DiffusionRank: A Possible Penicillin for Web Spamming

Haixuan Yang, Irwin King, and Michael R. Lyu

Department of Computer Science & Engineering
The Chinese University of Hong Kong

SIGIR2007, Amsterdam, Netherlands
July 25, 2007
State of the Web

- Web is easily **manipulated** for commercial gains
  - About 70% of all pages in the .biz domain are spam [Alexandros Ntoulas et al., 2006]
  - About 35% of the pages in the .us domain belong to spam category [Alexandros Ntoulas et al., 2006]
- Web spamming techniques
  - Link Stuffing
  - Keyword Stuffing
- PageRank becomes the target of many spamming techniques
State of the Web

- Web is easily *manipulated* for commercial gains
  - About 70% of all pages in the .biz domain are spam [Alexandros Ntoulas et al., 2006]
  - About 35% of the pages in the .us domain belong to spam category [Alexandros Ntoulas et al., 2006]

- Web spamming techniques
  - Link Stuffing
  - Keyword Stuffing

- PageRank becomes the target of many spamming techniques
State of the Web

- Web is easily manipulated for commercial gains
  - About 70% of all pages in the .biz domain are spam [Alexandros Ntoulas et al., 2006]
  - About 35% of the pages in the .us domain belong to spam category [Alexandros Ntoulas et al., 2006]

- Web spamming techniques
  - Link Stuffing
  - Keyword Stuffing

- PageRank becomes the target of many spamming techniques
Spam, Spam, Spam Everywhere

State of the Web

- Web is easily manipulated for commercial gains
  - About 70% of all pages in the .biz domain are spam [Alexandros Ntoulas et al., 2006]
  - About 35% of the pages in the .us domain belong to spam category [Alexandros Ntoulas et al., 2006]

- Web spamming techniques
  - Link Stuffing
  - Keyword Stuffing
  - PageRank becomes the target of many spamming techniques
State of the Web

- Web is easily **manipulated** for commercial gains
  - About 70% of all pages in the .biz domain are spam [Alexandros Ntoulas et al., 2006]
  - About 35% of the pages in the .us domain belong to spam category [Alexandros Ntoulas et al., 2006]

- Web spamming techniques
  - **Link Stuffing**
  - **Keyword Stuffing**

- PageRank becomes the **target** of many spamming techniques
State of the Web

- Web is easily manipulated for commercial gains
  - About 70% of all pages in the .biz domain are spam [Alexandros Ntoulas et al., 2006]
  - About 35% of the pages in the .us domain belong to spam category [Alexandros Ntoulas et al., 2006]
- Web spamming techniques
  - Link Stuffing
  - Keyword Stuffing
  - PageRank becomes the target of many spamming techniques

Haixuan Yang, Irwin King, and Michael R. Lyu
DiffusionRank: A Possible Penicillin for Web Spamming
SIGIR2007, Amsterdam
Web is easily manipulated for commercial gains

- About 70% of all pages in the .biz domain are spam [Alexandros Ntoulas et al., 2006]
- About 35% of the pages in the .us domain belong to spam category [Alexandros Ntoulas et al., 2006]

Web spamming techniques

- Link Stuffing
- Keyword Stuffing

PageRank becomes the target of many spamming techniques
PageRank

- Calculate the importance of a Web page based on the link structure
- Recursively defined by the in-coming links

\[ x_i = \sum_{(j,i) \in E} a_{ij} x_j \quad a_{ij} = \frac{1}{d^+(j)} \]

\[ x = Ax \quad x = [(1 - \alpha)g1^T + \alpha A]x \]

Issues

Haixuan Yang, Irwin King, and Michael R. Lyu

DiffusionRank: A Possible Penicillin for Web Spamming
PageRank

- Calculate the importance of a Web page based on the link structure
- Recursively defined by the in-coming links

\[ x_i = \sum_{(j,i) \in E} a_{ij} x_j \quad a_{ij} = 1/d^+(j) \]

\[ x = Ax \]

\[ x = [(1 - \alpha)g1^T + \alpha A]x \]
PageRank

- Calculate the importance of a Web page based on the **link structure**
- Recursively defined by the **in-coming links**

\[ x_i = \sum_{(j,i) \in E} a_{i,j} x_j \quad a_{ij} = 1/d^+(j) \]

\[ x = Ax \quad x = [(1 - \alpha)g1^T + \alpha A]x \]
PageRank

- Calculate the importance of a Web page based on the link structure
- Recursively defined by the in-coming links

\[ x_i = \sum_{(j,i) \in E} a_{i,j} x_j \quad a_{ij} = \frac{1}{d^+(j)} \]
\[ x = Ax \]
\[ x = [(1 - \alpha)g1^T + \alpha A]x \]

- Issues
  - Incomplete information of the Web structure (previous work)
  - Susceptible to Web spamming
PageRank

- Calculate the importance of a Web page based on the link structure
- Recursively defined by the in-coming links
  \[ x_i = \sum_{(j,i) \in E} a_{ij} x_j \quad a_{ij} = 1/d^+(j) \]
  \[ x = Ax \]
  \[ x = [(1 - \alpha)g1^T + \alpha A]x \]

- Issues
  - Incomplete information of the Web structure (previous work)
  - Susceptible to Web spamming
PageRank

- Calculate the importance of a Web page based on the link structure
- Recursively defined by the in-coming links

\[ x_i = \sum_{(j,i) \in E} a_{i,j} x_j \quad a_{ij} = 1/d^+(j) \]
\[ x = Ax \]

- Issues
  - Incomplete information of the Web structure (previous work)
  - Susceptible to Web spamming
An Example of Web Manipulation

Perfect World

\[
\begin{align*}
    x_i &= \sum_{(j,i) \in E} 0.85 a_{ij} x_j + 0.15/n \\
    a_{ij} &= 1/d^+(j)
\end{align*}
\]

PageRank Results:

2 > 5 > 3 > 4 > 1 > 6

Real World

Node 1’s value can be increased greatly!

PageRank Results:

1 > 2 > 5 > 3 > 4 > 6

Haixuan Yang, Irwin King, and Michael R. Lyu

DiffusionRank: A Possible Penicillin for Web Spamming

SIGIR2007, Amsterdam
Spam, Spam, Spam Everywhere

Why Spamming Is Easy?

- Web is overly democratic—All pages are treated equal
- Input independent—For any given non-zero initial input, the iteration will converge to the same stable distribution

Web Spam Is Easy
PageRank can be easily manipulated by having link stuffing!
Why Spamming Is Easy?

- Web is overly **democratic**—All pages are treated equal.
- Input **independent**—For any given non-zero initial input, the iteration will converge to the same stable distribution.

Web Spam Is Easy

PageRank can be easily manipulated by having link stuffing!
Variations of PageRank

- PageRank [L. Page et al., 1998]
- Ranking the Web frontier [N. Eiron et al., 2004]
- Generalize PageRank by damping functions [R. A. Baeza-Yates et al., 2006]
- TrustRank [Z. Gyöngyi et al., 2004]
Variations of PageRank

- PageRank [L. Page et al., 1998]
- Ranking the Web frontier [N. Eiron et al., 2004]
- Generalize PageRank by damping functions [R. A. Baeza-Yates et al., 2006]
- TrustRank [Z. Gyöngyi et al., 2004]
Variations of PageRank

- PageRank [L. Page et al., 1998]
- Ranking the Web frontier [N. Eiron et al., 2004]
- Generalize PageRank by damping functions [R. A. Baeza-Yates et al., 2006]
- TrustRank [Z. Gyöngyi et al., 2004]
Variations of PageRank

- PageRank [L. Page et al., 1998]
- Ranking the Web frontier [N. Eiron et al., 2004]
- Generalize PageRank by damping functions [R. A. Baeza-Yates et al., 2006]
- TrustRank [Z. Gyöngyi et al., 2004]
TrustRank

Main characteristics

- The seed set is selected according to the inverse PageRank.
- The biased PageRank is employed by setting $g$ to be the distribution shared by all the trusted pages found in the first part.
- Advantage—it can combat Web spam.
- Disadvantage—it does not follow the actual users’ behaviors by setting a biased $g$.

$$x = [(1 - \alpha)g1^T + \alpha A]x \ (1 - \alpha)g1^T + \alpha A$$
TrustRank

- **Main characteristics**
  - The seed set is selected according to the inverse PageRank
  - The biased PageRank is employed by setting \( g \) to be the distribution shared by all the trusted pages found in the first part

- **Advantage**—can combat Web spam
- **Disadvantage**—it does not follow the actual users’ behaviors by setting a biased \( g \)

\[
x = [(1 - \alpha)g1^T + \alpha A]x \quad (1 - \alpha)g1^T + \alpha A
\]
TrustRank

- Main characteristics
  - The seed set is selected according to the inverse PageRank
  - The biased PageRank is employed by setting $g$ to be the distribution shared by all the trusted pages found in the first part

- Advantage—can combat Web spam

- Disadvantage—it does not follow the actual users’ behaviors by setting a biased $g$

$$x = [(1 - \alpha)g1^T + \alpha A]x \ (1 - \alpha)g1^T + \alpha A$$

Haixuan Yang, Irwin King, and Michael R. Lyu

DiffusionRank: A Possible Penicillin for Web Spamming
TrustRank

- **Main characteristics**
  - The seed set is selected according to the inverse PageRank.
  - The biased PageRank is employed by setting $g$ to be the distribution shared by all the trusted pages found in the first part.

- **Advantage**—can combat Web spam
  - **Disadvantage**—it does not follow the actual users’ behaviors by setting a biased $g$

\[
x = [(1 - \alpha) g^T + \alpha A] x \quad (1 - \alpha) g^T + \alpha A
\]
TrustRank

- Main characteristics
  - The seed set is selected according to the inverse PageRank
  - The biased PageRank is employed by setting $g$ to be the distribution shared by all the trusted pages found in the first part

- Advantage—can combat Web spam

- Disadvantage—it does not follow the actual users’ behaviors by setting a biased $g$

$$\mathbf{x} = [(1 - \alpha)\mathbf{g}\mathbf{1}^T + \alpha\mathbf{A}]\mathbf{x} \quad (1 - \alpha)\mathbf{g}\mathbf{1}^T + \alpha\mathbf{A}$$
Heat Diffusion Model

Assumptions

- Pages are not equal
- Different initial temperature distributions will give rise to different temperature distributions after a fixed time period

Haixuan Yang, Irwin King, and Michael R. Lyu
DiffusionRank: A Possible Penicillin for Web Spamming
SIGIR2007, Amsterdam
**Variations of PageRank**

### Heat Diffusion Model

#### Assumptions

- Pages are **not equal**
- Different initial temperature distributions will give rise to **different temperature distributions** after a fixed time period

---

**Haixuan Yang, Irwin King, and Michael R. Lyu**  
**SIGIR2007, Amsterdam**

**DiffusionRank: A Possible Penicillin for Web Spamming**
Our Contributions

- Propose a novel DiffusionRank
  - Provide a new viewpoint on ranking problems
  - Use random graphs
- Theoretically we show that DiffusionRank generalizes PageRank
  - When the thermal conductivity tends to infinity, DiffusionRank becomes PageRank
  - A finite thermal conductivity setting makes DiffusionRank have the effect of anti-spam
Our Contributions

- Propose a novel DiffusionRank
  - Provide a new viewpoint on ranking problems
  - Use random graphs
- Theoretically we show that DiffusionRank generalizes PageRank
  - When the thermal conductivity tends to infinity, DiffusionRank becomes PageRank
  - A finite thermal conductivity setting makes DiffusionRank have the effect of anti-spam
Our Contributions

- Propose a novel DiffusionRank
- Provide a new viewpoint on ranking problems
- Use random graphs
- Theoretically we show that DiffusionRank generalizes PageRank
  - When the thermal conductivity tends to infinity, DiffusionRank becomes PageRank
  - A finite thermal conductivity setting makes DiffusionRank have the effect of anti-spam
Our Contributions

- Propose a novel DiffusionRank
  - Provide a new viewpoint on ranking problems
  - Use random graphs

- Theoretically we show that DiffusionRank generalizes PageRank
  - When the thermal conductivity tends to infinity, DiffusionRank becomes PageRank
  - A finite thermal conductivity setting makes DiffusionRank have the effect of anti-spam
Our Contributions

- Propose a novel DiffusionRank
  - Provide a new viewpoint on ranking problems
  - Use random graphs
- Theoretically we show that DiffusionRank generalizes PageRank
  - When the thermal conductivity tends to infinity, DiffusionRank becomes PageRank
  - A finite thermal conductivity setting makes DiffusionRank have the effect of anti-spam
Our Contributions

- Propose a novel DiffusionRank
  - Provide a new viewpoint on ranking problems
  - Use random graphs
- Theoretically we show that DiffusionRank generalizes PageRank
  - When the thermal conductivity tends to infinity, DiffusionRank becomes PageRank
  - A finite thermal conductivity setting makes DiffusionRank have the effect of anti-spam
On DiffusionRank

DiffusionRank Defined

- **Undirected Graph**—the amount of the heat flow from \( j \) to \( i \) is proportional to the heat difference between \( i \) and \( j \)

\[
f(1) = e^{\gamma H} f(0), \quad H_{ij} = \begin{cases} 
-d(v_j), & j = i, \\
1, & (v_j, v_i) \in E, \\
0, & \text{otherwise}.
\end{cases}
\]

- **Directed Graph**—there is extra energy imposed on the link \((j, i)\) such that the heat flow only from \( j \) to \( i \) if there is no link \((i, j)\)

\[
f(1) = e^{\gamma H} f(0), \quad H_{ij} = \begin{cases} 
-1, & j = i, \\
1/d_j, & (v_j, v_i) \in E, \\
0, & \text{otherwise}.
\end{cases}
\]

- **Randomized Directed Graph**—the heat flow is proportional to the probability of the link \((j, i)\)

\[
f(1) = e^{\gamma R} f(0), \quad R_{ij} = \begin{cases} 
-1, & j = i, \\
p_{ji}/RD^+(v_j), & \text{otherwise},
\end{cases}
\]
On DiffusionRank

DiffusionRank Defined

- **Undirected Graph**—the amount of the heat flow from \( j \) to \( i \) is proportional to the heat difference between \( i \) and \( j \)

\[
\mathbf{f}(1) = e^{\gamma \mathbf{H}}\mathbf{f}(0), \quad H_{ij} = \begin{cases} 
-d(v_j), & j = i, \\
1, & (v_j, v_i) \in E, \\
0, & \text{otherwise}.
\end{cases}
\]

- **Directed Graph**—there is extra energy imposed on the link \((j, i)\) such that the heat flow only from \( j \) to \( i \) if there is no link \((i, j)\)

\[
\mathbf{f}(1) = e^{\gamma \mathbf{H}}\mathbf{f}(0), \quad H_{ij} = \begin{cases} 
-1, & j = i, \\
1/d_j, & (v_j, v_i) \in E, \\
0, & \text{otherwise}.
\end{cases}
\]

- **Randomized Directed Graph**—the heat flow is proportional to the probability of the link \((j, i)\)

\[
\mathbf{f}(1) = e^{\gamma \mathbf{R}}\mathbf{f}(0), \quad R_{ij} = \begin{cases} 
-1, & j = i, \\
p_{ji}/RD^+(v_j), & \text{otherwise}.
\end{cases}
\]
DiffusionRank Defined

- **Undirected Graph**—the amount of the heat flow from $j$ to $i$ is proportional to the heat difference between $i$ and $j$

$$f(1) = e^{\gamma H}f(0), \quad H_{ij} = \begin{cases} -d(v_j), & j = i, \\ 1, & (v_j, v_i) \in E, \\ 0, & \text{otherwise}. \end{cases}$$

- **Directed Graph**—there is extra energy imposed on the link $(j, i)$ such that the heat flow only from $j$ to $i$ if there is no link $(i, j)$

$$f(1) = e^{\gamma H}f(0), \quad H_{ij} = \begin{cases} -1, & j = i, \\ 1/d_j, & (v_j, v_i) \in E, \\ 0, & \text{otherwise}. \end{cases}$$

- **Randomized Directed Graph**—the heat flow is proportional to the probability of the link $(j, i)$

$$f(1) = e^{\gamma R}f(0), \quad R_{ij} = \begin{cases} -1, & j = i, \\ p_{ji}/RD^+(v_j), & \text{otherwise}. \end{cases}$$
Issues on DiffusionRank

- Temperature distribution $f(1)$ is the ranking vector

$$f(1) = e^{\gamma R}f(0)$$

$$P = \alpha \cdot A + (1 - \alpha) \cdot g \cdot 1^T$$

$$R = -I + P$$

- Initial temperature $f(0)$ setting:
  - Select $L$ trusted pages with highest inverse PageRank score
  - The temperatures of these $L$ pages are 1, and 0 for all others
Issues on DiffusionRank

- Temperature distribution $f(1)$ is the ranking vector

$$f(1) = e^{\gamma R} f(0)$$

$$R_{ij} = \begin{cases} 
-1, & j = i, \\
p_{ji} / RD^+(v_j), & \text{otherwise.}
\end{cases}$$

$$P = \alpha \cdot A + (1 - \alpha) \cdot g \cdot 1^T$$

$$g = \frac{1}{n} \cdot 1$$

- Initial temperature $f(0)$ setting:
  - Select $L$ trusted pages with highest inverse PageRank score
  - The temperatures of these $L$ pages are 1, and 0 for all others
Issues on DiffusionRank

Temperature distribution $f(1)$ is the ranking vector

$$f(1) = e^{\gamma}Rf(0)$$

$$P = \alpha \cdot A + (1 - \alpha) \cdot g \cdot 1^T$$

$$R = -I + P$$

Initial temperature $f(0)$ setting:

- Select $L$ trusted pages with highest inverse PageRank score
- The temperatures of these $L$ pages are 1, and 0 for all others
Issues on DiffusionRank

- Temperature distribution $f(1)$ is the ranking vector
  
  $$f(1) = e^{\gamma R}f(0)$$

  $$P = \alpha \cdot A + (1 - \alpha) \cdot g \cdot 1^T$$

  $$R = -I + P$$

- Initial temperature $f(0)$ setting:
  - Select $L$ trusted pages with highest inverse PageRank score
  - The temperatures of these $L$ pages are 1, and 0 for all others
Summary of DiffusionRank

- It is not over-democratic—Some pages will be born with a high temperature while others with a low temperature.
- It is not input-independent—Different initial temperature distribution will result in a different temperature distribution after a fixed time period.
- It models actual users’ behaviors—Heat diffusion model is established on a random graph describing actual users’ behaviors.
- It has the advantage of anti-manipulation.
Summary of DiffusionRank

- **It is not over-democratic**—Some pages will be born with a high temperature while others with a low temperature.
- **It is not input-independent**—Different initial temperature distribution will result in a different temperature distribution after a fixed time period.
- **It models actual users’ behaviors**—Heat diffusion model is established on a random graph describing actual users’ behaviors.
- **It has the advantage of anti-manipulation**.
Summary of DiffusionRank

- **It is not over-democratic**—Some pages will be born with a high temperature while others with a low temperature.

- **It is not input-independent**—Different initial temperature distribution will result in a different temperature distribution after a fixed time period.

- **It models actual users’ behaviors**—Heat diffusion model is established on a random graph describing actual users’ behaviors.

- **It has the advantage of anti-manipulation**.
Summary of DiffusionRank

- It is not over-democratic—Some pages will be born with a high temperature while others with a low temperature.
- It is not input-independent—Different initial temperature distribution will result in a different temperature distribution after a fixed time period.
- It models actual users’ behaviors—Heat diffusion model is established on a random graph describing actual users’ behaviors.
- It has the advantage of anti-manipulation.
Computational Considerations

- Approximation of the heat kernel $e^{\gamma R}$

$$f(1) = (I + \frac{\gamma}{N} R)^N f(0) \quad f(1) = e^{\gamma R} f(0)$$

$$(I + \frac{\gamma}{N} R)^N \rightarrow e^{\gamma R} \quad \text{when} \quad N \rightarrow \infty$$

- How to set $N$?

When $\gamma = 1$, $N \geq 30$, the absolute value of real eigenvalues of $(I + \frac{\gamma}{N} R)^N - e^{\gamma R}$ are less than 0.01

When $\gamma = 1$, $N \geq 100$, they are less than 0.005

We use $N = 100$ in the paper
Computational Considerations

- Approximation of the heat kernel $e^{\gamma R}$

$$f(1) = (I + \frac{\gamma}{N} R)^N f(0)$$

$$f(1) = e^{\gamma R} f(0)$$

$(I + \frac{\gamma}{N} R)^N \rightarrow e^{\gamma R}$ when $N \rightarrow \infty$

- How to set $N$?

---

Haixuan Yang, Irwin King, and Michael R. Lyu

DiffusionRank: A Possible Penicillin for Web Spamming

SIGIR2007, Amsterdam
Computational Considerations

- Approximation of the heat kernel $e^{\gamma R}$

$$f(1) = (I + \frac{\gamma}{N} R)^N f(0) \quad f(1) = e^{\gamma R} f(0)$$

$$(I + \frac{\gamma}{N} R)^N \to e^{\gamma R} \quad \text{when} \ N \to \infty$$

- How to set $N$?

Haixuan Yang, Irwin King, and Michael R. Lyu

DiffusionRank: A Possible Penicillin for Web Spamming
On DiffusionRank

Computational Considerations

- Approximation of the heat kernel $e^{\gamma R}$

$$f(1) = (I + \frac{\gamma}{N} R)^N f(0)$$

$$f(1) = e^{\gamma R} f(0)$$

$$(I + \frac{\gamma}{N} R)^N \rightarrow e^{\gamma R} \quad \text{when } N \rightarrow \infty$$

- How to set $N$?
  - When $\gamma = 1$, $N \geq 30$, the absolute value of real eigenvalues of $(I + \frac{\gamma}{N} R)^N - e^{\gamma R}$ are less than 0.01
  - When $\gamma = 1$, $N \geq 100$, they are less than 0.005
  - We use $N = 100$ in the paper

Haixuan Yang, Irwin King, and Michael R. Lyu

DiffusionRank: A Possible Penicillin for Web Spamming

SIGIR2007, Amsterdam
Computational Considerations

- Approximation of the heat kernel $e^{\gamma R}$

\[
\begin{align*}
 f(1) &= \left( I + \frac{\gamma}{N} R \right)^N f(0) \\
 (I + \frac{\gamma}{N} R)^N &\to e^{\gamma R} \quad \text{when } N \to \infty
\end{align*}
\]

- How to set $N$?
  - When $\gamma = 1$, $N \geq 30$, the absolute value of real eigenvalues of $(I + \frac{\gamma}{N} R)^N - e^{\gamma R}$ are less than 0.01
  - When $\gamma = 1$, $N \geq 100$, they are less than 0.005
  - We use $N = 100$ in the paper
Computational Considerations

- Approximation of the heat kernel $e^{\gamma R}$

$$
\begin{align*}
\mathbf{f}(1) &= (\mathbf{I} + \frac{\gamma}{N} \mathbf{R})^N \mathbf{f}(0) \\
\mathbf{f}(1) &= e^{\gamma R} \mathbf{f}(0) \\
(\mathbf{I} + \frac{\gamma}{N} \mathbf{R})^N &\rightarrow e^{\gamma R} \quad \text{when } N \rightarrow \infty
\end{align*}
$$

- How to set $N$?
  - When $\gamma = 1$, $N \geq 30$, the absolute value of real eigenvalues of $(\mathbf{I} + \frac{\gamma}{N} \mathbf{R})^N - e^{\gamma R}$ are less than 0.01
  - When $\gamma = 1$, $N \geq 100$, they are less than 0.005
  - We use $N = 100$ in the paper
Computational Considerations

- Approximation of the heat kernel $e^{\gamma R}$

$$f(1) = (I + \frac{\gamma}{N} R)^N f(0) \quad f(1) = e^{\gamma R} f(0)$$

$$(I + \frac{\gamma}{N} R)^N \to e^{\gamma R} \quad \text{when } N \to \infty$$

- How to set $N$?
  - When $\gamma = 1$, $N \geq 30$, the absolute value of real eigenvalues of
    $$(I + \frac{\gamma}{N} R)^N - e^{\gamma R}$$
    are less than 0.01
  - When $\gamma = 1$, $N \geq 100$, they are less than 0.005
  - We use $N = 100$ in the paper
The Thermal Conductivity, $\gamma$

1. $\gamma = 0$
   The ranking value is most robust to manipulation since no heat is diffused, but the Web structure is completely ignored.

2. $\gamma = \infty$
   DiffusionRank becomes PageRank, it can be manipulated easily.

3. $\gamma = 1$
   DiffusionRank works well in practice.
**Importance of $\gamma$**

1. $\gamma = 0$
   - The ranking value is most robust to manipulation since no heat is diffused, but the Web structure is completely ignored.

2. $\gamma = \infty$
   - DiffusionRank becomes PageRank, it can be manipulated easily.

3. $\gamma = 1$
   - DiffusionRank works well in practice.
Importance of $\gamma$

1. $\gamma = 0$
   The ranking value is most robust to manipulation since no heat is diffused, but the Web structure is completely ignored.

2. $\gamma = \infty$
   DiffusionRank becomes PageRank, it can be manipulated easily.

3. $\gamma = 1$
   DiffusionRank works well in practice.
Applications of DiffusionRank

On DiffusionRank

Applications of DiffusionRank

Group-to-group Relations

Classification

The amount of heat flow from all pages in one department to another

Temperature distribution at time 1:
(0.17, 0.16, 0.17, 0.16, 0.16, 0.12, 0.02,
−0.07, −0.18, −0.22, −0.24, −0.24)

Haixuan Yang, Irwin King, and Michael R. Lyu

DiffusionRank: A Possible Penicillin for Web Spamming

SIGIR2007, Amsterdam
Applications of DiffusionRank

Group-to-group Relations

The amount of heat flow from all pages in one department to another

Classification

Temperature distribution at time 1:
(0.17, 0.16, 0.17, 0.16, 0.16, 0.12, 0.02, −0.07, −0.18, −0.22, −0.24, −0.24)
Experimental Set-Up

- Dataset
  - A toy graph (6 nodes)
  - A middle-size graph (18,542 nodes)
  - A large-size graph crawled from CUHK (607,170 nodes)

- Normalize the rank scores: the sum is the number of nodes

- Parameter settings

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td># iterations</td>
<td>100</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>thermal conductivity</td>
<td>1 (best)</td>
</tr>
<tr>
<td>$L$</td>
<td># trusted pages</td>
<td>1</td>
</tr>
<tr>
<td>$g$</td>
<td>random jump distribution</td>
<td>uniformly (w/o a priori)</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>probability following actual links</td>
<td>0.85</td>
</tr>
</tbody>
</table>
Experiment I

- Tendency of DiffusionRank
  Rank value difference between \( \{A_i\} \) and \( \{B_i\} \): \( \sum |A_i - B_i| \)

- Compare with TrustRank and PageRank on variation of rank values
  When the number of newly added nodes for manipulation is increased

- Compare with TrustRank and PageRank on variation of order difference
  Order difference between \( \{A_i\} \) and \( \{B_i\} \) is measured by the number of all occurrences of the following cases:
  \[ |A_i - A_j| > 0.1 \& (A_i - A_j) \times (B_i - B_j) < 0 \]
  \[ |B_i - B_j| > 0.1 \& (A_i - A_j) \times (B_i - B_j) < 0 \]
Experiment II

- Inverse PageRank scores:
  
  \[ 4 > 3 > 1 > 2 > 6 > 5 \]

- If node 4 has not been manipulated, then node 4 can be trusted, otherwise node 3 should be trusted.
Variation of Rank Values on the Toy Data Set
Experiments

Variation of Rank Values on Two Larger Datasets

Haixuan Yang, Irwin King, and Michael R. Lyu
DiffusionRank: A Possible Penicillin for Web Spamming

SIGIR2007, Amsterdam
Variation of Order Difference on the Larger Dataset

Haixuan Yang, Irwin King, and Michael R. Lyu
DiffusionRank: A Possible Penicillin for Web Spamming
SIGIR2007, Amsterdam
Variation of Order Difference on the Larger Dataset

Haixuan Yang, Irwin King, and Michael R. Lyu

DiffusionRank: A Possible Penicillin for Web Spamming

SIGIR2007, Amsterdam
Conclusion

- DiffusionRank combats Web spamming
- DiffusionRank is a generalization of PageRank when $\gamma = \infty$
- DiffusionRank can be employed to detect group-to-group relations
- DiffusionRank can be used for classification

Future Work

- Investigate the actual users' behaviors for random jumps, $g$
- What are the optimal values for $L$
- Commercial applications $\$$

Haixuan Yang, Irwin King, and Michael R. Lyu

DiffusionRank: A Possible Penicillin for Web Spamming

SIGIR2007, Amsterdam
Looking Into the Crystal Ball...

**Conclusion**

- DiffusionRank combats *Web spamming*
- DiffusionRank is a *generalization* of PageRank when $\gamma = \infty$
- DiffusionRank can be employed to detect *group-to-group relations*
- DiffusionRank can be used for *classification*

**Future Work**

- Investigate the actual users' behaviors for random jumps, $g$
- What are the optimal values for $L$
- Commercial applications $\$$
Looking Into the Crystal Ball...

**Conclusion**
- DiffusionRank combats **Web spamming**
- DiffusionRank is a **generalization** of PageRank when $\gamma = \infty$
- DiffusionRank can be employed to detect **group-to-group relations**
- DiffusionRank can be used for **classification**

**Future Work**
- Investigate the actual users’ behaviors for random jumps, $g$
- What are the optimal values for $L$
- Commercial applications $$

**Haixuan Yang, Irwin King, and Michael R. Lyu**
**SIGIR2007, Amsterdam**
**DiffusionRank: A Possible Penicillin for Web Spamming**
Conclusion

- DiffusionRank combats Web spamming
- DiffusionRank is a generalization of PageRank when $\gamma = \infty$
- DiffusionRank can be employed to detect group-to-group relations
- DiffusionRank can be used for classification

Future Work

- Investigate the actual users' behaviors for random jumps, $g$
- What are the optimal values for $L$
- Commercial applications $\$$

Haixuan Yang, Irwin King, and Michael R. Lyu

DiffusionRank: A Possible Penicillin for Web Spamming

SIGIR2007, Amsterdam
Looking Into the Crystal Ball...

Conclusion

- DiffusionRank combats Web spamming
- DiffusionRank is a generalization of PageRank when $\gamma = \infty$
- DiffusionRank can be employed to detect group-to-group relations
- DiffusionRank can be used for classification

Future Work

- Investigate the actual users’ behaviors for random jumps, $g$
- What are the optimal values for $L$
- Commercial applications $$
Conclusion

- DiffusionRank combats Web spamming
- DiffusionRank is a generalization of PageRank when $\gamma = \infty$
- DiffusionRank can be employed to detect group-to-group relations
- DiffusionRank can be used for classification

Future Work

- Investigate the actual users’ behaviors for random jumps, $g$
- What are the optimal values for $L$
- Commercial applications $$
DiffusionRank combats Web spamming
DiffusionRank is a generalization of PageRank when $\gamma = \infty$
DiffusionRank can be employed to detect group-to-group relations
DiffusionRank can be used for classification

Future Work
- Investigate the actual users’ behaviors for random jumps, $g$
- What are the optimal values for $L$
- Commercial applications $\$$

Haixuan Yang, Irwin King, and Michael R. Lyu
DiffusionRank: A Possible Penicillin for Web Spamming
SIGIR2007, Amsterdam