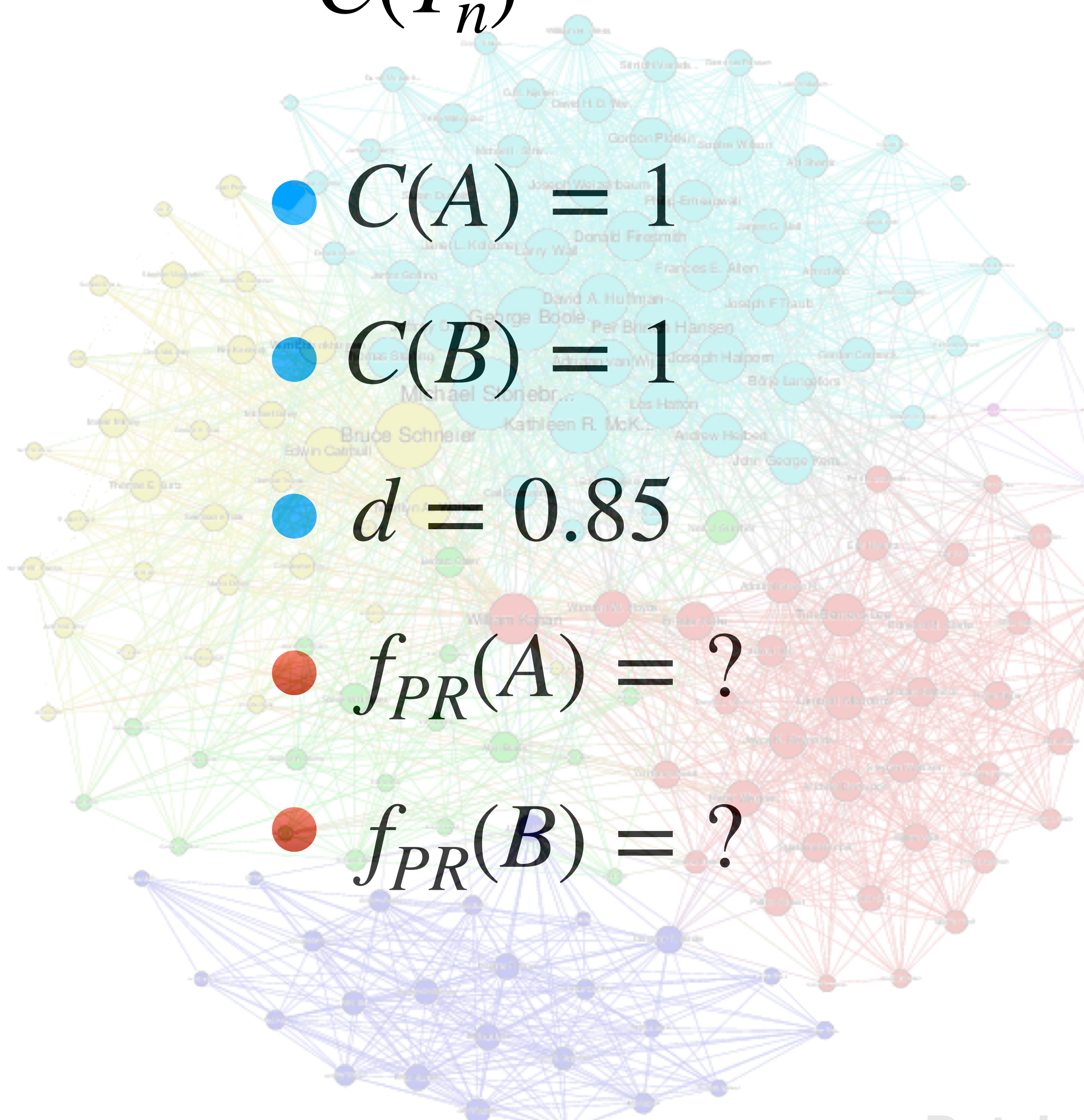


- $d \in [0,1]$ is a **dumping factor**
  $d \approx 0.85$
- $C(T_i)$ is a number of links going out of T_i

PageRank Sample

$$f_{PR}(A) = (1-d) + d \cdot \left(\frac{f_{PR}(T_1)}{C(T_1)} + \ldots + \frac{f_{PR}(T_n)}{C(T_n)} \right)$$

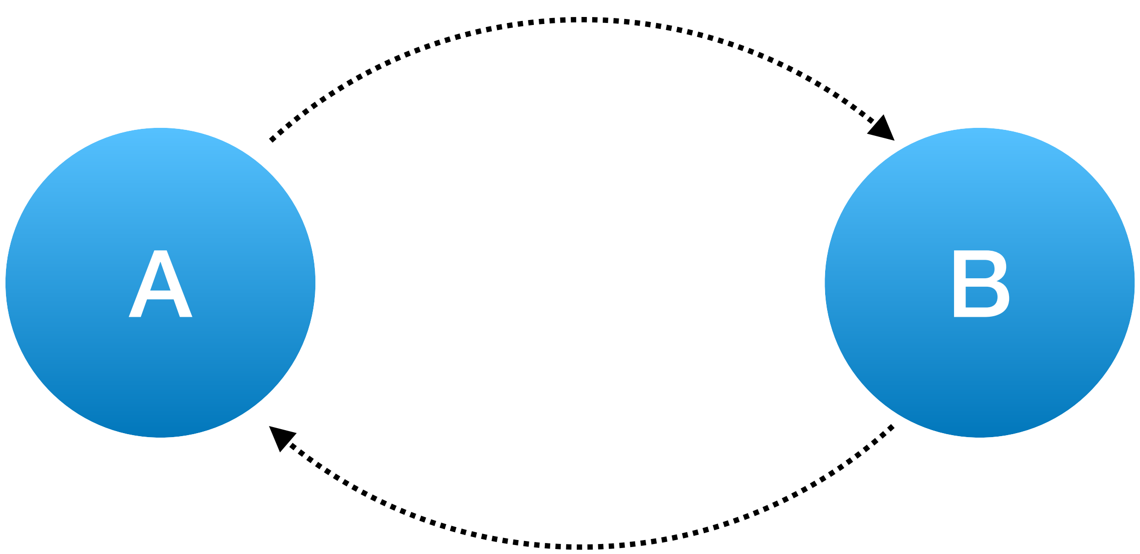




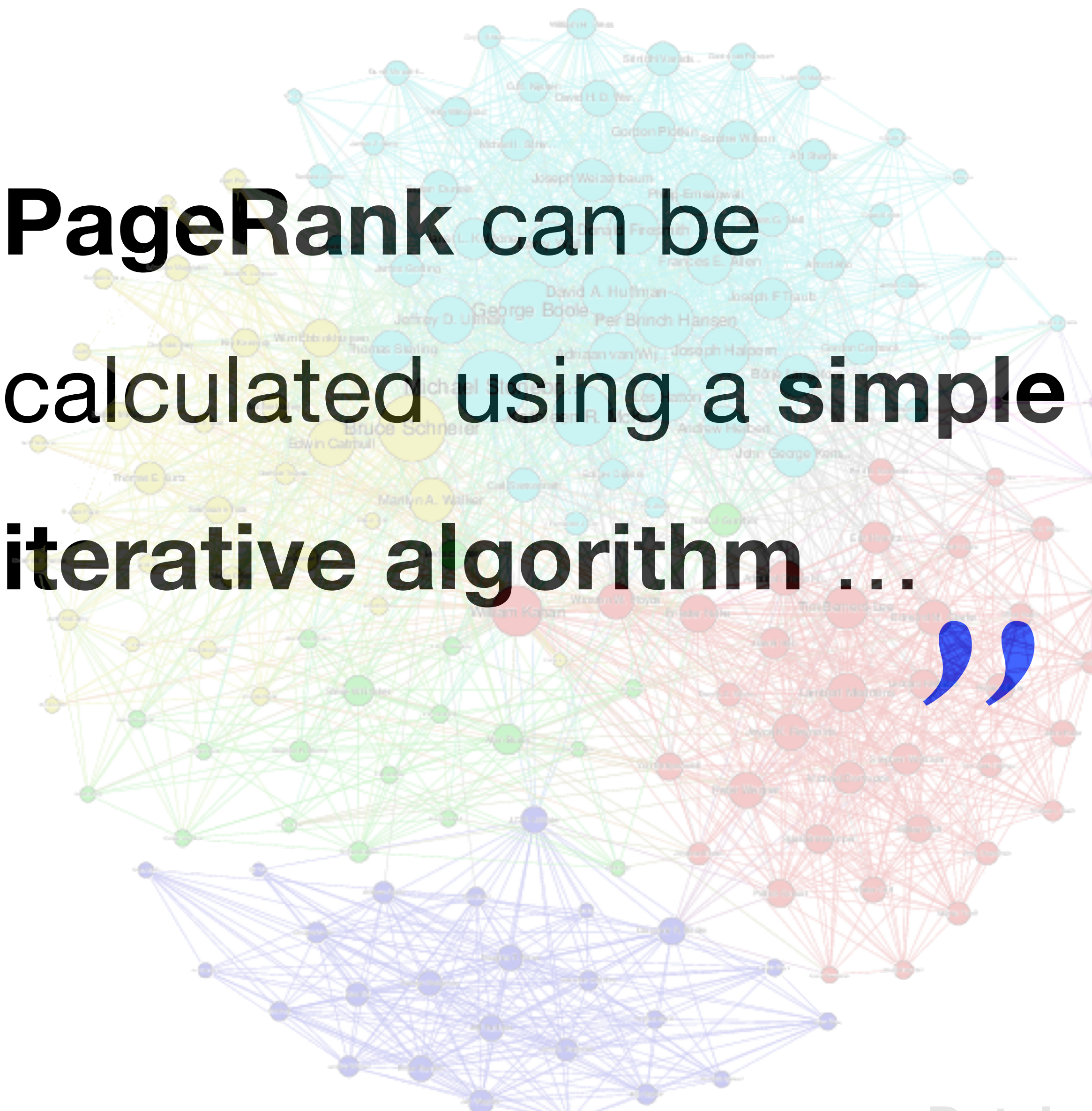
- $C(A) = 1$
- $C(B) = 1$
- $d = 0.85$
- $f_{PR}(A) = ?$
- $f_{PR}(B) = ?$

PageRank Sample

- $f_{PR}(A) = ?$
- $f_{PR}(B) = ?$



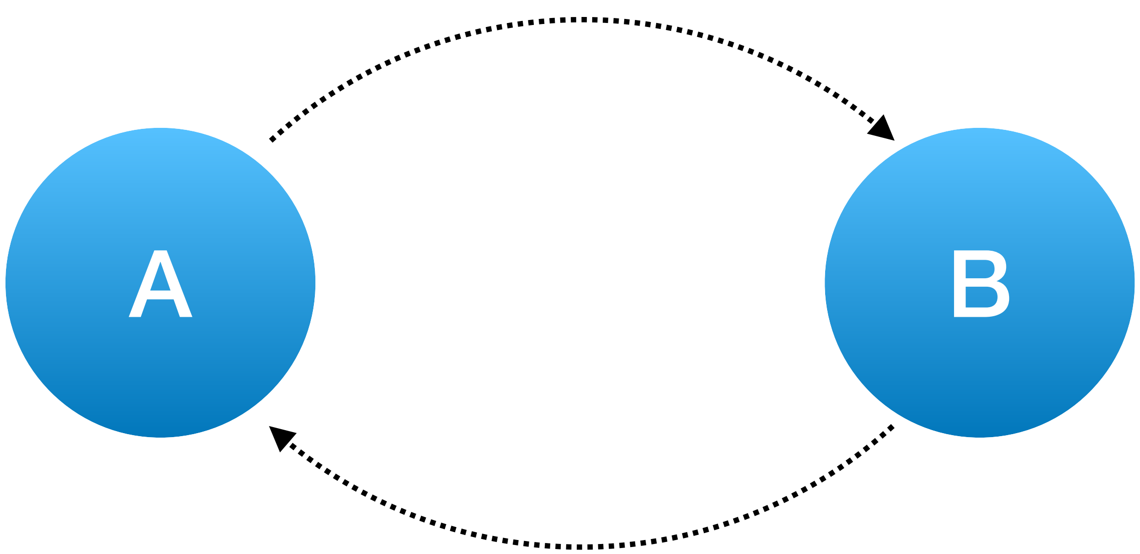
Google article

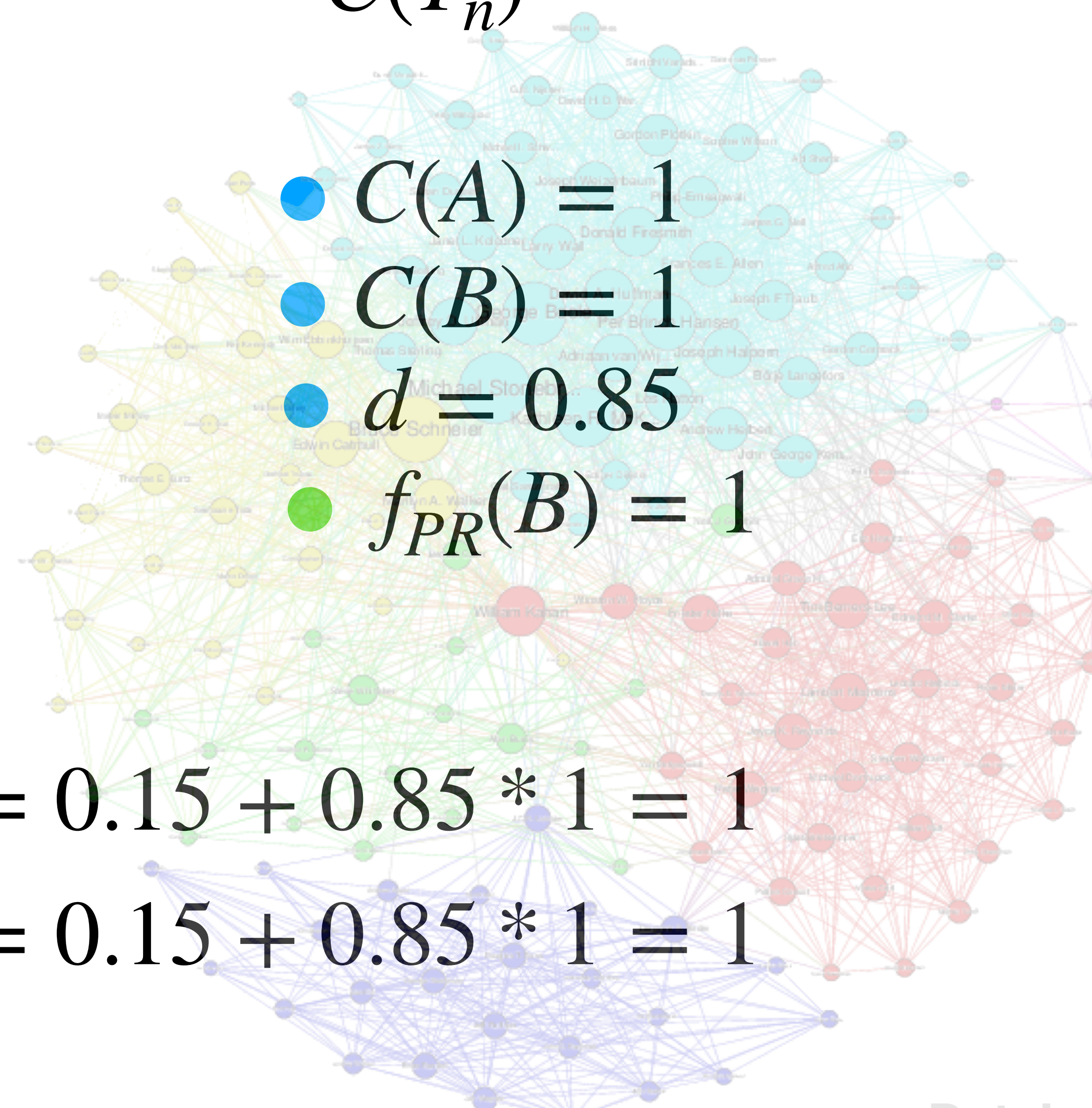


> “**PageRank** can be calculated using a **simple iterative algorithm** …”

PageRank Sample

$$f_{PR}(A) = (1 - d) + d \cdot \left(\frac{f_{PR}(T_1)}{C(T_1)} + \ldots + \frac{f_{PR}(T_n)}{C(T_n)} \right)$$





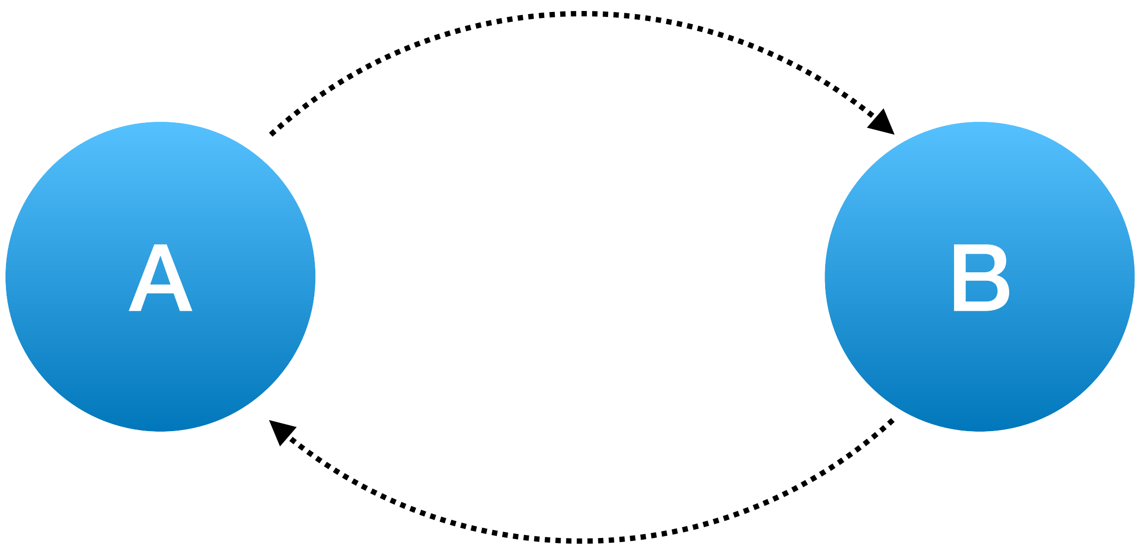
- $C(A) = 1$
- $C(B) = 1$
- $d = 0.85$
- $f_{PR}(B) = 1$

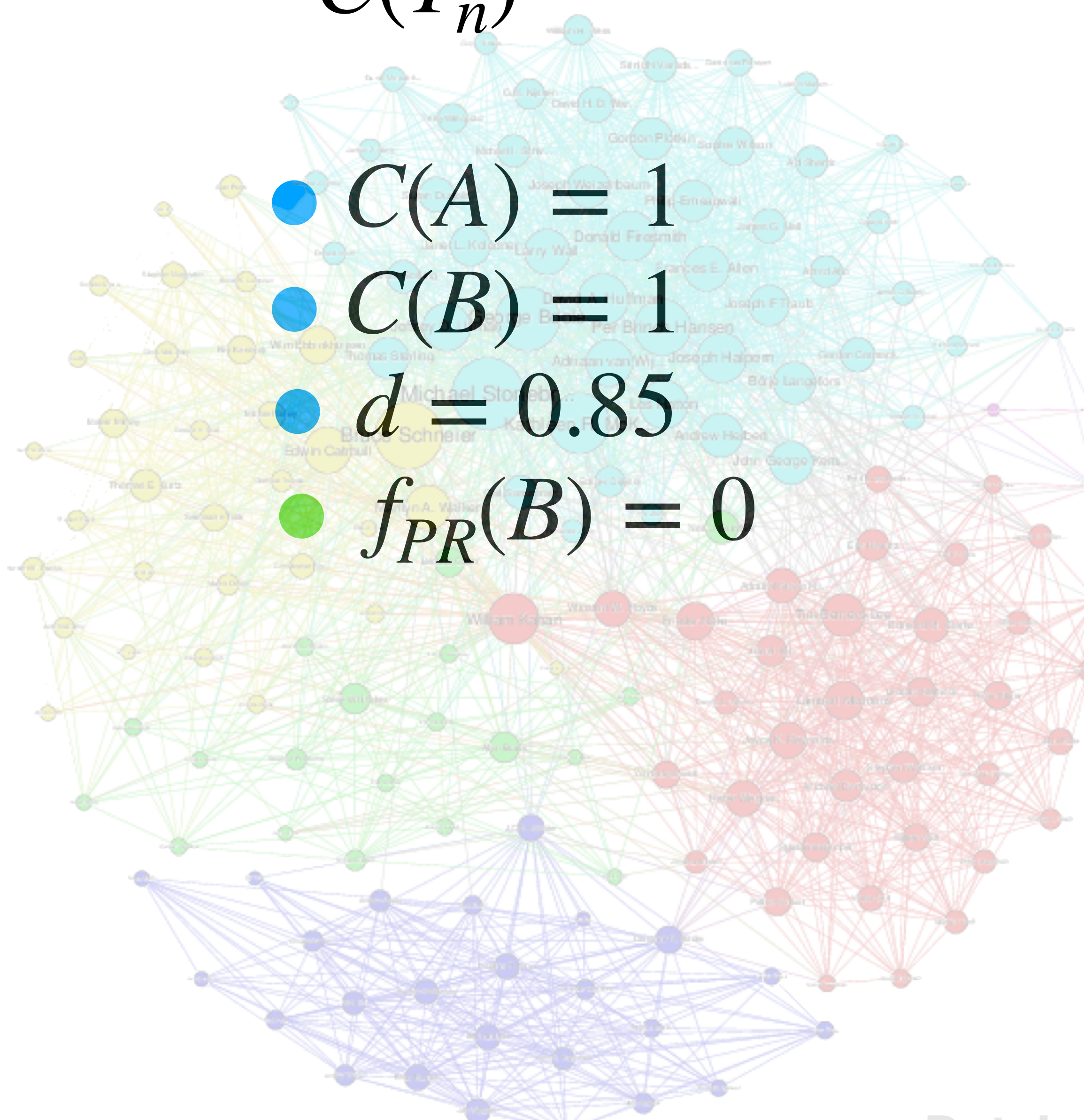
$$f_{PR}(A) = 0.15 + 0.85 * 1 = 1$$

$$f_{PR}(B) = 0.15 + 0.85 * 1 = 1$$

PageRank Sample

$$f_{PR}(A) = (1 - d) + d \cdot \left(\frac{f_{PR}(T_1)}{C(T_1)} + \ldots + \frac{f_{PR}(T_n)}{C(T_n)} \right)$$





- $C(A) = 1$
- $C(B) = 1$
- $d = 0.85$
- $f_{PR}(B) = 0$

PageRank Sample

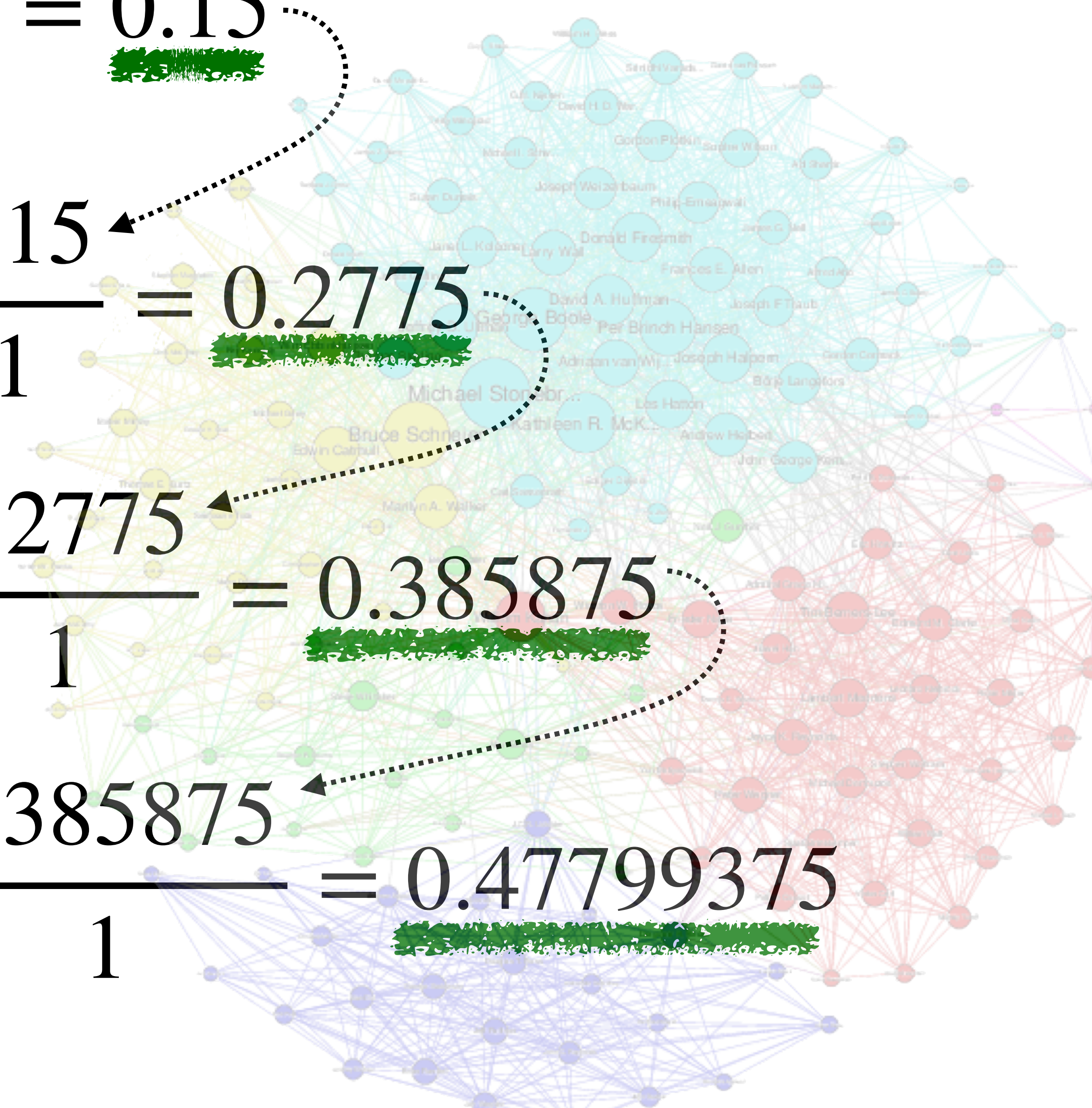
- $f_{PR}(B) = 0$

I
- $f_{PR}(A) = 0.15 + 0.85 * \frac{0}{1} = 0.15$

$$f_{PR}(B) = 0.15 + 0.85 * \frac{0.15}{1} = 0.2775$$

II
- $f_{PR}(A) = 0.15 + 0.85 * \frac{0.2775}{1} = 0.385875$

$$f_{PR}(B) = 0.15 + 0.85 * \frac{0.385875}{1} = 0.47799375$$



PageRank Sample

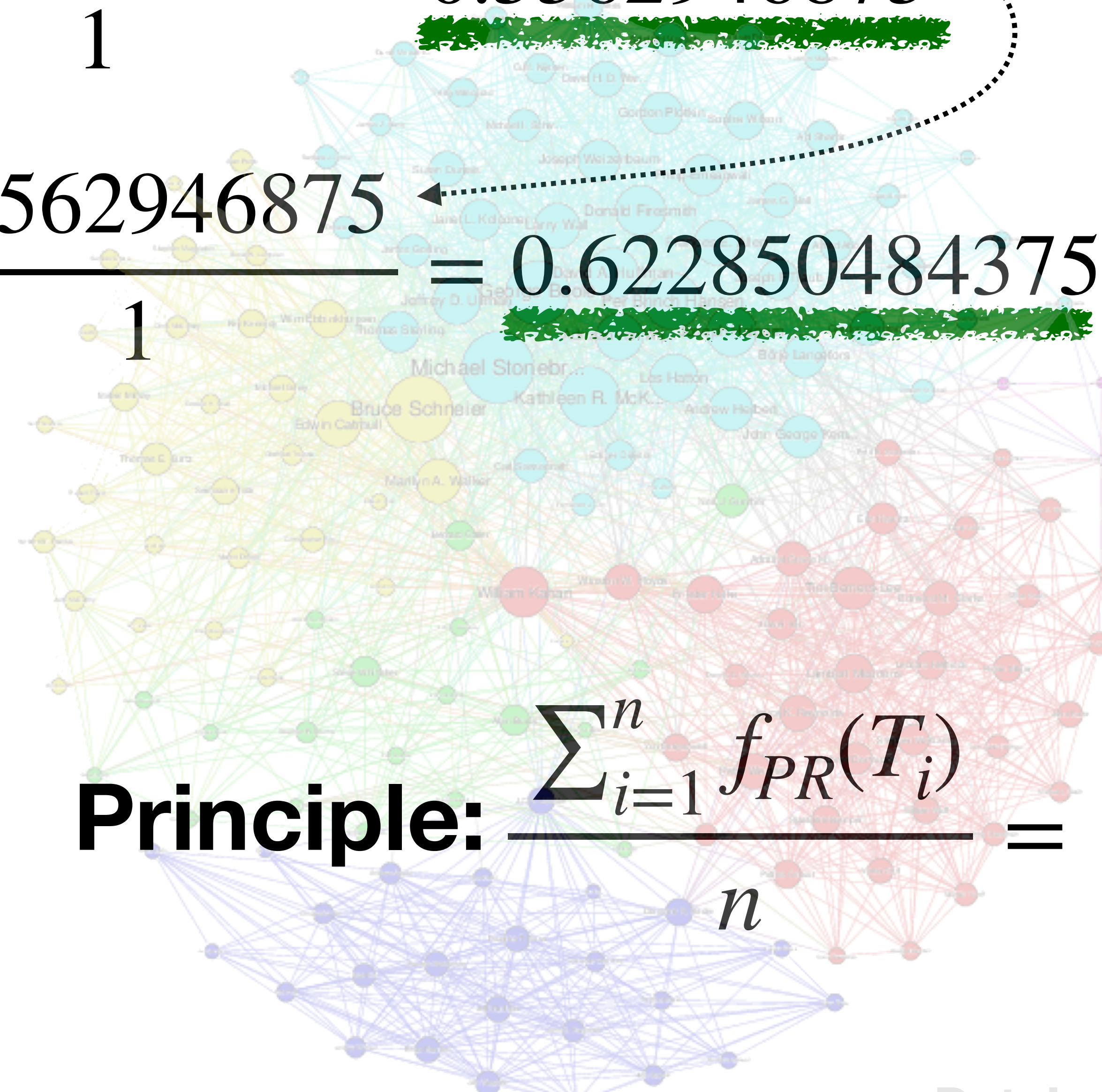
III • $f_{PR}(A) = 0.15 + 0.85 * \dfrac{0.47799375}{1} = \underline{0.5562946875}$

$f_{PR}(B) = 0.15 + 0.85 * \dfrac{0.5562946875}{1} = \underline{0.622850484375}$

IV • …

$f_{PR}(A) =$

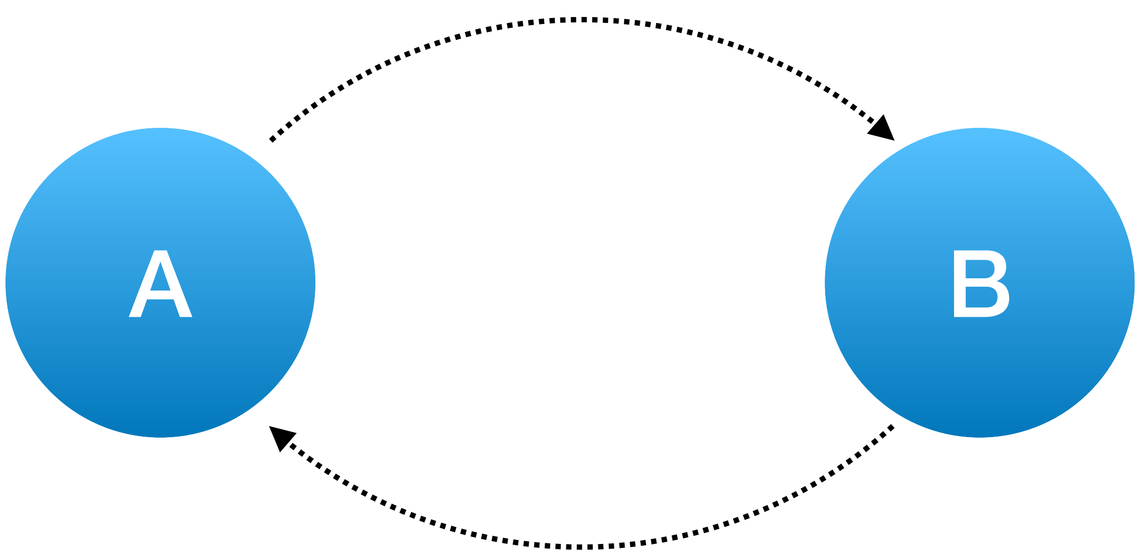
$f_{PR}(B) =$

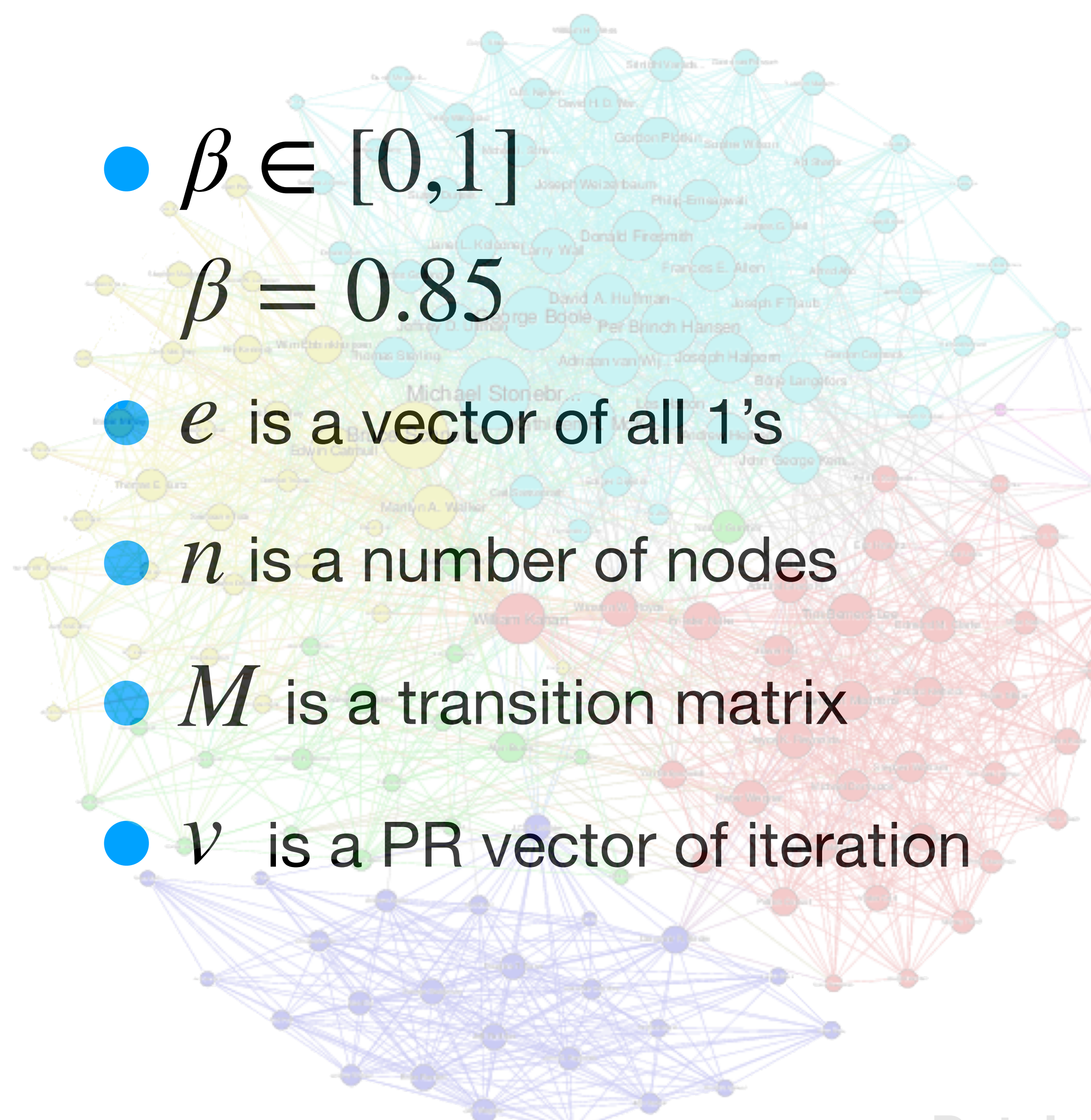


Principle: $\dfrac{\sum_{i=1}^{n} f_{PR}(T_i)}{n} = 1$

PageRank Matrix Sample

$$v' = \beta M v + \frac{(1-\beta)}{n} e$$



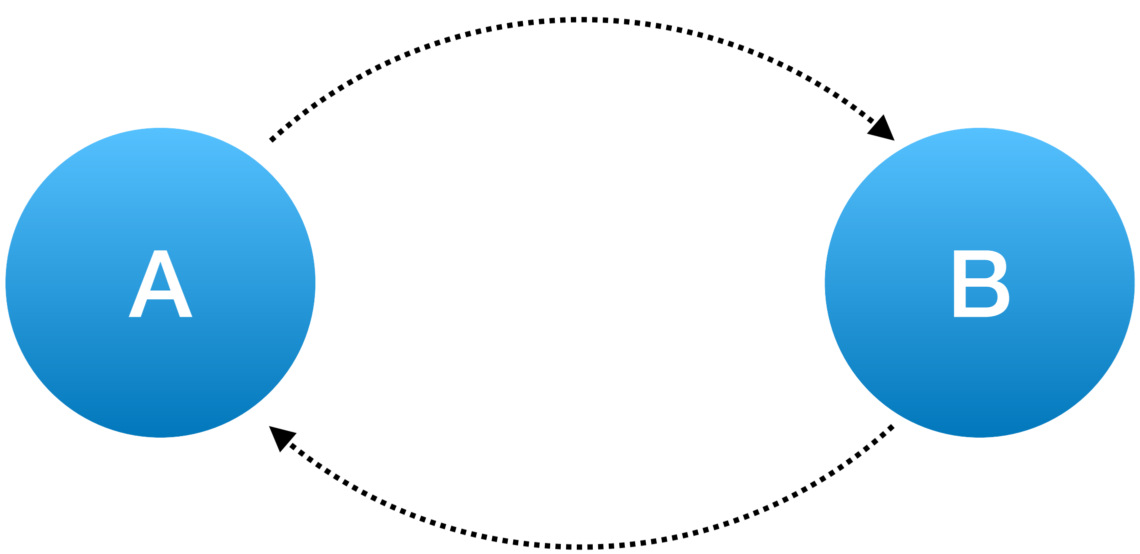


- $\beta \in [0,1]$

  $\beta = 0.85$
- e is a vector of all 1's
- n is a number of nodes
- M is a transition matrix
- v is a PR vector of iteration

PageRank Matrix Sample

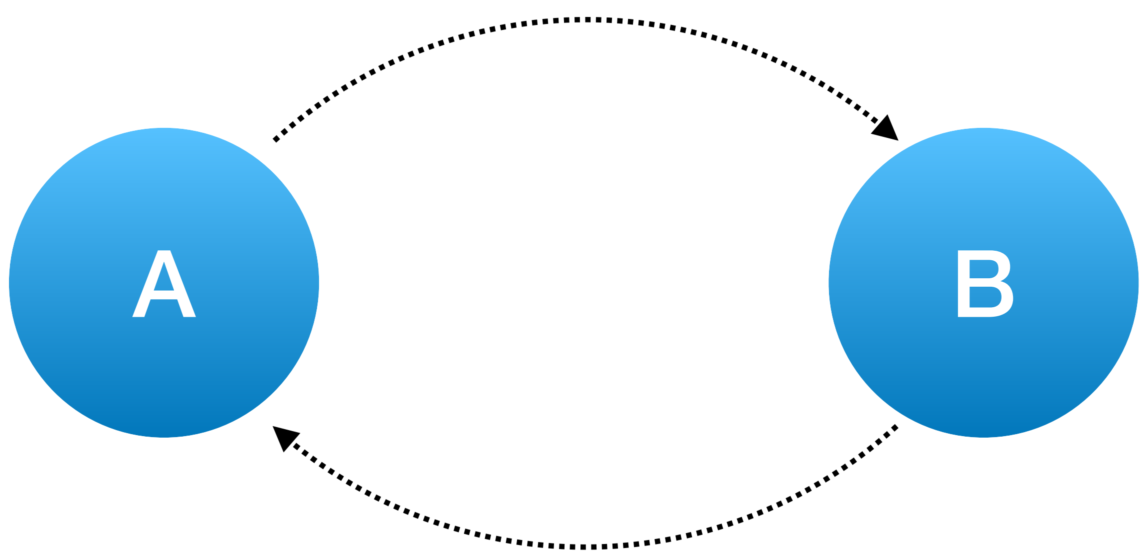
$$v' = \beta M v + \frac{(1-\beta)}{n} e$$

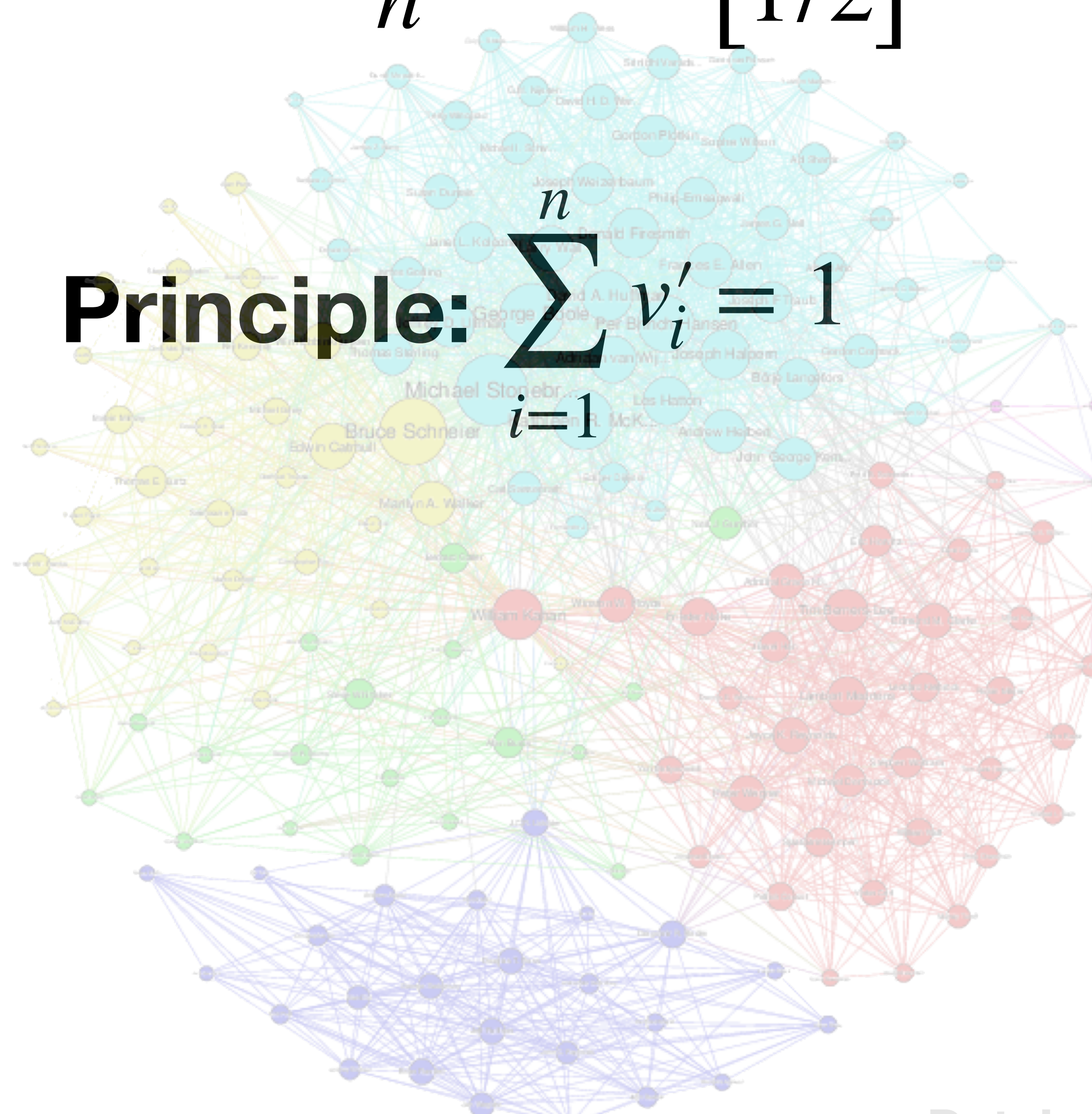


- $M = \begin{array}{c} \begin{array}{cc} \text{A} & \text{B} \end{array} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array}$
- $\beta = 0.85$
- $n = 2$
- $e = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
- $v = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$

PageRank Matrix Sample

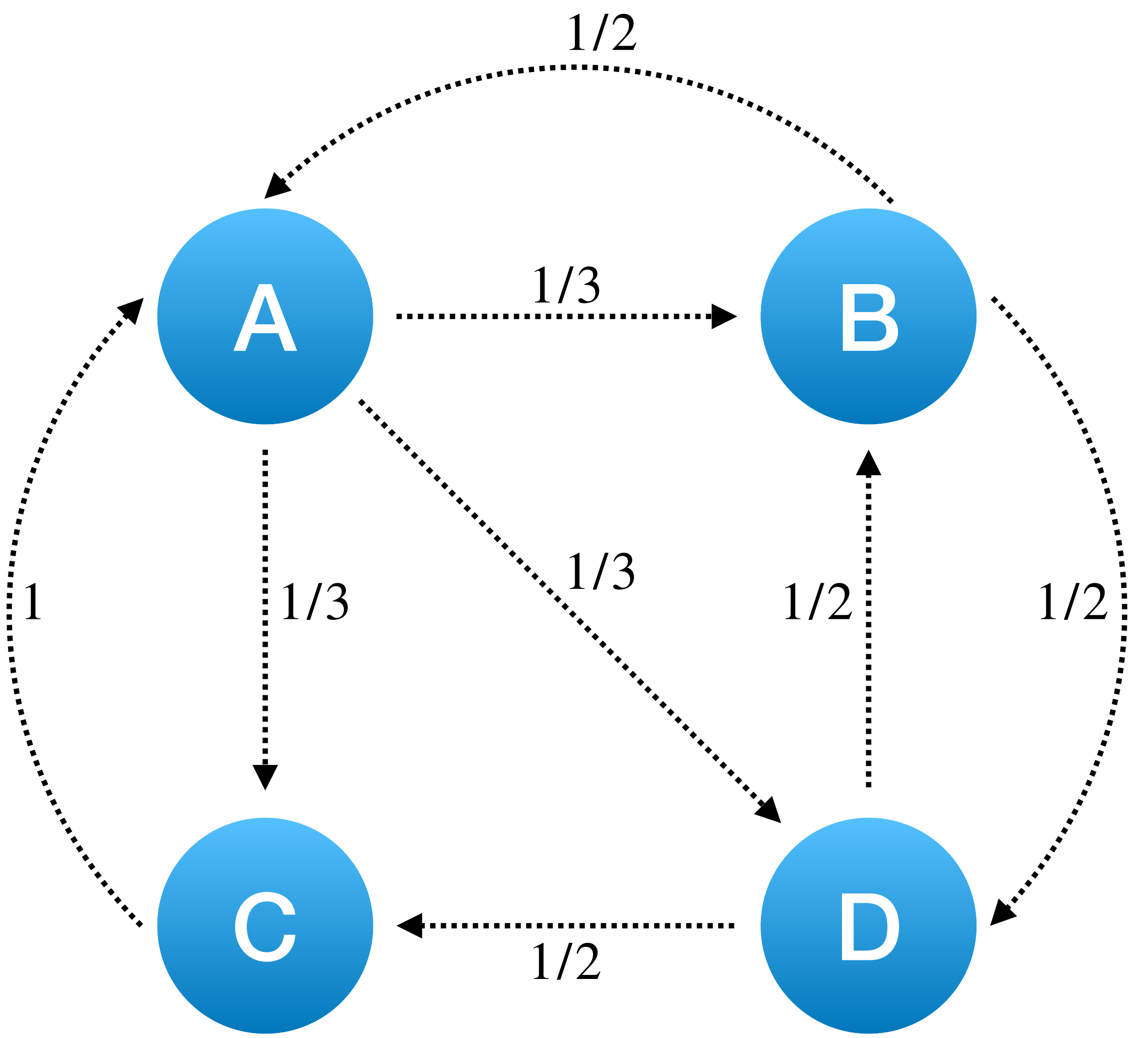
$$v' = \beta M v + \frac{(1-\beta)}{n} e = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$





Principle: $\sum_{i=1}^{n} v'_i = 1$

PageRank Matrix Sample



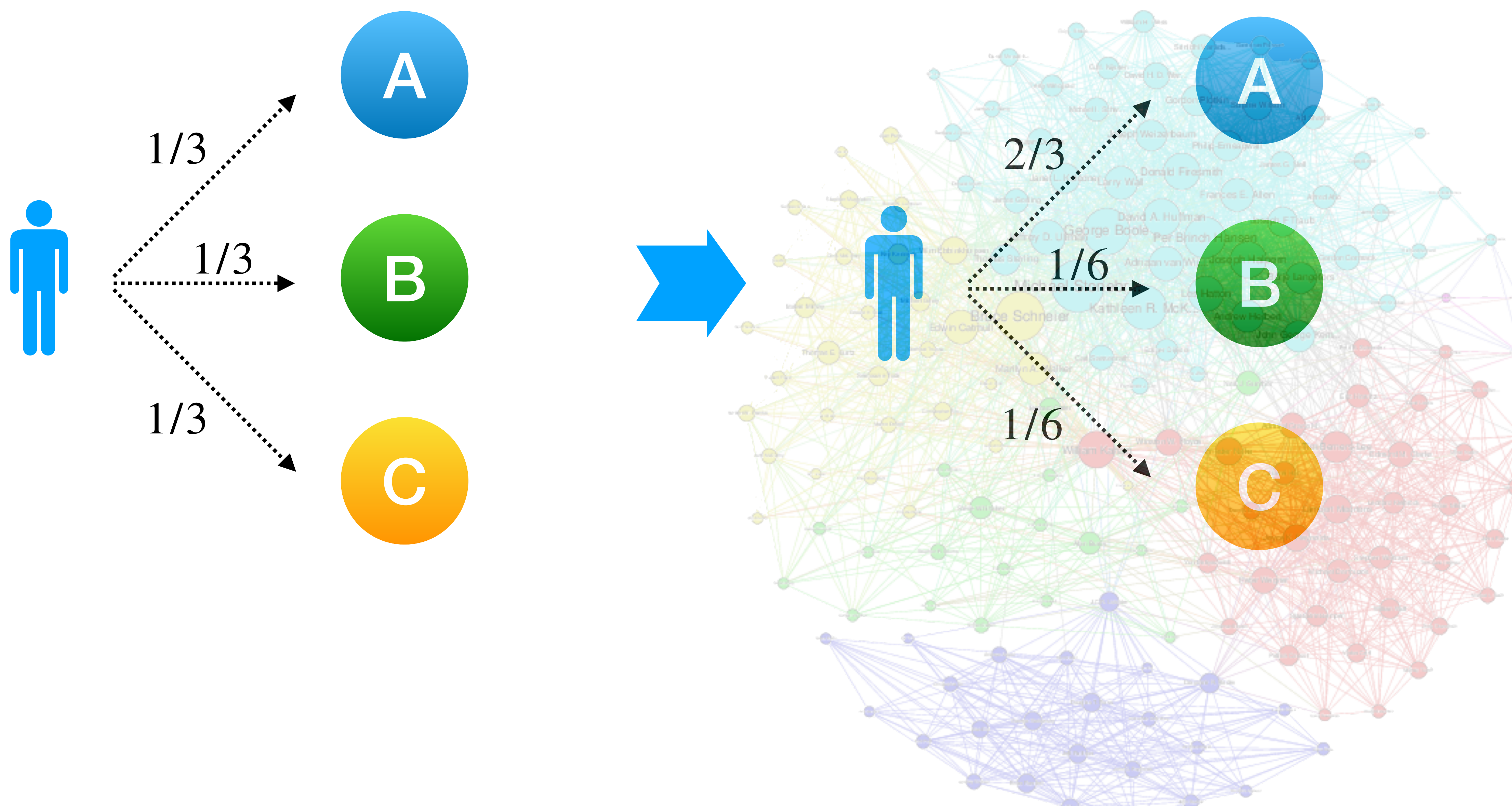
$$v' = \beta M v + \frac{(1-\beta)}{n} e$$

$$v_i = \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}, \begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix}, \ldots, \begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix}$$

$$v' = \beta M v + \frac{(1-\beta)}{n} e \overset{?}{=} M^n \cdot v_0$$

Why are equals

Topic-Sensitive Page Rank



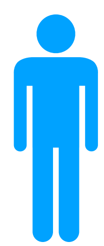
Topic-Sensitive Page Rank

I am “googling” - **jaguar**

Topic-Sensitive Page Rank

I am “googling” - **jaguar**

private Page Rank



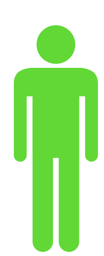
$$\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$$

private PageRank

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}$$

private PageRank

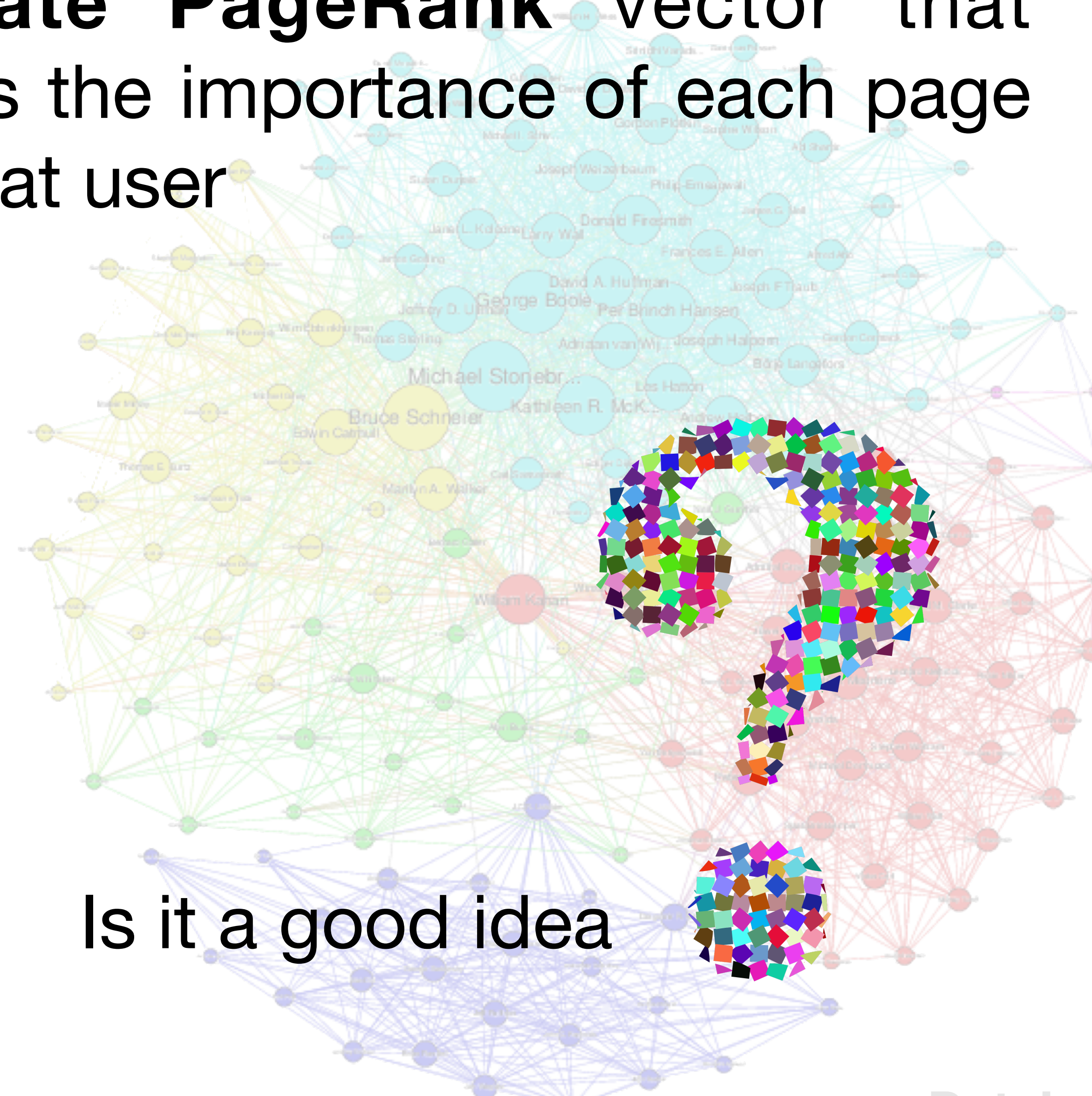
...



$$\begin{bmatrix} z_1 \\ z_2 \\ \dots \\ z_n \end{bmatrix}$$

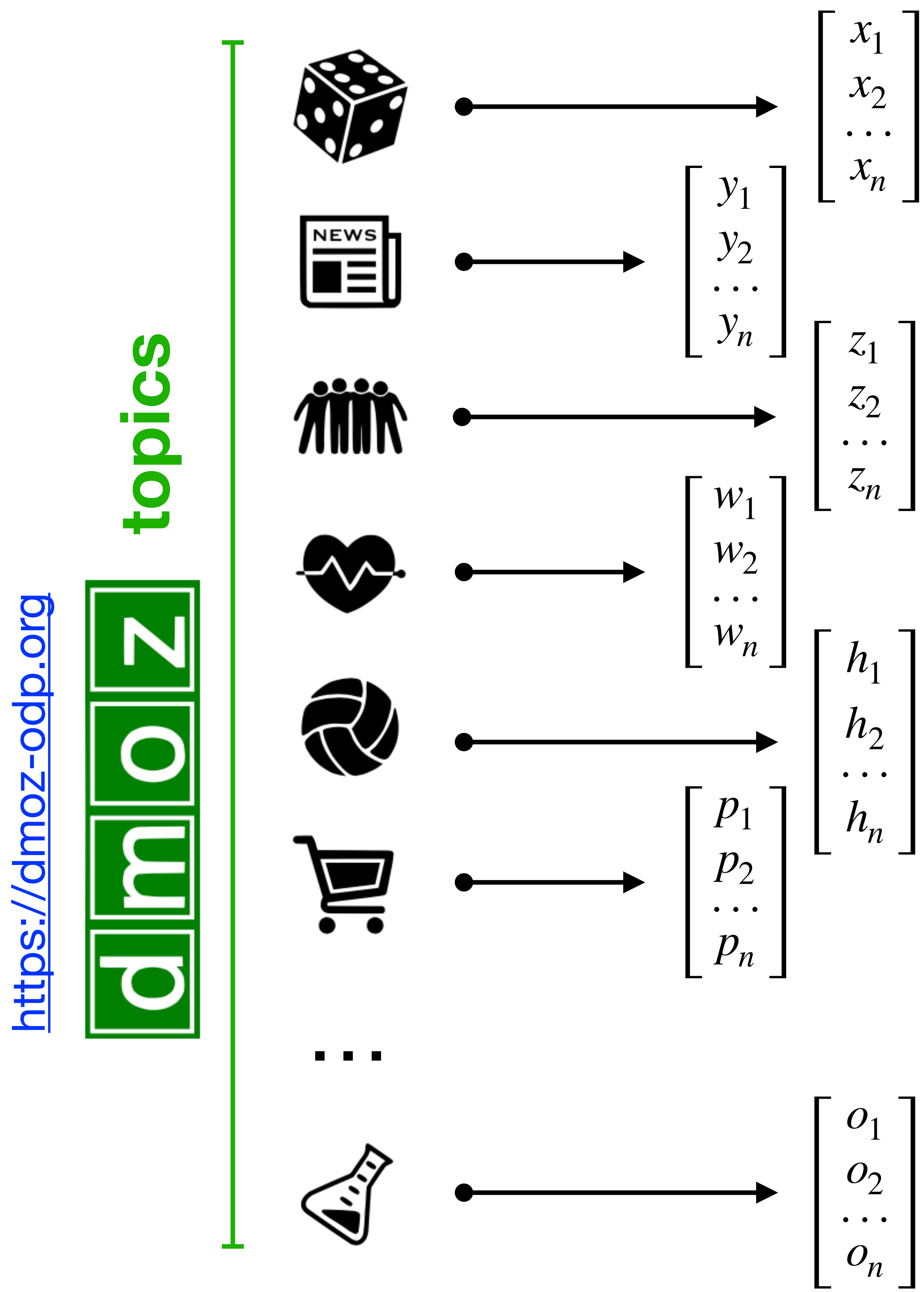
private PageRank

Ideally, each user would have a **private PageRank** vector that gives the importance of each page to that user

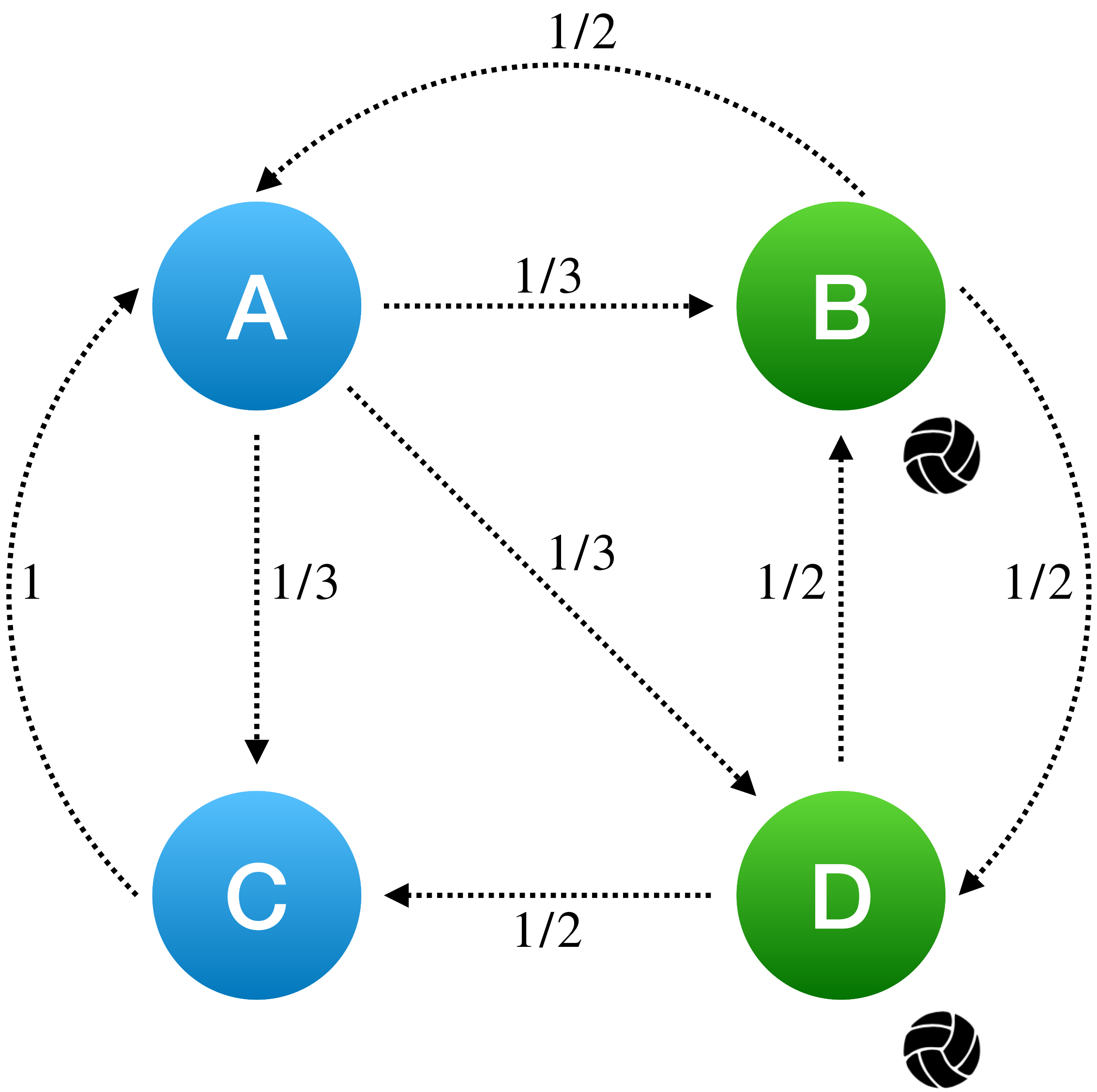


Is it a good idea

Topic-Sensitive Page Rank



Biased Random Walks

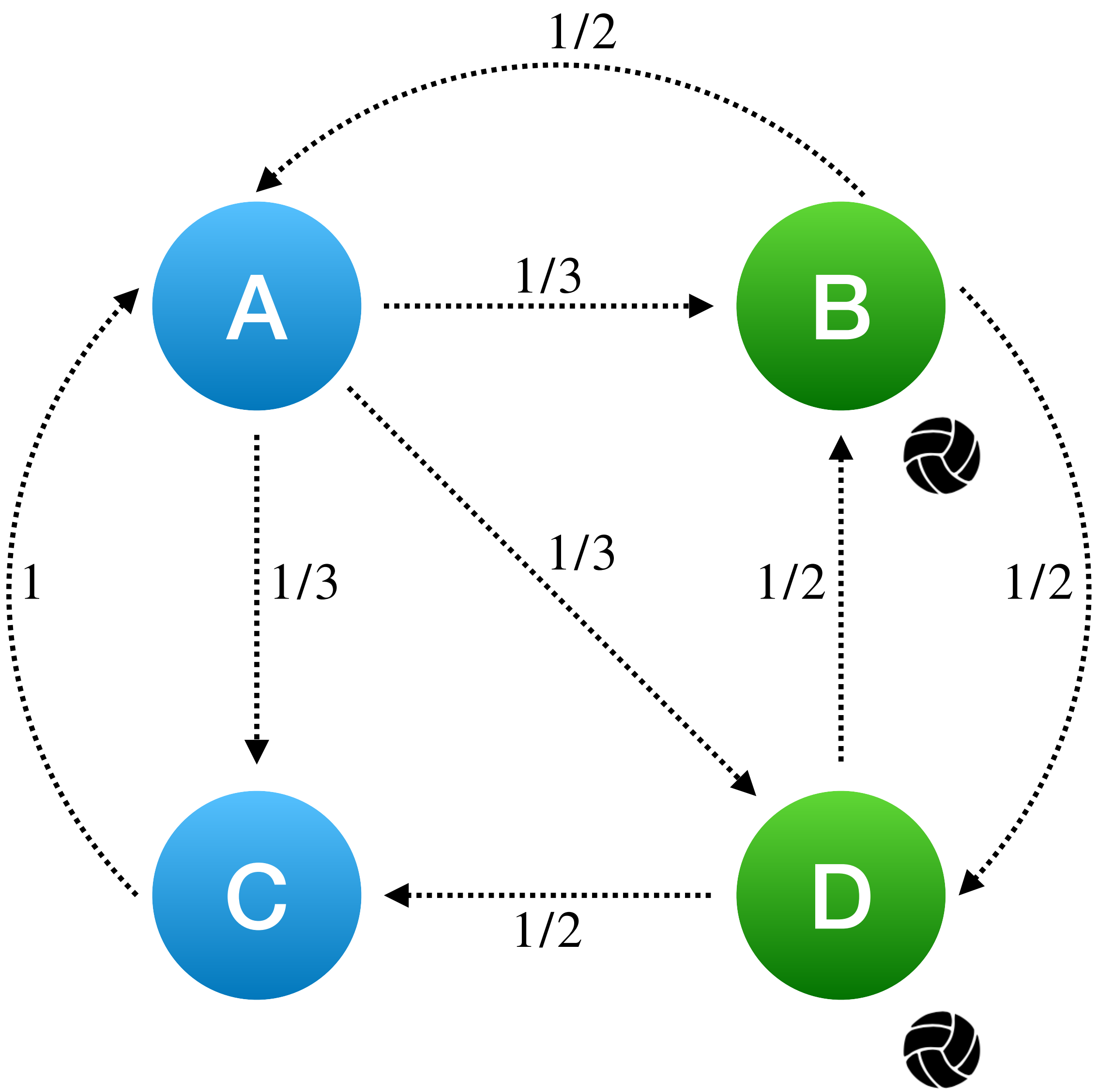


$B, D \to$ ***sport*** topic

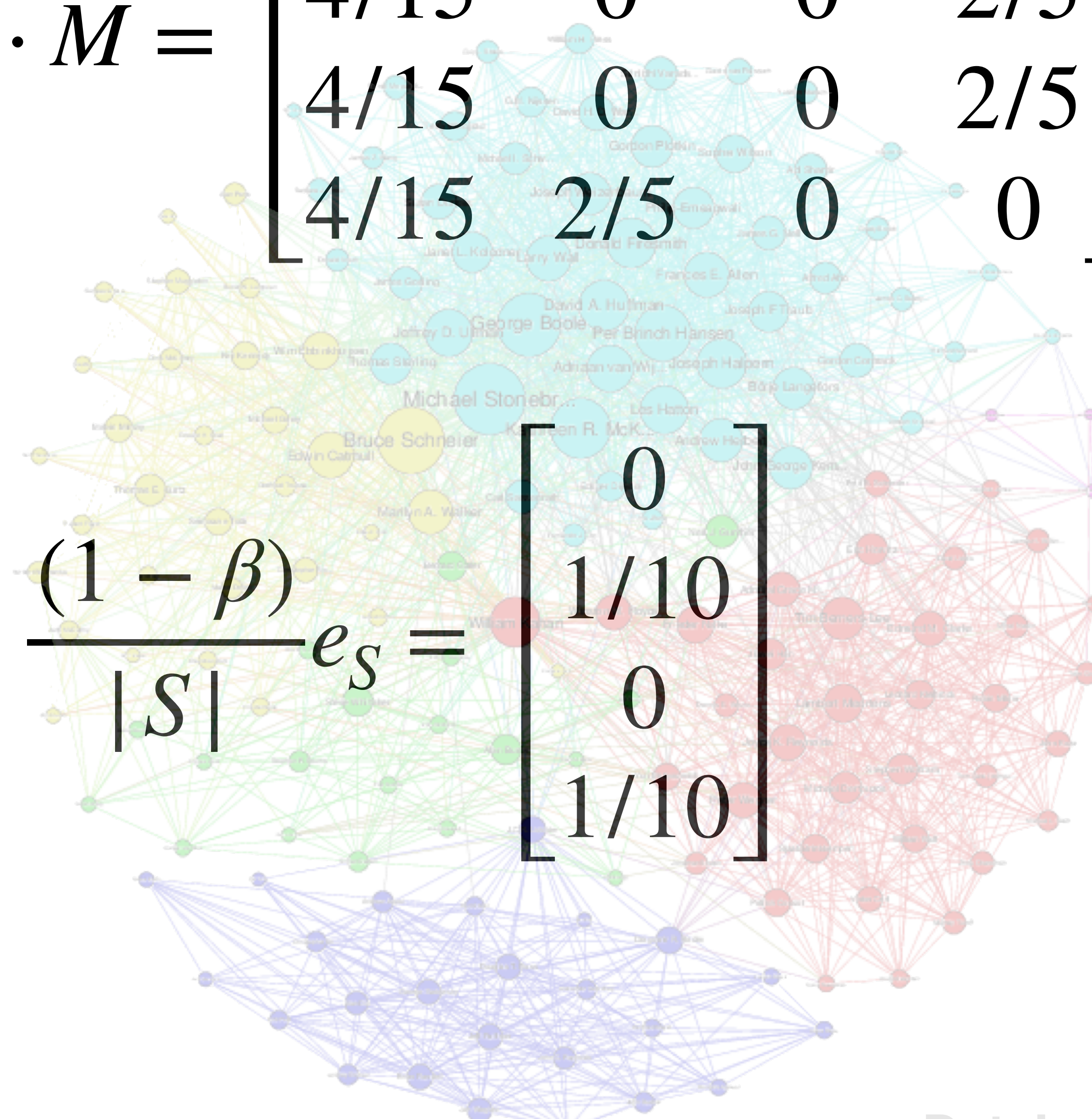
$$v' = \beta M v + \frac{(1-\beta)}{|S|} e_S$$

- $S = \{B, D\}$
- $|S| = 2$
- $e_S = \begin{array}{cc} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} & \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \end{array} \end{array}$

Biased Random Walks

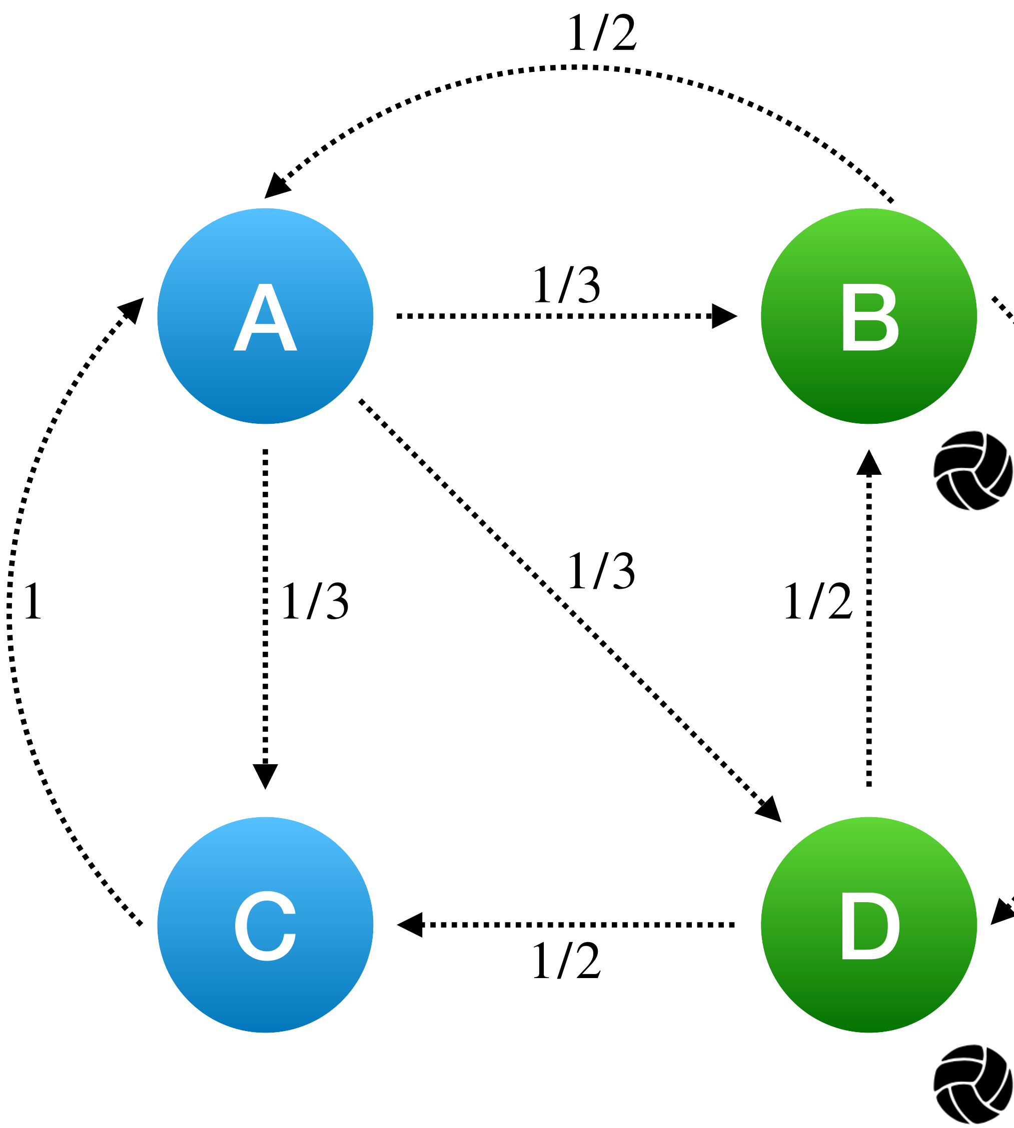


$$\beta \cdot M = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix}$$



$$\frac{(1-\beta)}{|S|} e_S = \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

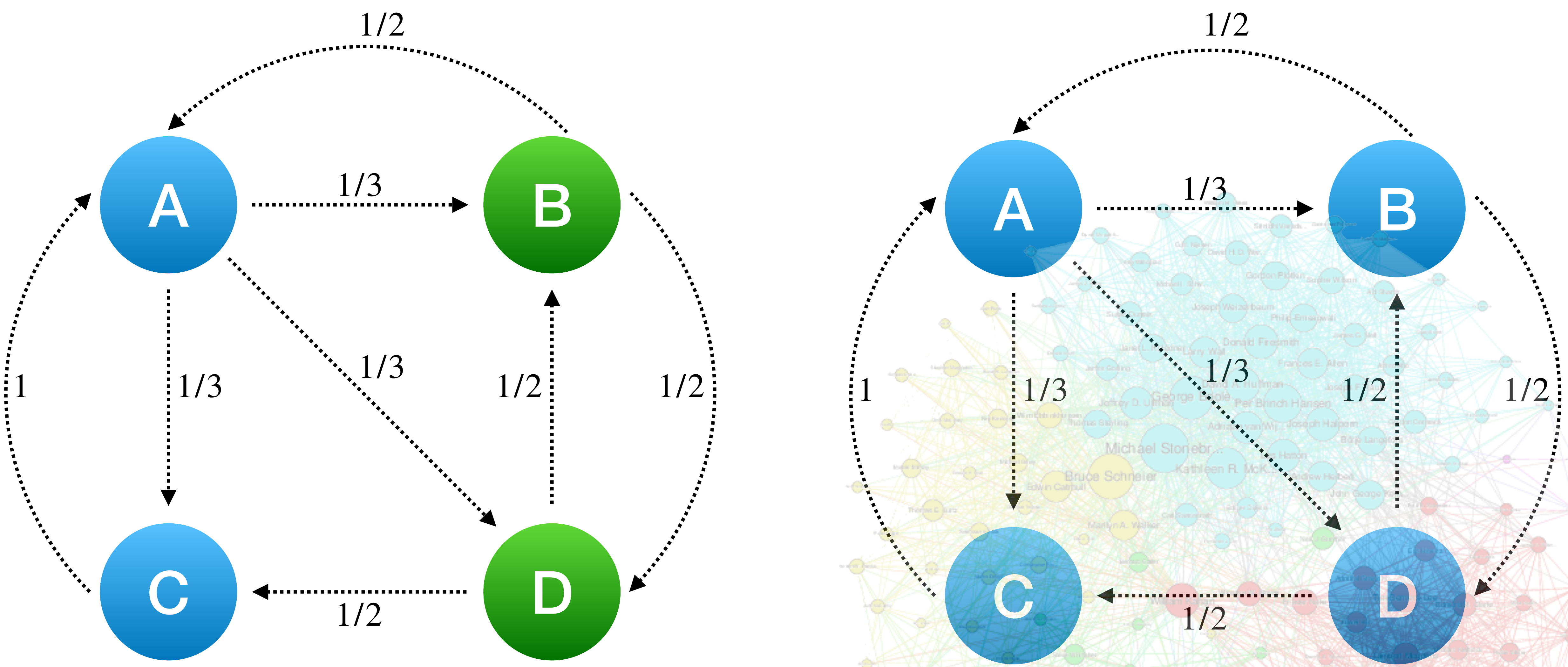
Biased Random Walks



$$v' = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} v + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

$$\begin{array}{c} \color{blue}{A} \\ \color{green}{B} \\ \color{blue}{C} \\ \color{green}{D} \end{array} \begin{bmatrix} 0/2 \\ 1/2 \\ 0/2 \\ 1/2 \end{bmatrix}, \begin{bmatrix} 2/10 \\ 3/10 \\ 2/10 \\ 3/10 \end{bmatrix}, \begin{bmatrix} 42/150 \\ 41/150 \\ 26/150 \\ 41/150 \end{bmatrix}, \ldots, \begin{bmatrix} 54/210 \\ 59/210 \\ 38/210 \\ 59/210 \end{bmatrix}$$

Biased Random Walks



$$\begin{array}{c} A \\ B \\ C \\ D \end{array} \begin{bmatrix} 54/210 \\ 59/210 \\ 38/210 \\ 59/210 \end{bmatrix} \qquad \begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix} \begin{array}{c} A \\ B \\ C \\ D \end{array}$$

$$59/210 > 2/9$$

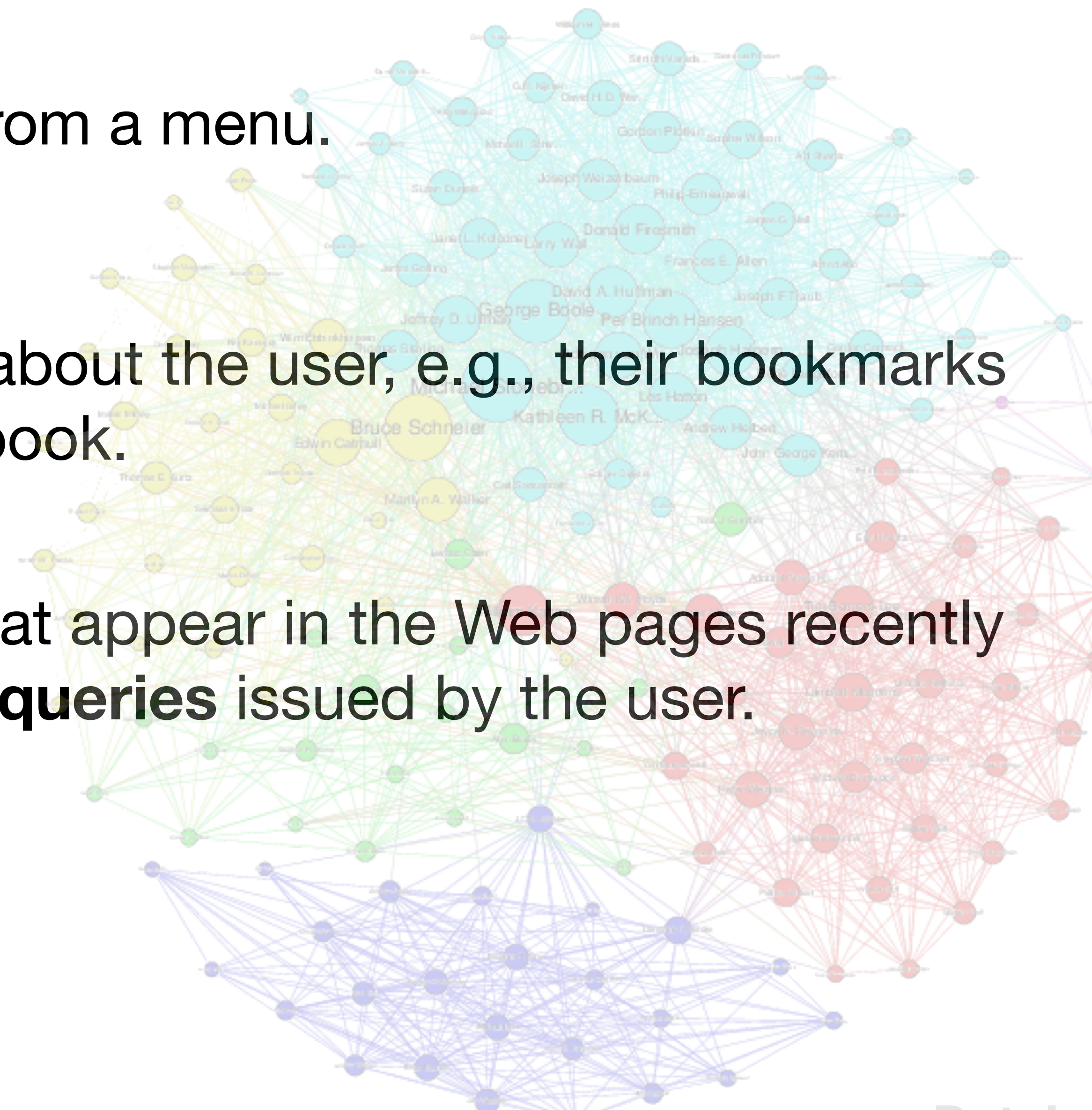
How to use

- Decide on the **topics** for which we shall create specialized PageRank vectors.
- Pick a **teleport set** for each of these topics, and use that set to compute the ***topic-sensitive PageRank*** vector for that topic.
- Find a way of determining the topic or set of topics that are most **relevant** for a particular search query.
- Use the PageRank vectors for that topic or topics in the ordering of the responses to the search query.

How to use

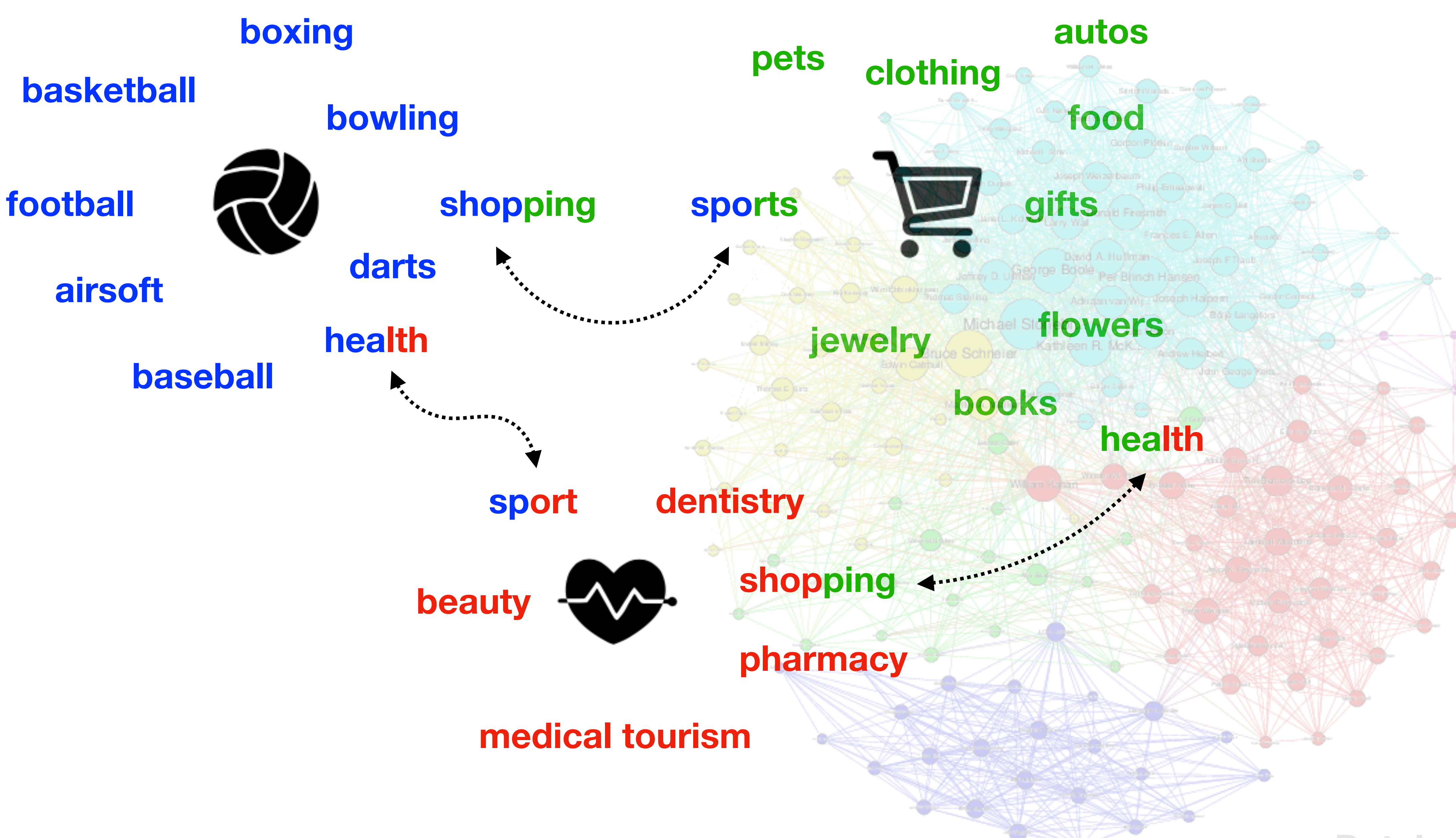
- Find a way of determining the topic or set of topics that are most **relevant** for a particular search query.

  - Allow the user to select a topic from a menu.
  - Mine the topic(s) by information about the user, e.g., their bookmarks or their stated interests on Facebook.
  - Mine the topic(s) by the words that appear in the Web pages recently searched by the **user**, or **recent queries** issued by the user.



Mining Topics

dmoz



Mining Topics

$$S_1 = \{word_1, word_2, \ldots, word_k\}$$

The Data Laboratory

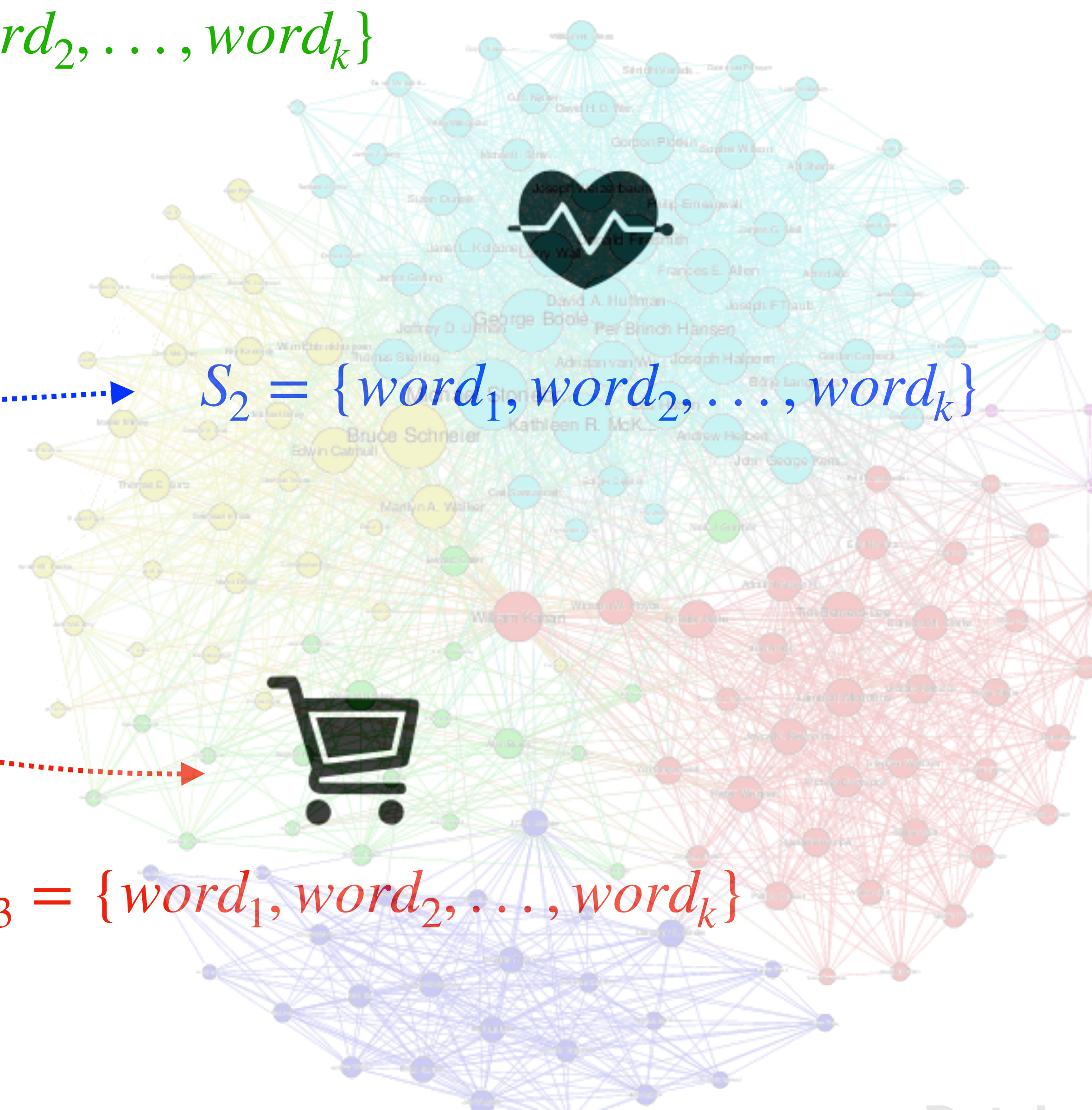
Hi there!

We are highly "SCI-IT-motivated" students from Kazan Fe

We are here to understand a real world by different aspec

> **"Data is the new oil"**
>
> *Clive Humby*

$$P = \{word_1, word_2, \ldots, word_k\}$$



$$S_2 = \{word_1, word_2, \ldots, word_k\}$$

$$S_3 = \{word_1, word_2, \ldots, word_k\}$$

Mining Topics

$$J(P,S) = \frac{|P \cap S|}{|P \cup S|}$$

$$K(P,S) = \frac{|P \cap S|^2}{|P| \cdot |S|}$$

The Data Laboratory

Hi there!

We are highly "SCI-IT-motivated" students from Kazan F

We are here to understand a real world by different aspec

> **"Data is the new oil"**
>
> *Clive Humby*

$$P = \{word_1, word_2, \ldots, word_k\}$$

$$S_1 = \{word_1, word_2, \ldots, word_k\}$$

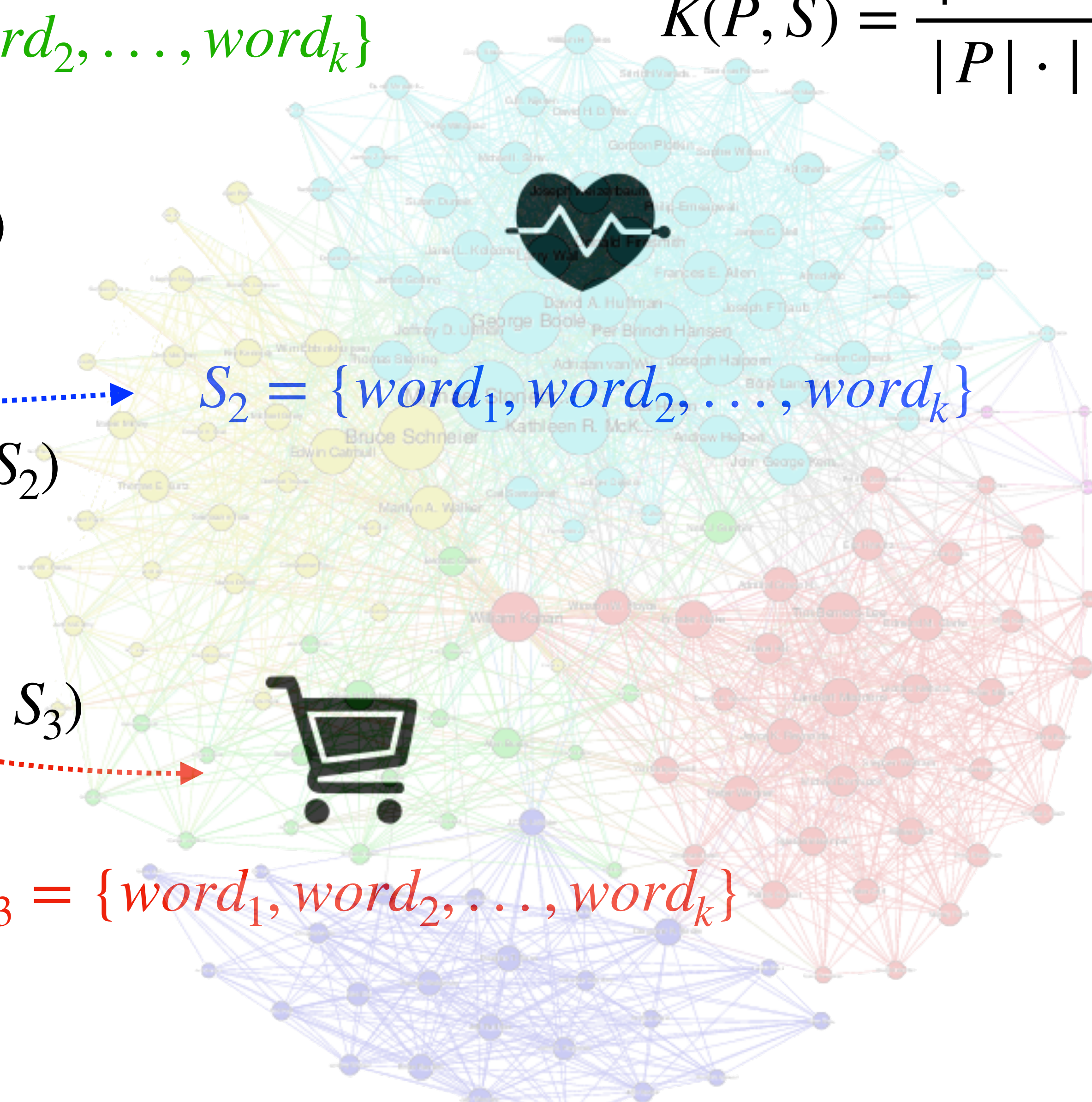
$J(P, S_1)$ $K(P, S_1)$

$$S_2 = \{word_1, word_2, \ldots, word_k\}$$

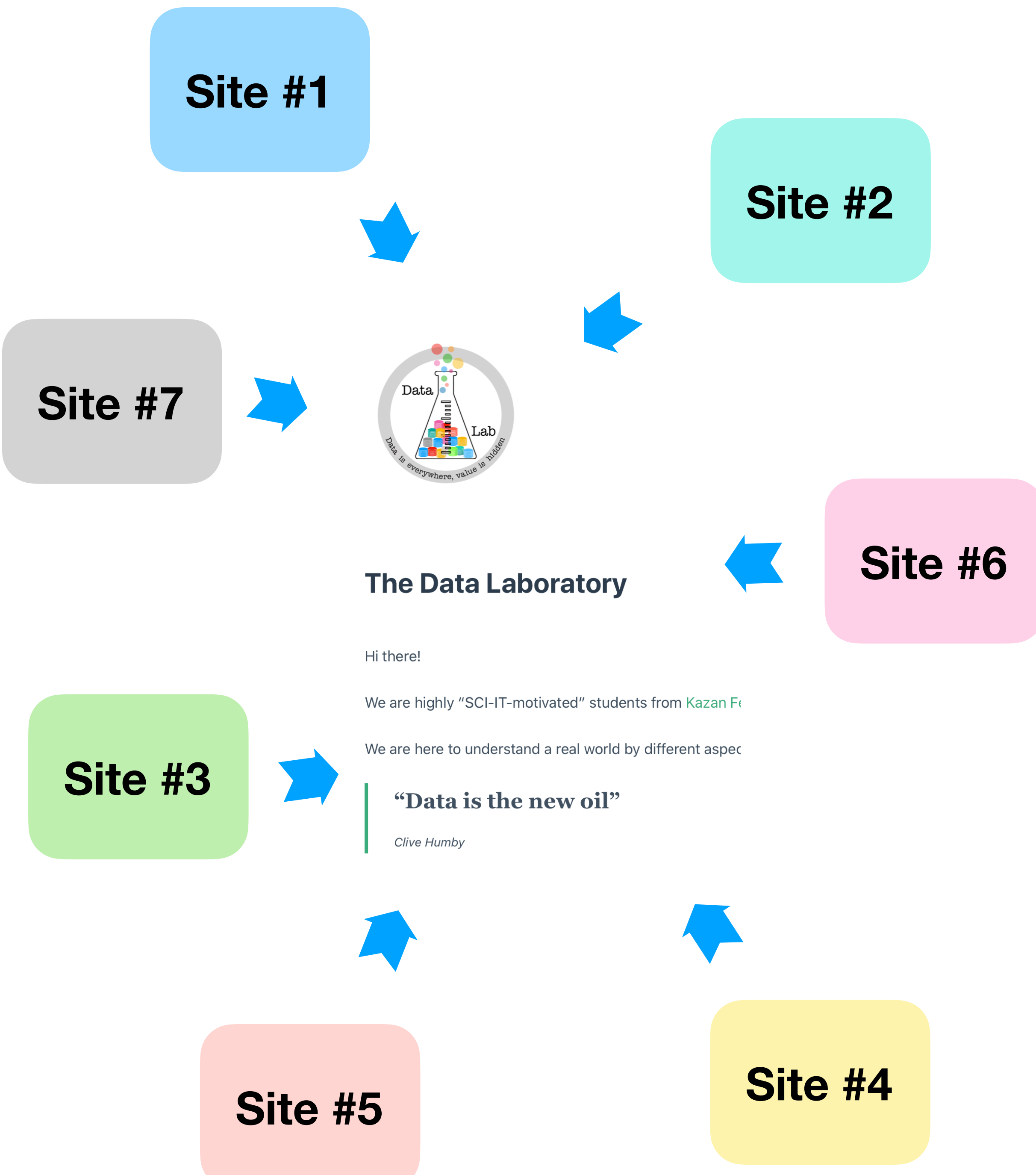
$J(P, S_2)$ $K(P, S_2)$

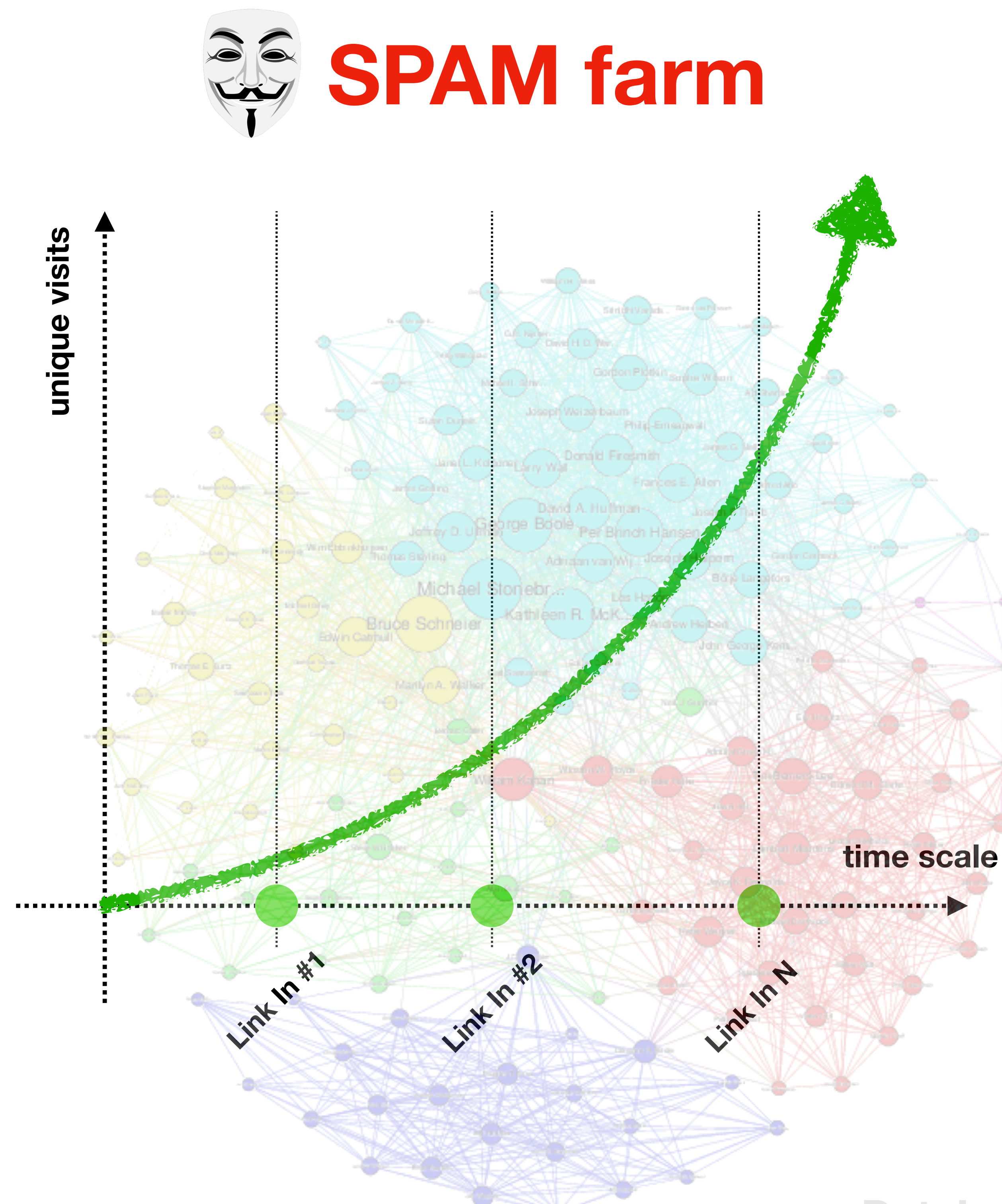
$$S_3 = \{word_1, word_2, \ldots, word_k\}$$

$J(P, S_3)$ $K(P, S_3)$



Spam Farm


Site #1
Site #2
Site #7
Data
Lab
Site #6
The Data Laboratory
Hi there!
We are highly “SCI-IT-motivated” students from Kazan F
We are here to understand a real world by different aspec
“Data is the new oil”
Clive Humby
Site #3
Site #5
Site #4


SPAM farm
unique visits
time scale
Link In #1
Link In #2
Link In N

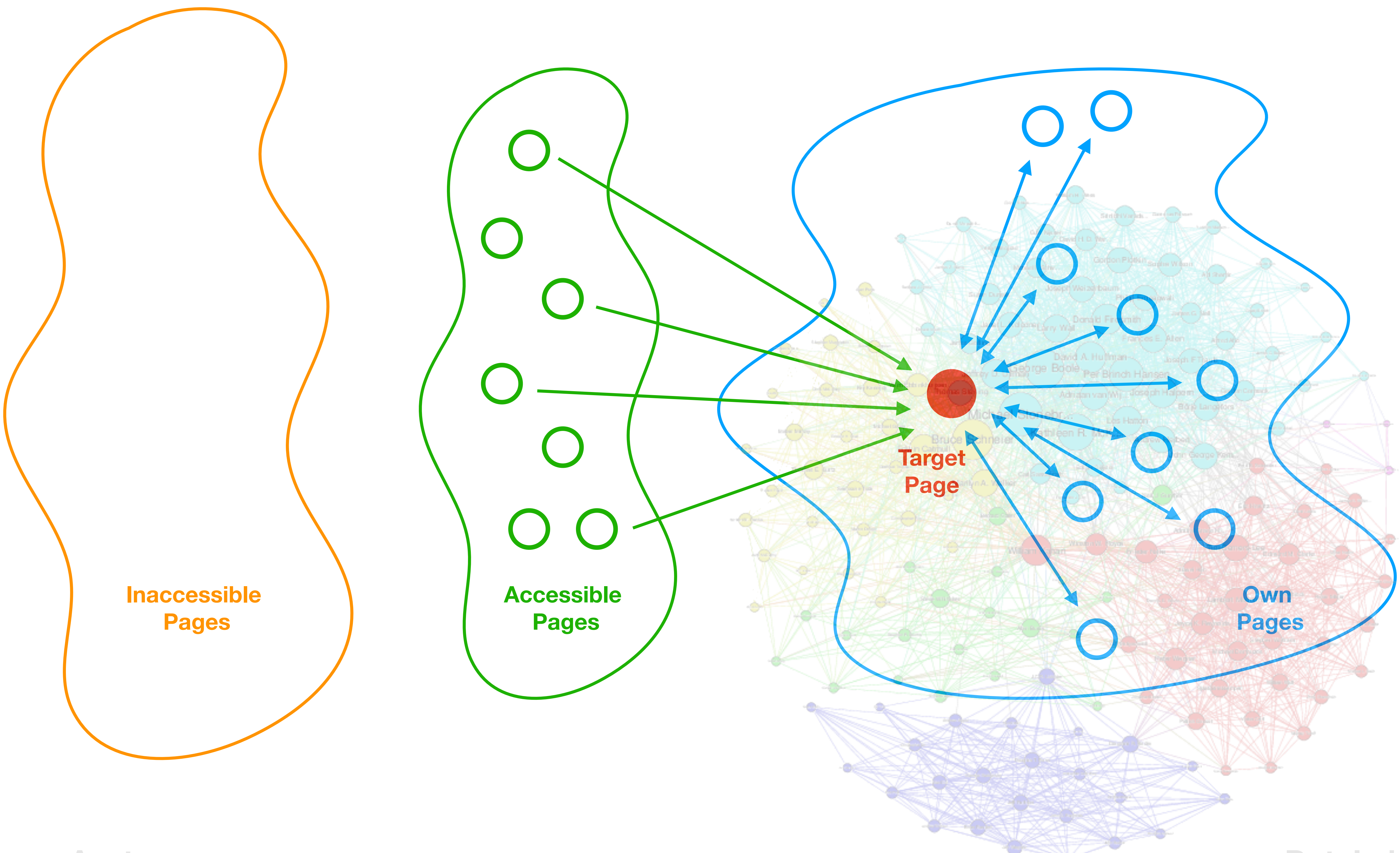
Architecture of a Spam Farm

Spam farm - of pages whose purpose is to increase the PageRank of a certain page or pages

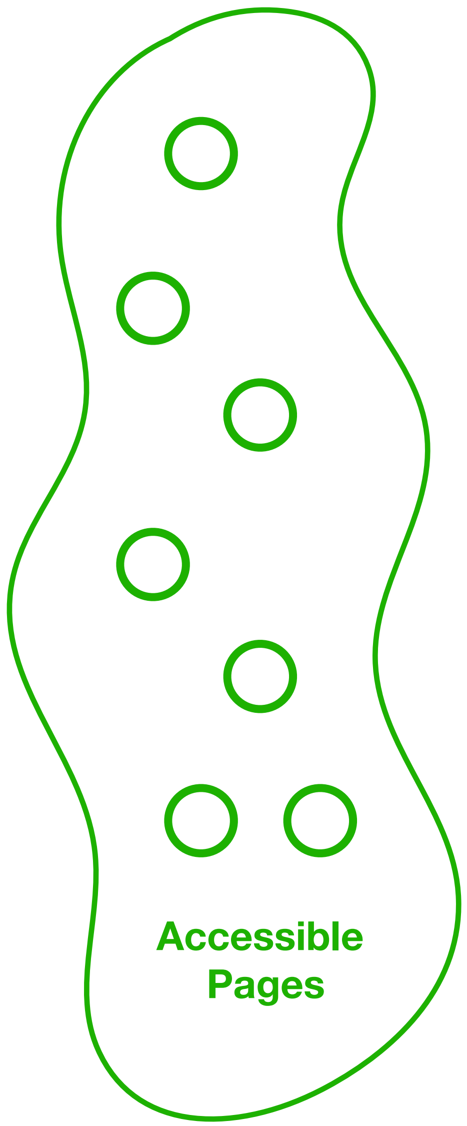
From the point of view of the Spammer the Web is divided into 3 areas

- **Inaccessible pages**: the pages that the spammer cannot affect. Most of the Web is in this part.
- **Accessible pages:** those pages that, while they are not controlled by the spammer, can be affected by the spammer.
- **Own pages:** the pages that the spammer owns and controls.

Architecture of a Spam Farm

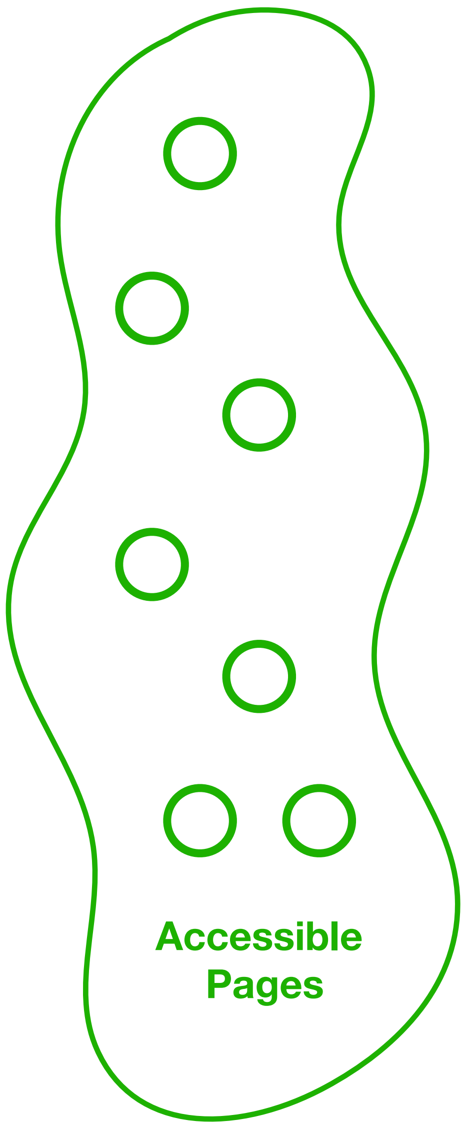


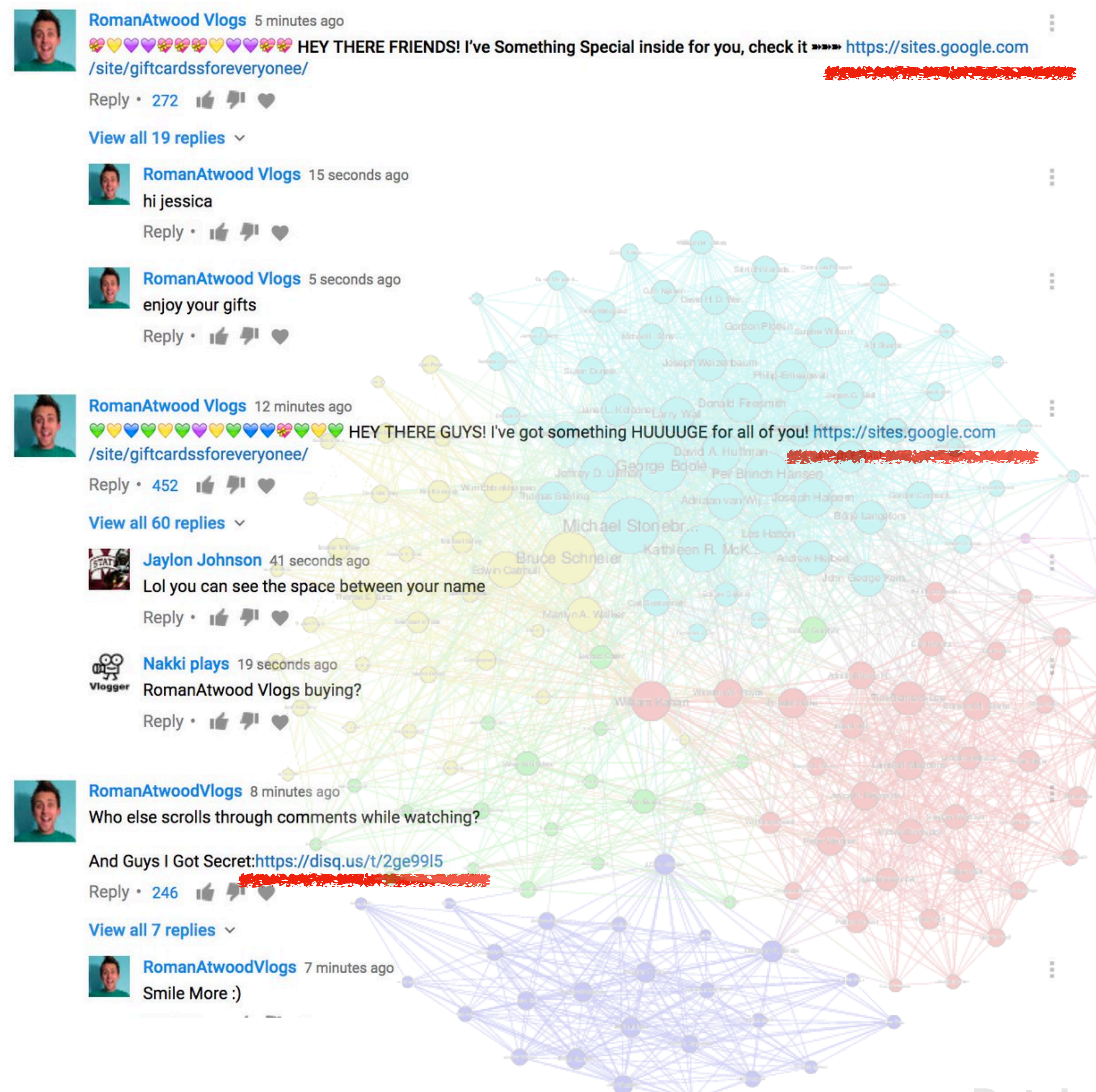
Architecture of a Spam Farm



Example a links from accessible pages

Architecture of a Spam Farm





Analysis of a Spam Farm

Let's $\beta = 0.85$ (taxation parameter of dumping factor)

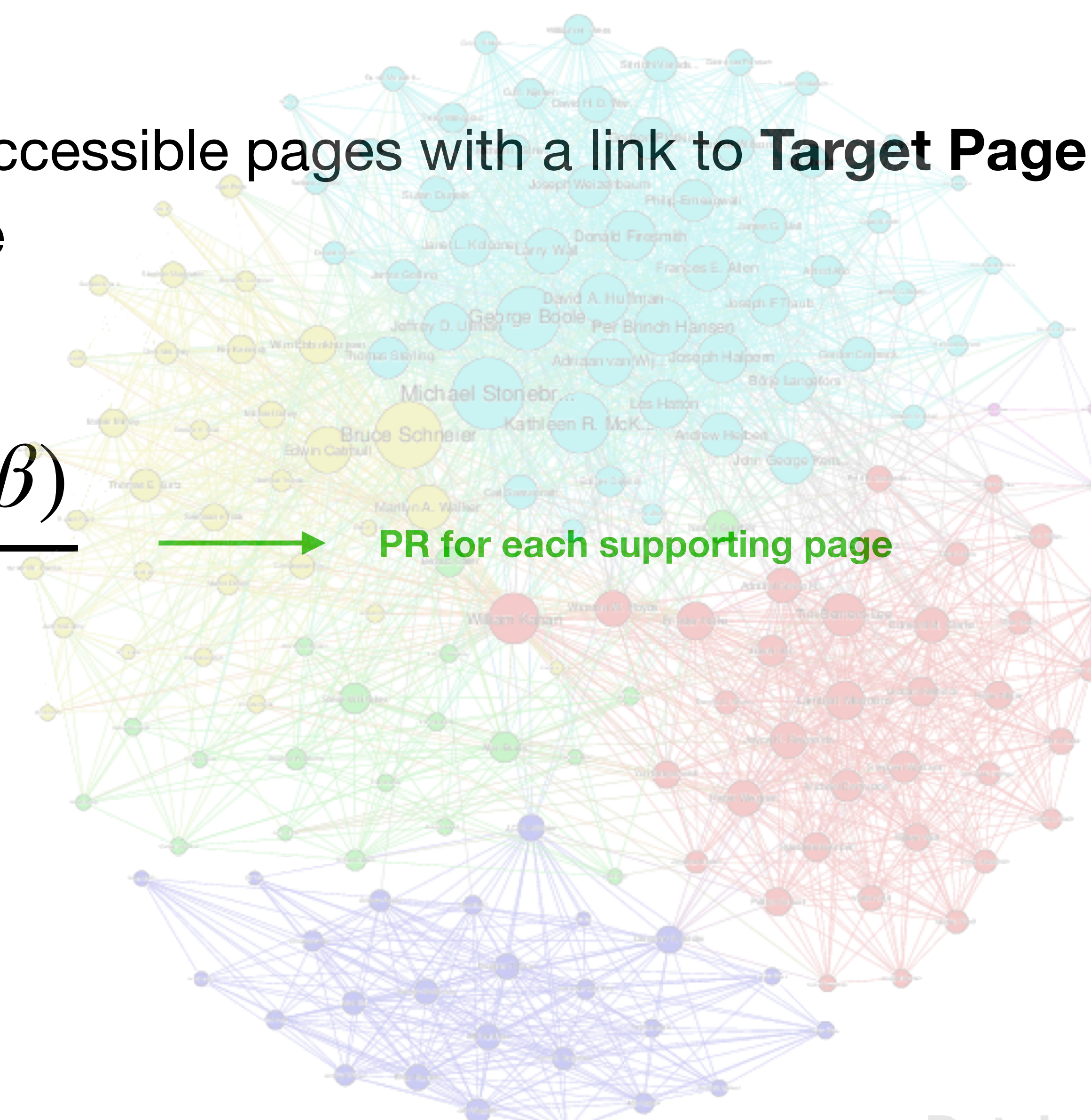
n - total pages in the Web

m - supporting pages from Spam Farm

x - is the sum of the PageRanks, over all accessible pages with a link to **Target Page**

y - the **unknown** PageRank of Target Page

$$f_{PR}(m_i) = \beta \cdot \frac{y}{m} + \frac{(1-\beta)}{n} \longrightarrow \text{PR for each supporting page}$$



Analysis of a Spam Farm

Page Rank y of **Target Page** comes from 3 sources

(1) - Contribution x from outside, as we have assumed.

(2) - β times the PageRank of every supporting page; that is

$$\beta \cdot (\beta \cdot \frac{y}{m} + \frac{(1-\beta)}{n})$$

(3) - $(1-\beta)/n$, the share of the fraction 1 – β of the PageRank that belongs to **Target Page**.

$$(1-\beta)/n \rightarrow 0$$

Analysis of a Spam Farm

y = (1) + (2) + (3)

$$y = x + \beta m(\frac{\beta y}{m} + \frac{(1-\beta)}{n}) + 0 = x + \beta^2 y + \beta(1-\beta)\frac{m}{n}$$

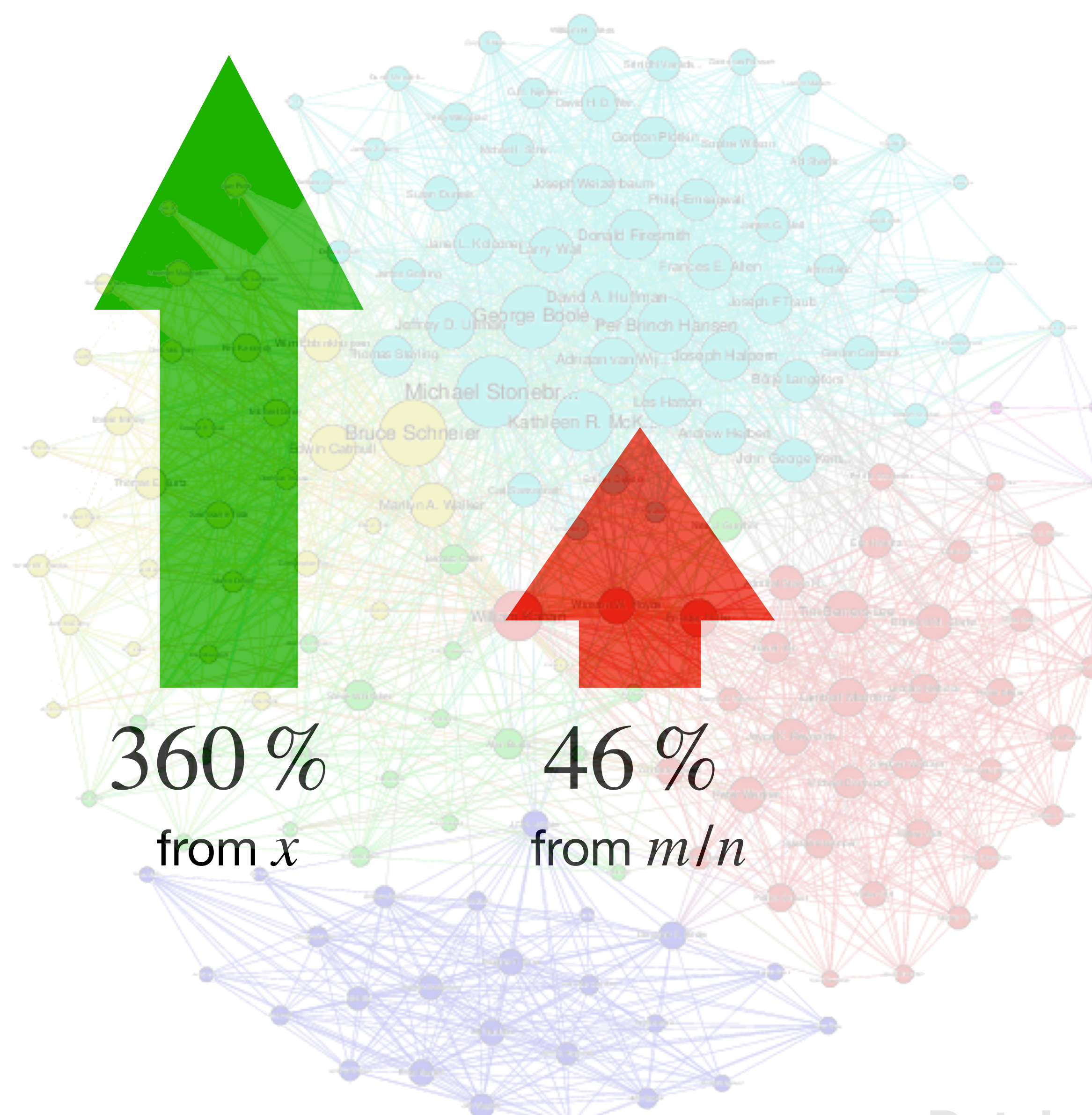
$$y = \frac{x}{1-\beta^2} + c\frac{m}{n}$$, where $c = \beta(1-\beta)/(1-\beta^2) = \beta/(1+\beta)$

Analysis of a Spam Farm

$$y = \frac{x}{1-\beta^2} + c\frac{m}{n}$$, where $c = \beta(1-\beta)/(1-\beta^2) = \beta/(1+\beta)$

$\beta = 0.85$

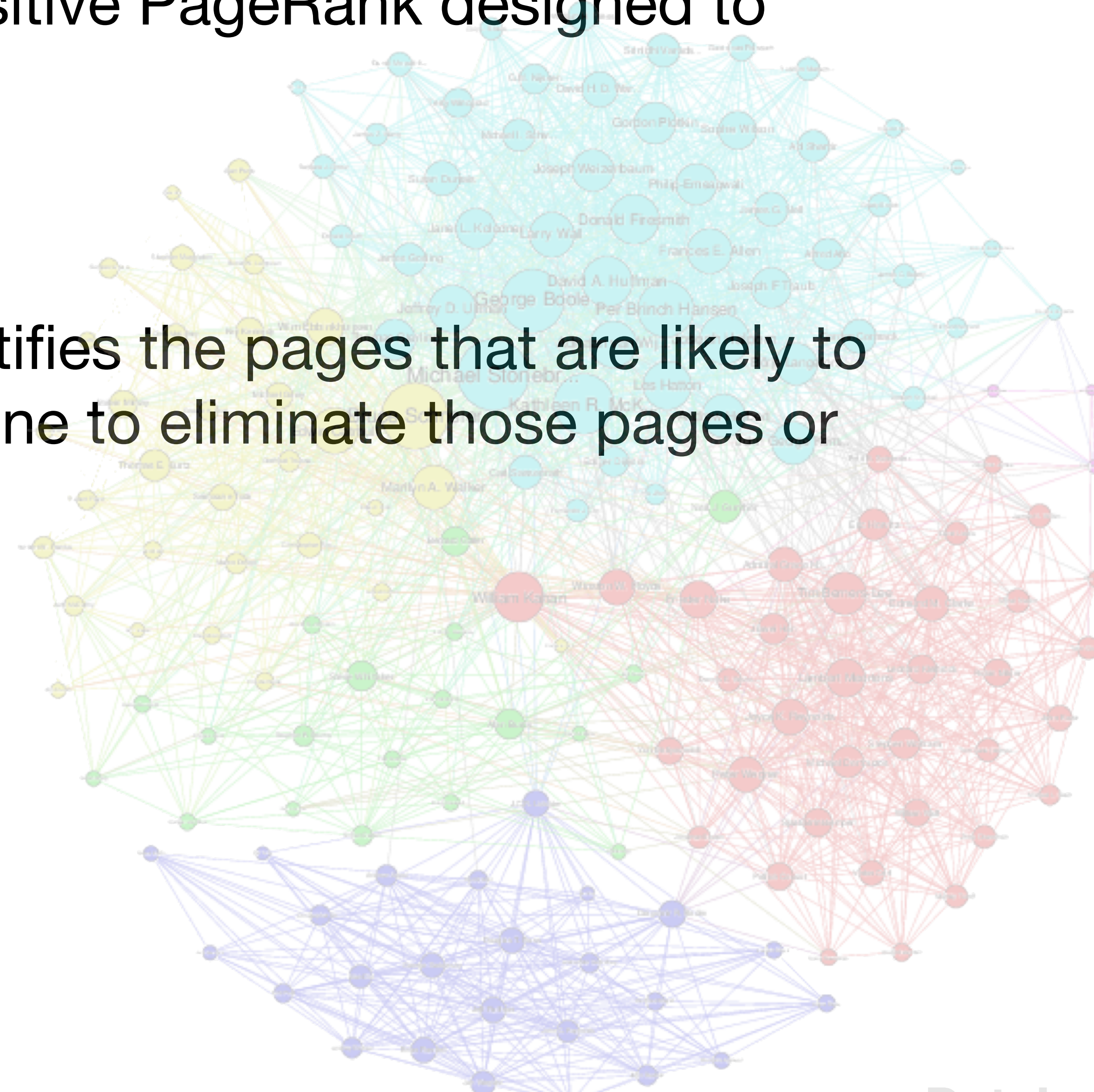
$$\frac{1}{1-\beta^2} = 3.6 \qquad \frac{\beta}{1+\beta} = 0.46$$



The Empire Strikes Back

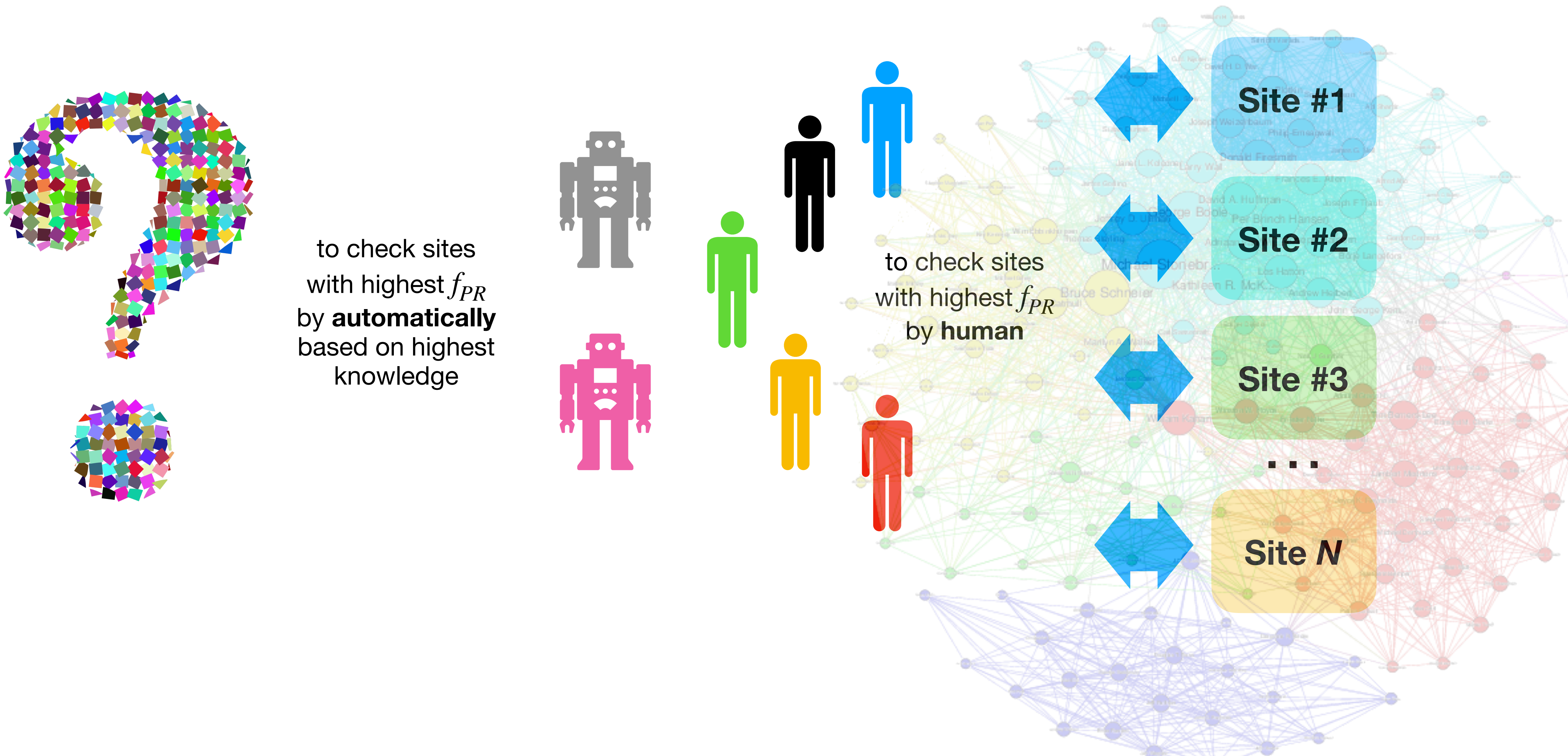
- ***TrustRank***, a variation of topic-sensitive PageRank designed to lower the score of spam pages

- ***Spam mass***, a calculation that identifies the pages that are likely to be spam and allows the search engine to eliminate those pages or to lower their PageRank strongly



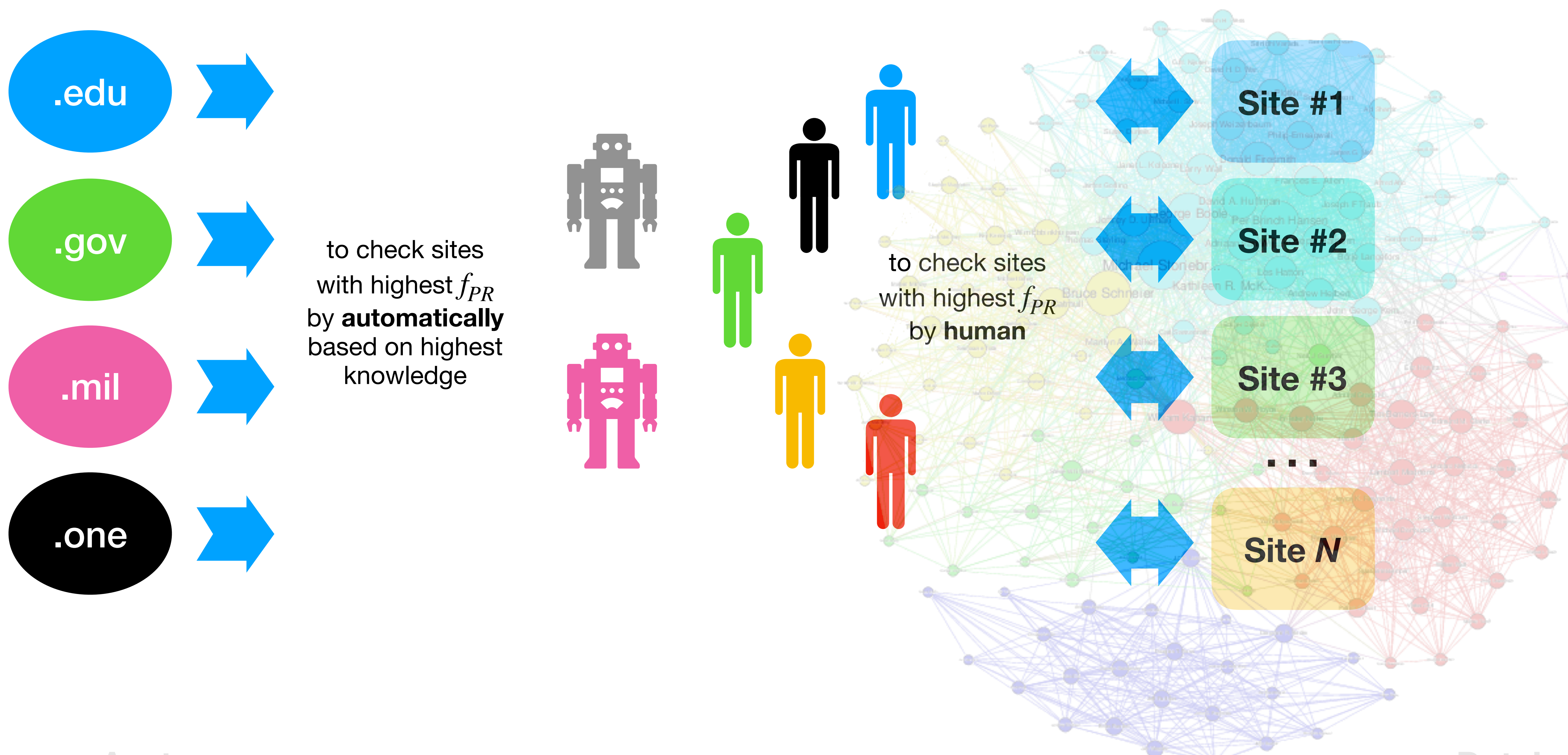
TrustRank

TrustRank is topic-sensitive PageRank, where the “topic” is a set of pages believed to be trustworthy (**not spam**)



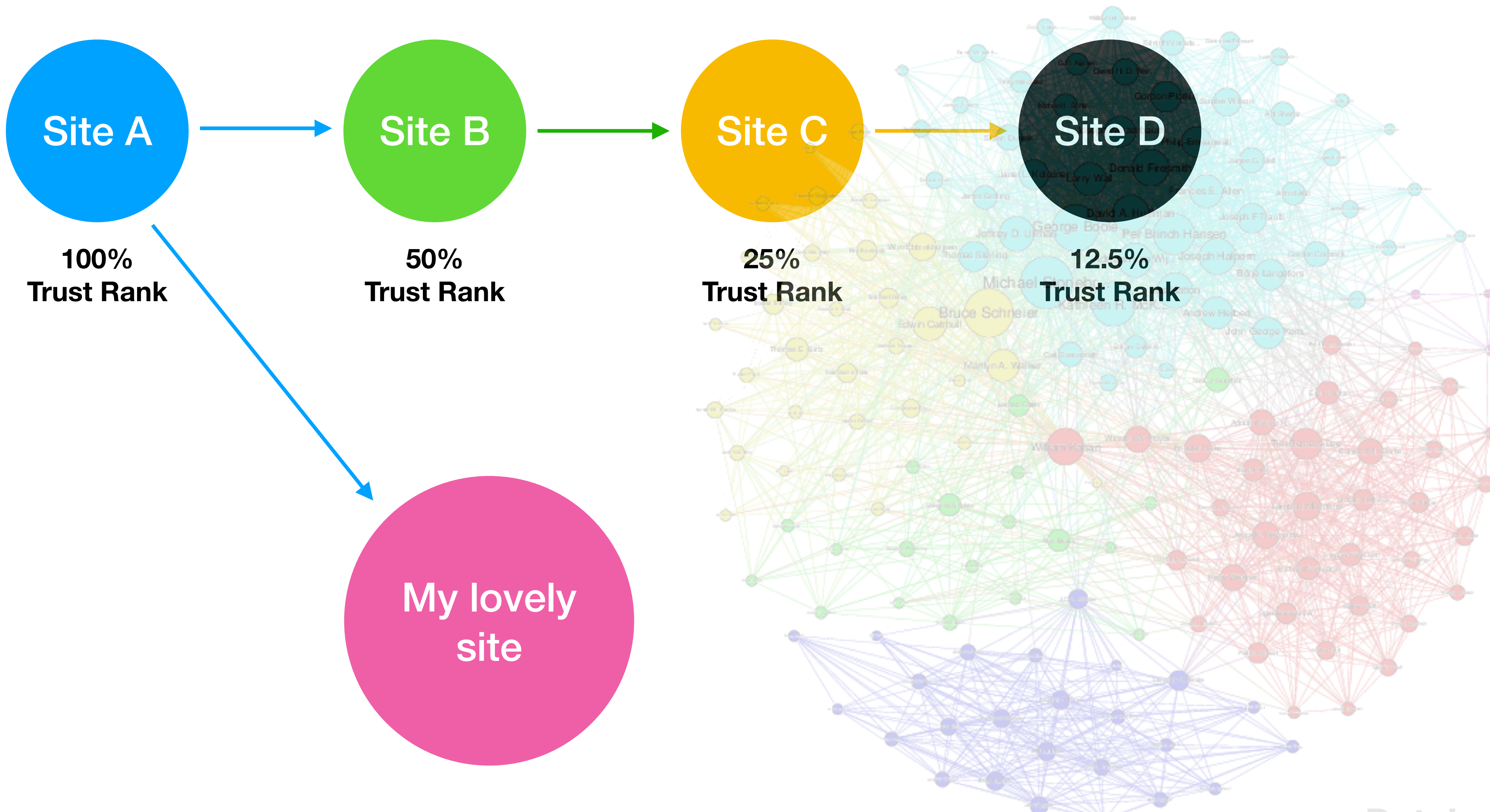
TrustRank

TrustRank is topic-sensitive PageRank, where the “topic” is a set of pages believed to be trustworthy (**not spam**)



TrustRank

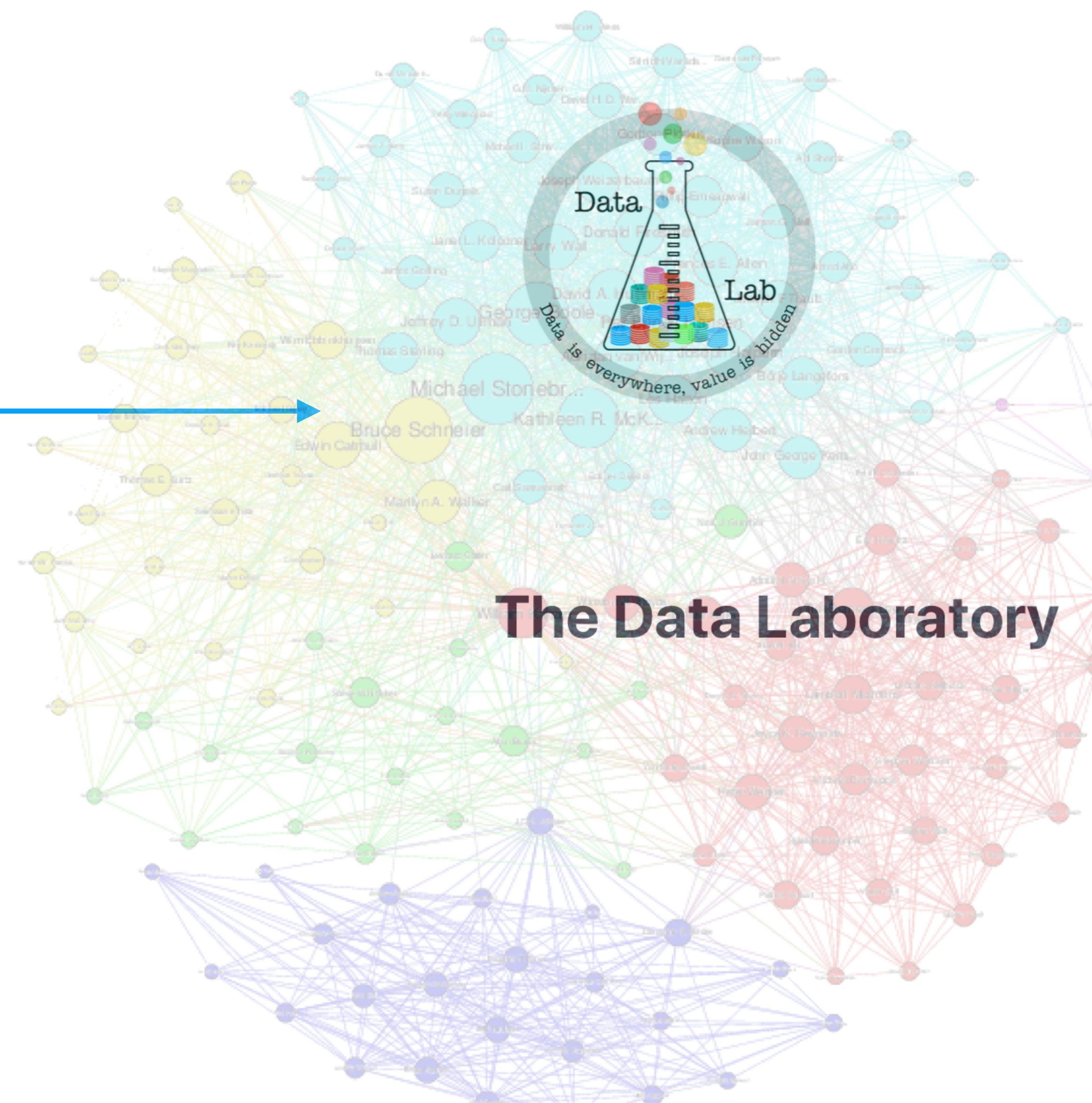
TrustRank is topic-sensitive PageRank, where the “topic” is a set of pages believed to be trustworthy (**not spam**)



TrustRank

TrustRank is topic-sensitive PageRank, where the "topic" is a set of pages believed to be trustworthy (**not spam**)

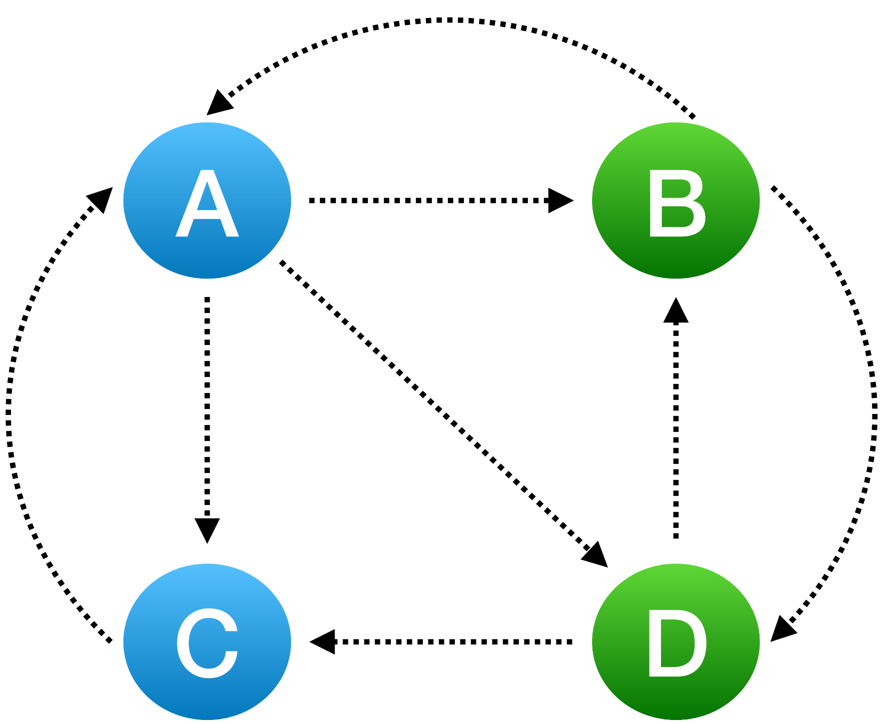
100%
Trust Rank



Spam Mass

Spam mass, a calculation that identifies the pages that are likely to be spam and allows the search engine **to eliminate** those pages or to lower their PageRank strongly

Let's page p has $f_{PR}(p) = r$ and $f_{TR}(p) = t$ then $f_{spam}(p) = \dfrac{r - t}{r}$

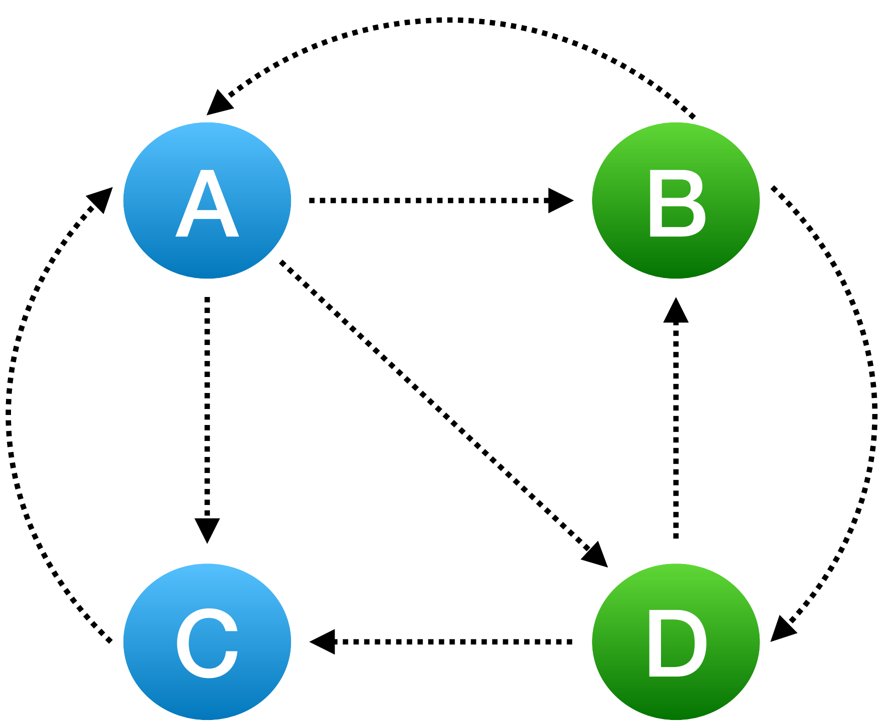


$S = \{B, D\}$ **- trusted pages**

Node	f_{PR}	f_{TR}	f_{spam}
A	$\frac{3}{9}$	$\frac{54}{210}$	≈ 0.229
B	$\frac{2}{9}$	$\frac{59}{210}$	≈ -0.264
C	$\frac{2}{9}$	$\frac{38}{210}$	≈ 0.186
D	$\frac{2}{9}$	$\frac{59}{210}$	≈ -0.264

Spam Mass

Spam mass, a calculation that identifies the pages that are likely to be spam and allows the search engine **to eliminate** those pages or to lower their PageRank strongly



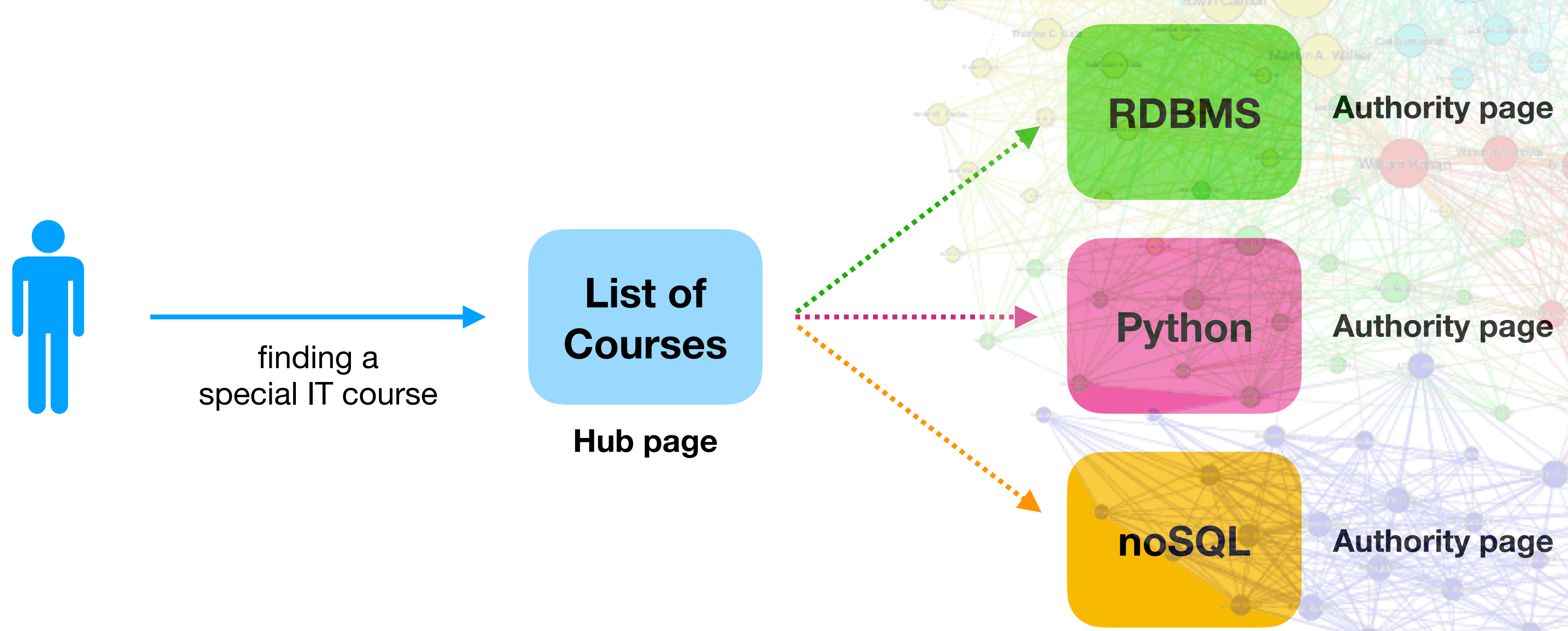
$S = \{B, D\}$ **- trusted pages**

Node	f_{spam}	
A	$\approx 0.229 \in [0,1]$	**- closely to 0 - not a spam**
B	$\approx -0.264 < 0$	**- trusted page**
C	$\approx 0.186 \in [0,1]$	**- closely to 0 - not a spam**
D	$\approx -0.264 < 0$	**- trusted page**

HITS Algorithm

HITS - hyperlink- induced topic search (Hubs and Authorities)

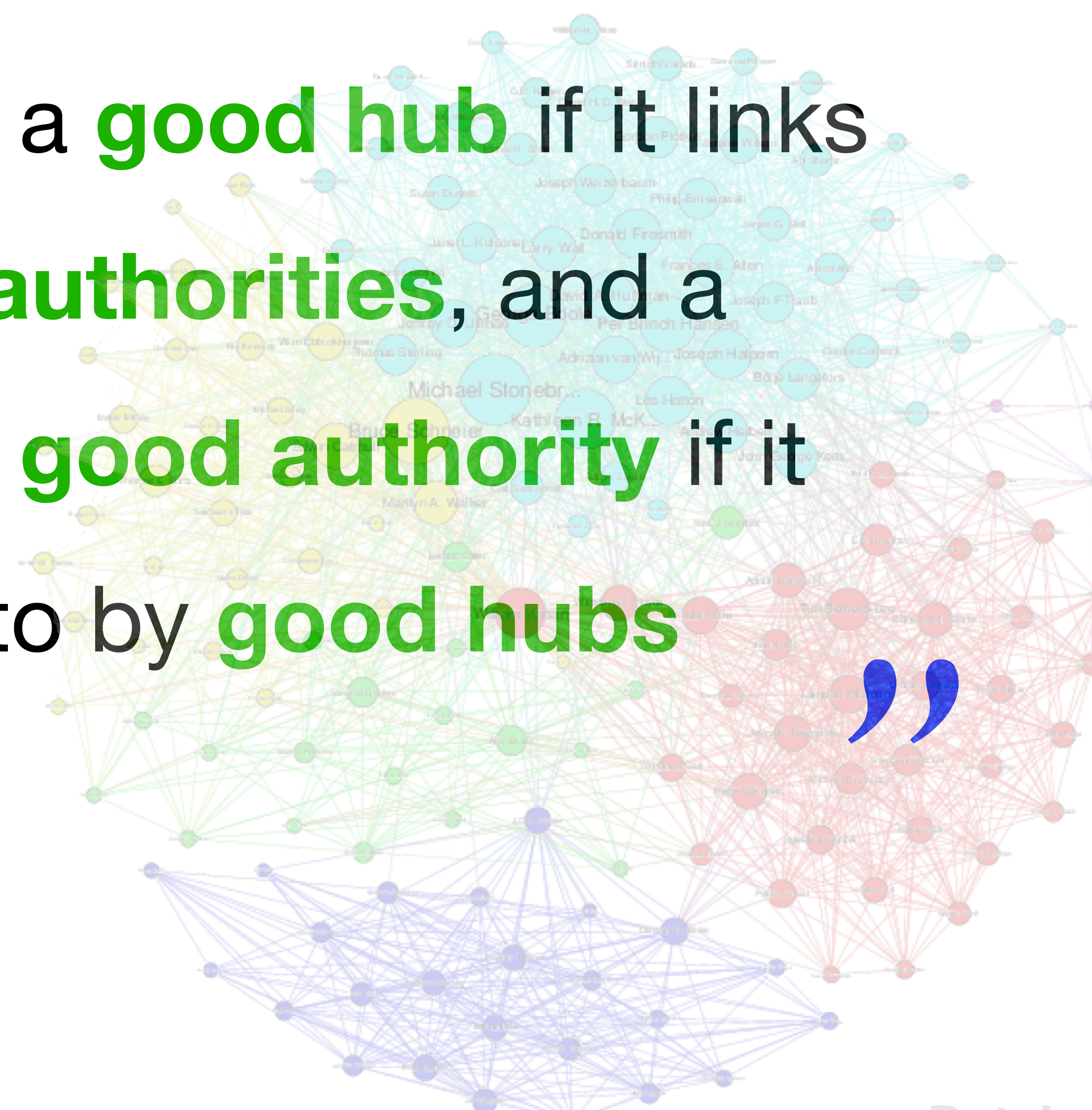
- Certain pages are valuable because they provide information about a topic. These pages are called **authorities**.
- Other pages are valuable not because they provide information about any topic, but because they tell you where to go to find out about that topic. These pages are called **hubs.**



HITS Algorithm

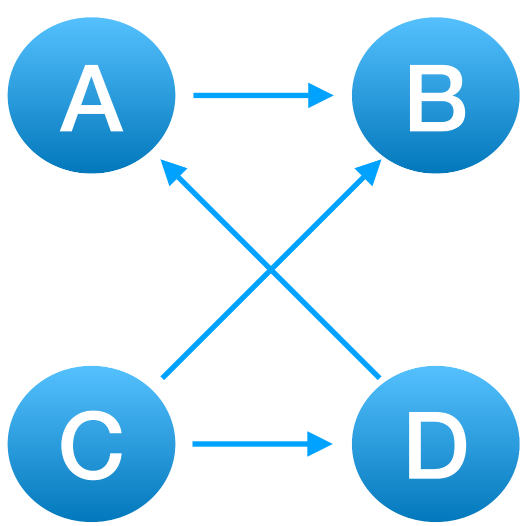
HITS - hyperlink- induced topic search (Hubs and Authorities)

“a page is a **good hub** if it links to **good authorities**, and a page is a **good authority** if it is linked to by **good hubs**”



HITS Algorithm

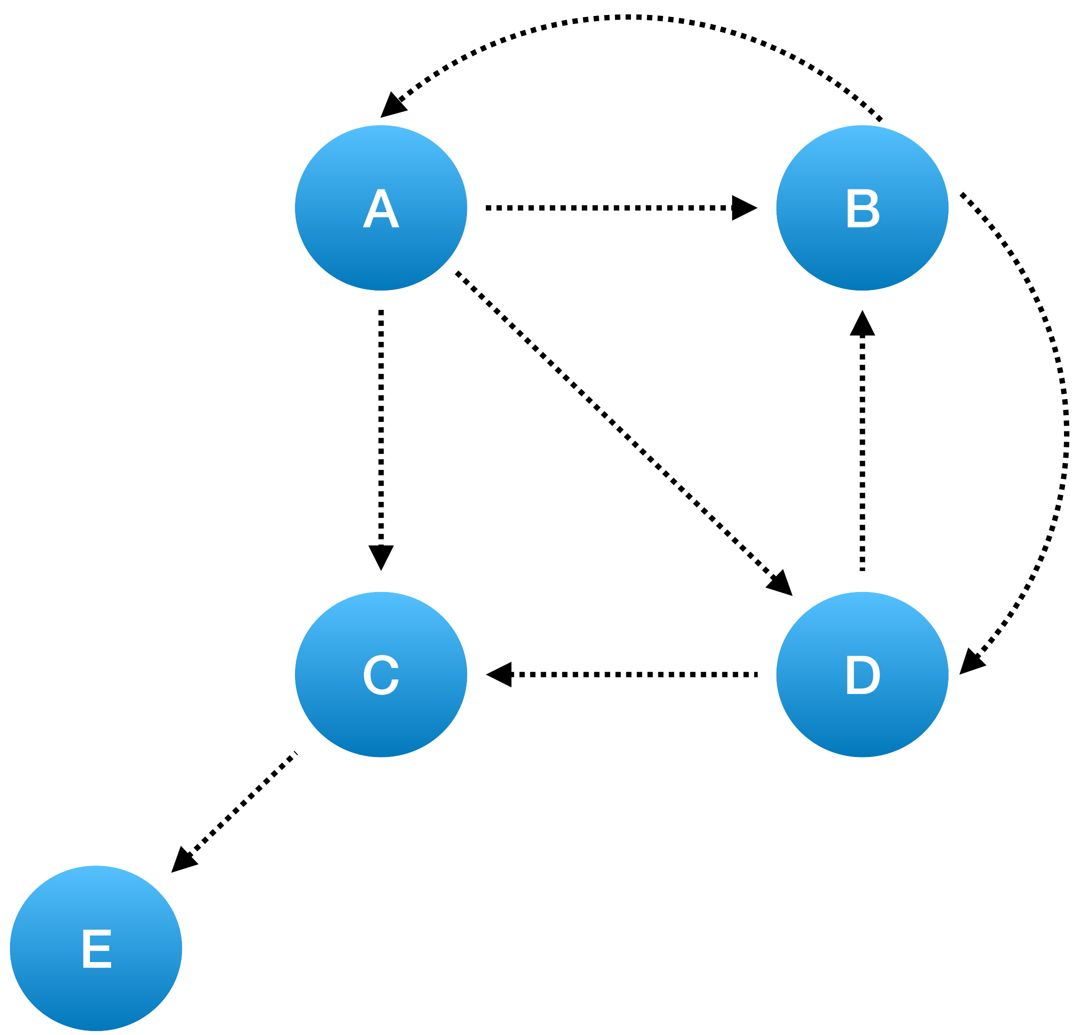
Link matrix L, $L_{ij} = \begin{cases} 1 & , \exists\, i \to j \\ 0 & \end{cases}$

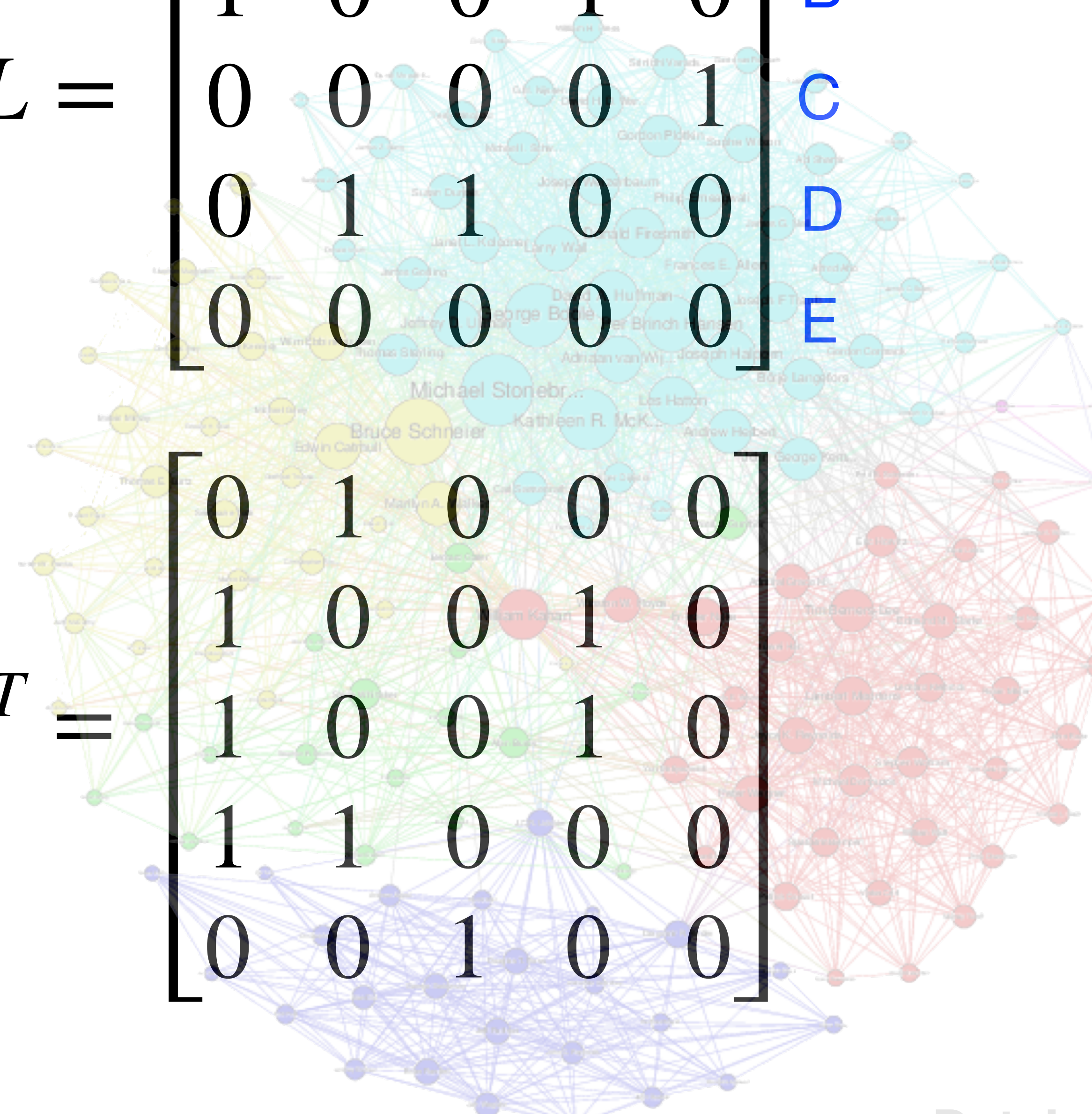


Transpose matrix L^T, $L^T_{ij} = \begin{cases} 1 & , \exists\, j \to i \\ 0 & \end{cases}$

$$L = \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{array}{c} A \\ B \\ C \\ D \end{array} \qquad L^T = \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array} \begin{array}{c} A \\ B \\ C \\ D \end{array}$$

HITS Algorithm





$$L = \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array}$$

$$L^T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

HITS Algorithm

$$h = \lambda L a$$

$$a = \mu L^T h$$

- λ - is an unknown constant representing the scaling factor needed
- μ - is an another scaling constant

initial vector

1st iteration

$$\underset{h}{\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}} \rightarrow \underset{L^Th}{\begin{bmatrix}1\\2\\2\\2\\1\end{bmatrix}} \rightarrow \underset{\substack{a\\ \mu = 1/2}}{\begin{bmatrix}1/2\\1\\1\\1\\1/2\end{bmatrix}} \rightarrow \underset{La}{\begin{bmatrix}3\\3/2\\1/2\\2\\0\end{bmatrix}}$$

2nd iteration

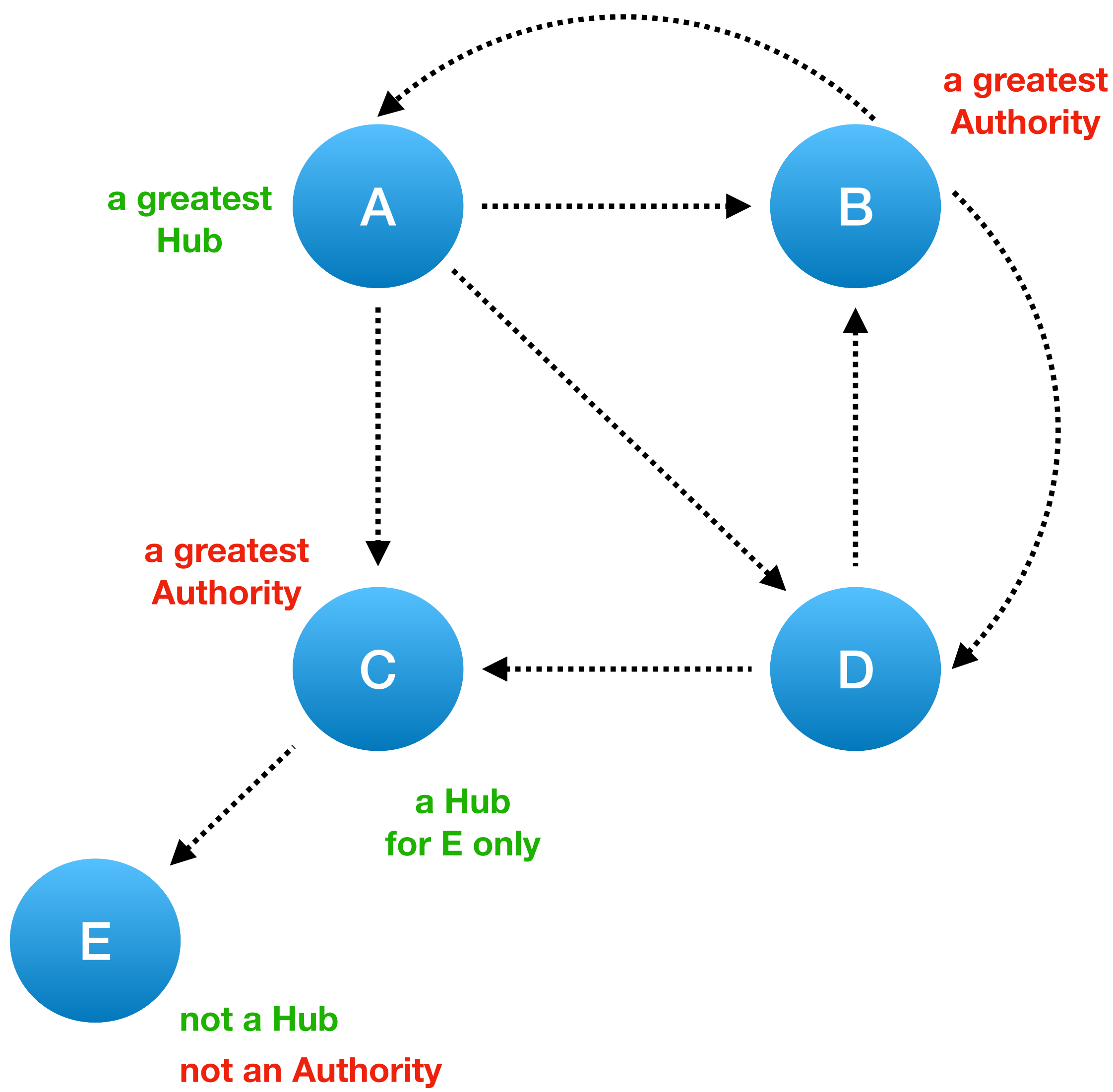
$$\rightarrow \underset{\substack{h\\ \lambda = 1/3}}{\begin{bmatrix}1\\1/2\\1/6\\2/3\\0\end{bmatrix}} \rightarrow \underset{L^Th}{\begin{bmatrix}1/2\\5/3\\5/3\\3/2\\1/6\end{bmatrix}} \rightarrow \underset{\substack{a\\ \mu = 3/5}}{\begin{bmatrix}3/10\\1\\1\\9/10\\1/10\end{bmatrix}} \rightarrow \underset{La}{\begin{bmatrix}29/10\\6/5\\1/10\\2\\0\end{bmatrix}}$$

3rd

$$\rightarrow \underset{\substack{h\\ \lambda = 10/29}}{\begin{bmatrix}1\\12/29\\1/29\\20/29\\0\end{bmatrix}}$$

- maximum number, need to make a scaling

HITS Algorithm



$$a = \begin{array}{c} \left[\begin{array}{c} 0.2087 \\ 1 \\ 1 \\ 0.7913 \\ 0 \end{array}\right] \end{array} \begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} \qquad h = \left[\begin{array}{c} 1 \\ 0.3583 \\ 0 \\ 0.7165 \\ 0 \end{array}\right] \begin{array}{c} A \\ B \\ C \\ D \\ E \end{array}$$

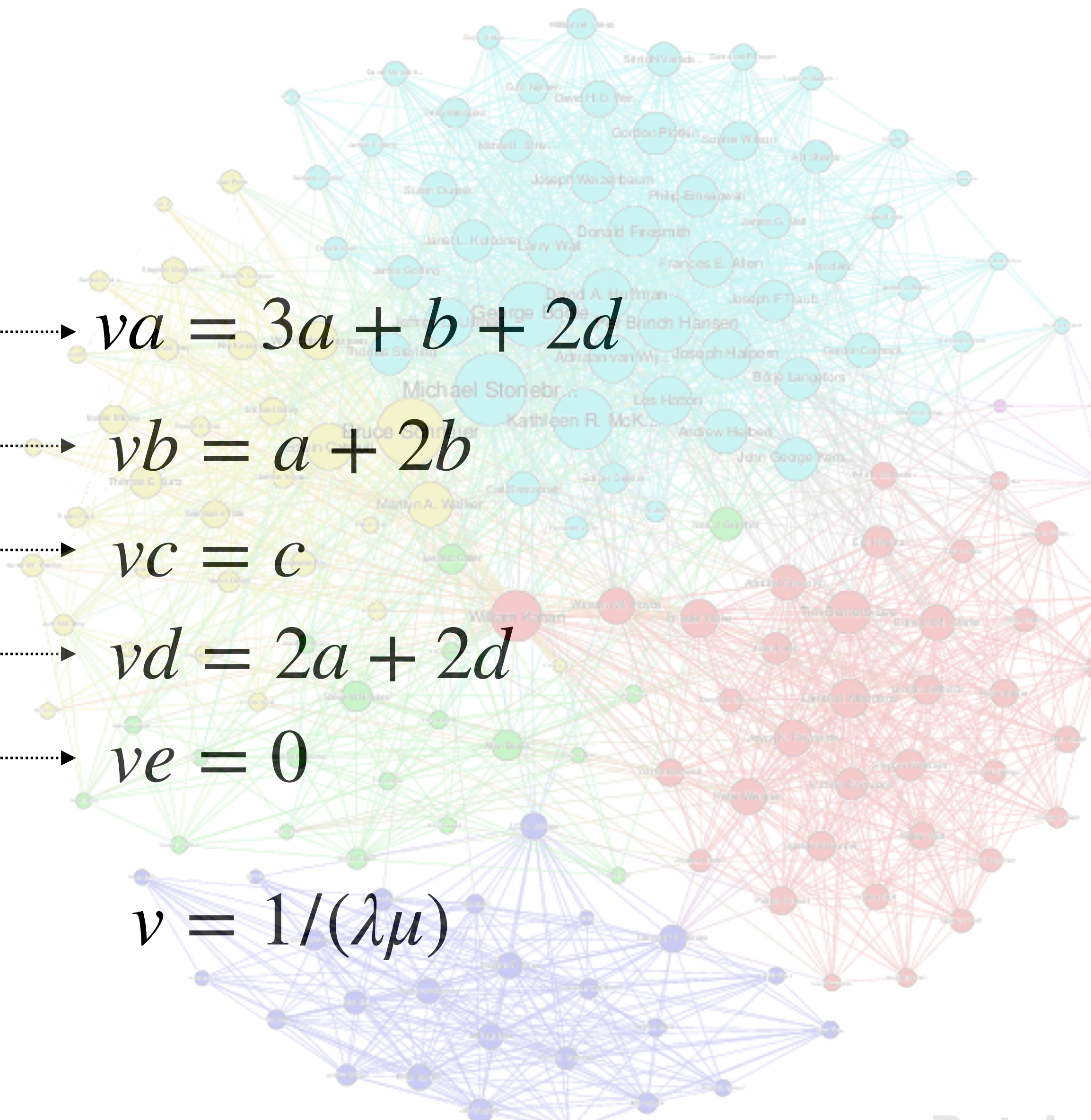
HITS Algorithm

$$h = \lambda L a \qquad \Rightarrow \qquad h = \lambda \mu \underline{L L^T} h$$

$$a = \mu L^T h \qquad \Rightarrow \qquad a = \lambda \mu L^T L a$$

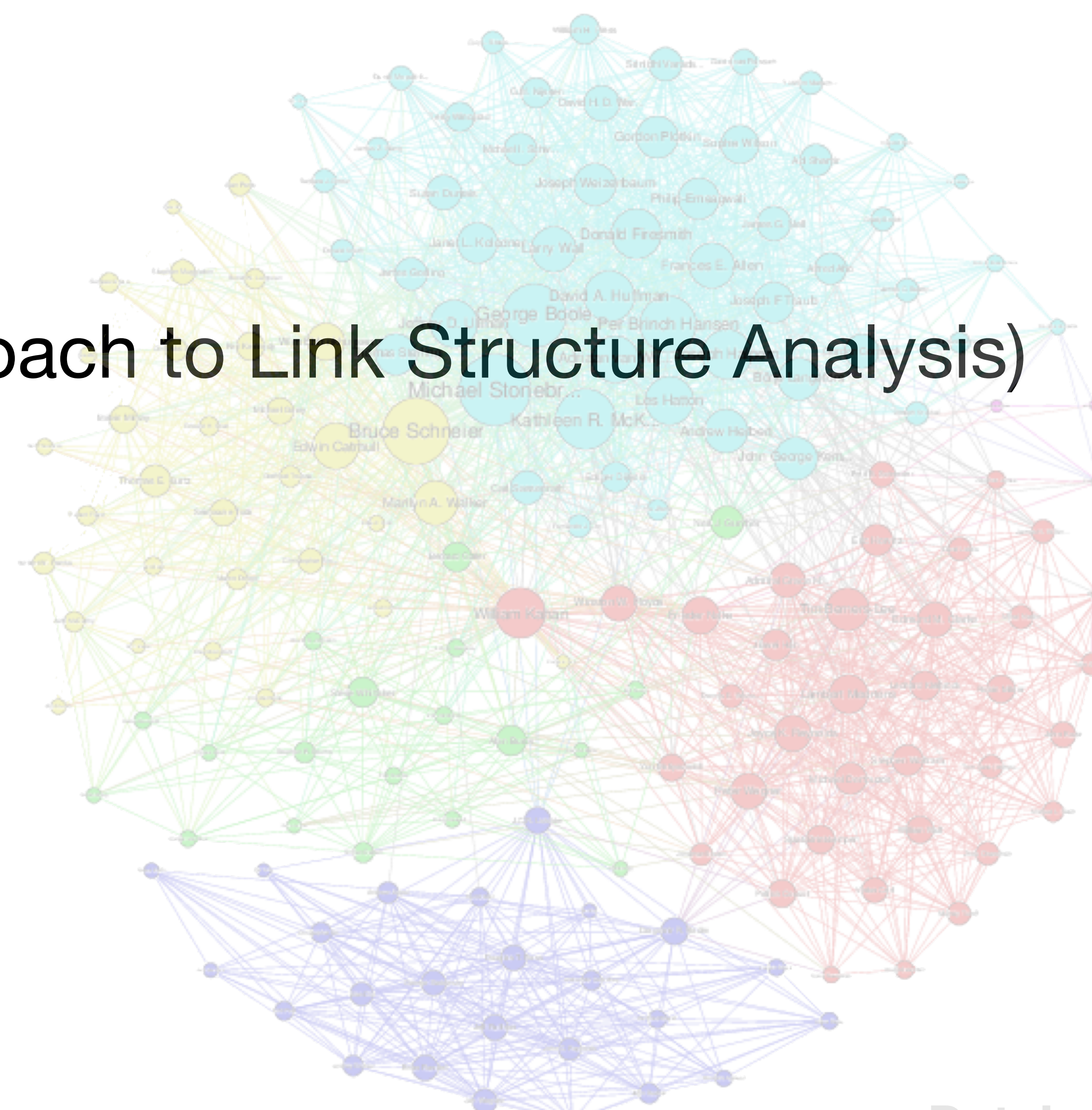
$$\underline{LL^T} = \begin{bmatrix} 3 & 1 & 0 & 2 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{array}{l} \to va = 3a + b + 2d \\ \to vb = a + 2b \\ \to vc = c \\ \to vd = 2a + 2d \\ \to ve = 0 \end{array}$$

$$v = 1/(\lambda\mu)$$

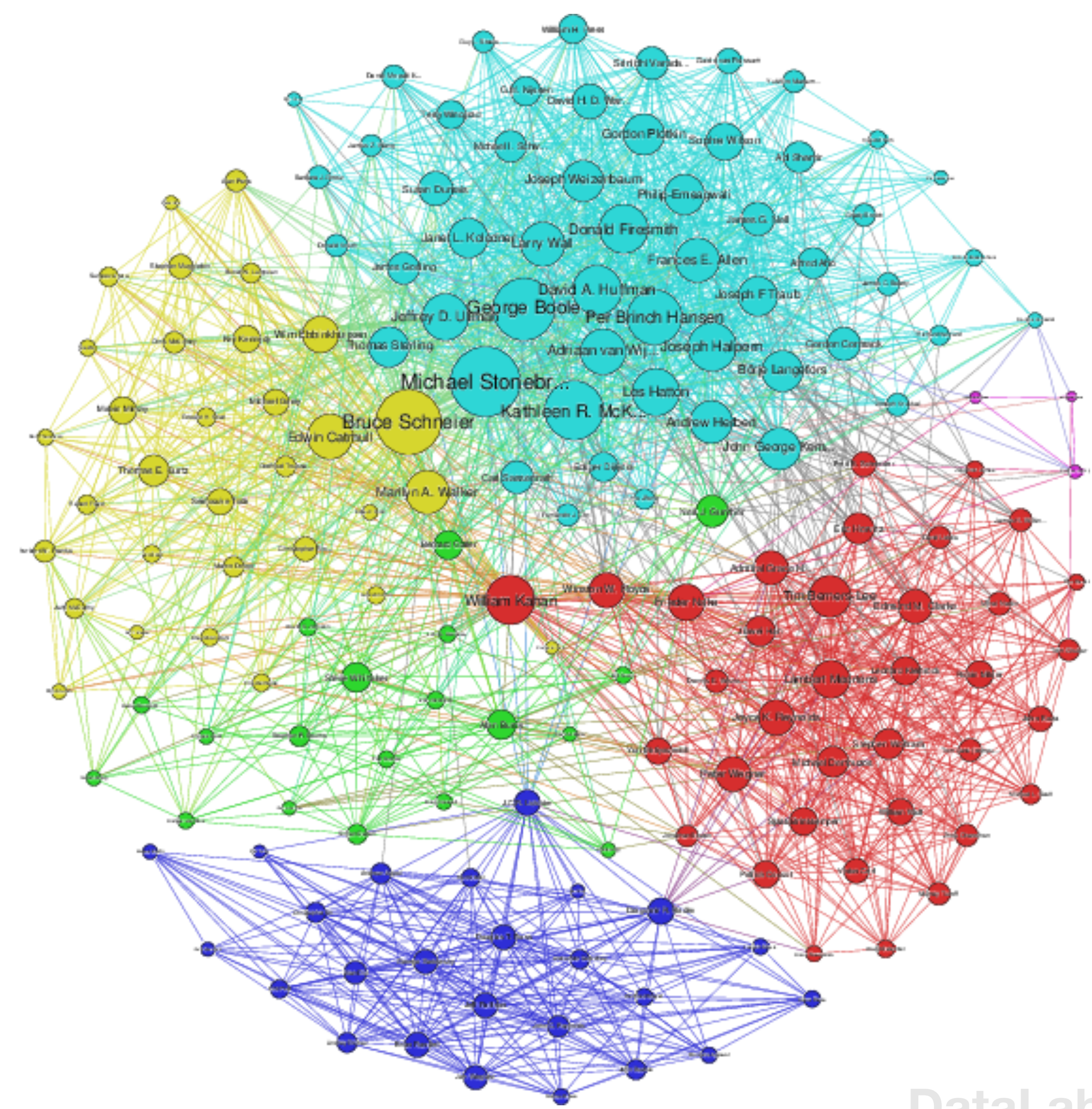


Other Algorithms

- Block Rank
- Host Rank
- SALSA (Stochastic Approach to Link Structure Analysis)
- Bad Rank
- Traffic Rank



Thanks!


Michael Stonebr...
Bruce Schneier
George Boole
Per Brinch Hansen
Kathleen R. McK...
Donald Firesmith
David A. Huffman
Joseph Weizenbaum
William Kahan
Marilyn A. Walker
Edwin Catmull