Evaluating scientific products by means of citation-based models: a first analysis and validation

Dario A. Bini

Dipartimento di Matematica, Università di Pisa
www.dm.unipi.it/~bini
Joint work with Gianna M. Del Corso and Francesco Romani

Due Giorni di Algebra Lineare, Bologna, March 6-7, 2008

The Problem

Research evaluation is a very hot topic.

Journals are often evaluated by means of the Impact Factor (IF) (IF is a measure of popularity more than of quality and reputation)

Publications are ranked according to the importance of the journal where they are published (IF)

The rank of a scholar is determined by the rank of his/her publications

The Problem

Research evaluation is a very hot topic.

Journals are often evaluated by means of the Impact Factor (IF)
(IF is a measure of popularity more than of quality and reputation)

Publications are ranked according to the importance of the journal where they are published (IF)

The rank of a scholar is determined by the rank of his/her publications

The Problem

Research evaluation is a very hot topic.

Journals are often evaluated by means of the Impact Factor (IF)
(IF is a measure of popularity more than of quality and reputation)

Publications are ranked according to the importance of the journal where they are published (IF)

The rank of a scholar is determined by the rank of his/her publications

Citations are the basis of most attempts to assess scientific impact.

Some recent metrics, mostly for ranking journals, are

- Impact Factor (E. Garfield 1972)
- Eigenfactor (C.T. Bergstrom 2007)
- AMS-MR - Mathematical Citation Quotient
- Invariant Method (I. Palacios-Huerta, O. Valij 2004)
- Y-factor (Bollan, Rodriguez, Van de Sompel 2006)
- H-index (J.E. Hirsch 2005)
- New concept of journal (Mizzaro 2003)

Citations are the basis of most attempts to assess scientific impact.

Some recent metrics, mostly for ranking journals, are

- Impact Factor (E. Garfield 1972)
- Eigenfactor (C.T. Bergstrom 2007)
- AMS-MR - Mathematical Citation Quotient
- Invariant Method (I. Palacios-Huerta, O. Valij 2004)
- Y-factor (Bollan, Rodriguez, Van de Sompel 2006)
- H-index (J.E. Hirsch 2005)
- New concept of journal (Mizzaro 2003)

Citations are the basis of most attempts to assess scientific impact.

Some recent metrics, mostly for ranking journals, are

- Impact Factor (E. Garfield 1972)
- Eigenfactor (C.T. Bergstrom 2007)
- AMS-MR - Mathematical Citation Quotient
- Invariant Method (I. Palacios-Huerta, O. Valij 2004)
- Y-factor (Bollan, Rodriguez, Van de Sompel 2006)
- H-index (J.E. Hirsch 2005)
- New concept of journal (Mizzaro 2003)

Citations are the basis of most attempts to assess scientific impact.

Some recent metrics, mostly for ranking journals, are

- Impact Factor (E. Garfield 1972)
- Eigenfactor (C.T. Bergstrom 2007)
- AMS-MR - Mathematical Citation Quotient
- Invariant Method (I. Palacios-Huerta, O. Valij 2004)
- Y-factor (Bollan, Rodriguez, Van de Sompel 2006)
- H-index (J.E. Hirsch 2005)
- New concept of journal (Mizzaro 2003)

Citations are the basis of most attempts to assess scientific impact.

Some recent metrics, mostly for ranking journals, are

- Impact Factor (E. Garfield 1972)
- Eigenfactor (C.T. Bergstrom 2007)
- AMS-MR - Mathematical Citation Quotient
- Invariant Method (I. Palacios-Huerta, O. Valij 2004)
- Y-factor (Bollan, Rodriguez, Van de Sompel 2006)
- H-index (J.E. Hirsch 2005)
- New concept of journal (Mizzaro 2003)

Citations are the basis of most attempts to assess scientific impact.

Some recent metrics, mostly for ranking journals, are

- Impact Factor (E. Garfield 1972)
- Eigenfactor (C.T. Bergstrom 2007)
- AMS-MR - Mathematical Citation Quotient
- Invariant Method (I. Palacios-Huerta, O. Valij 2004)
- Y-factor (Bollan, Rodriguez, Van de Sompel 2006)
- H-index (J.E. Hirsch 2005)
- New concept of journal (Mizzaro 2003)

Citations are the basis of most attempts to assess scientific impact.

Some recent metrics, mostly for ranking journals, are

- Impact Factor (E. Garfield 1972)
- Eigenfactor (C.T. Bergstrom 2007)
- AMS-MR - Mathematical Citation Quotient
- Invariant Method (I. Palacios-Huerta, O. Valij 2004)
- Y-factor (Bollan, Rodriguez, Van de Sompel 2006)
- H-index (J.E. Hirsch 2005)
- New concept of journal (Mizzaro 2003)

Our proposal

In the classical approach the ranking of journals is based on citations

The ranking of papers and authors follows from the rank of the journals where the research is published

We aim to provide an integrated ranking of authors, journals, papers areas, and institutions where each subject contributes to give importance to the other subjects in a suitable way

Mutual reinforcement among the different subjects, depending on the type of relationship

Our proposal

In the classical approach the ranking of journals is based on citations

The ranking of papers and authors follows from the rank of the journals where the research is published

We aim to provide an integrated ranking of authors, journals, papers areas, and institutions where each subject contributes to give importance to the other subjects in a suitable way

Mutual reinforcement among the different subjects, depending on the type of relationship

General principles

Subjects belonging to different classes:

- Papers
- Authors
- Journals
- Areas
- Institutions

All the subjects have a rank

Following the Google model we assume that:

The rank of a subject is the weighted rank given to it by all the subjects that are related to it; the relationship depends on the classes

General principles:

- a paper is important if is cited by important papers, is co-authored by important authors, if is published in an important journal
- an author is important if he/she is co-author of important authors, is author of important papers, publishes in important journals
- a journal is important if it has citations from important journals, publishes important papers, contains paper of important authors

	J	A	P
J	citation	publication	publication
A	publication	co-authorship	authorship
P	publication	authorship	citation

General principles:

- a paper is important if is cited by important papers, is co-authored by important authors, if is published in an important journal
- an author is important if he/she is co-author of important authors, is author of important papers, publishes in important journals
- a journal is important if it has citations from important journals, publishes important papers, contains paper of important authors

	J	A	P
J	citation	publication	publication
A	publication	co-authorship	authorship
P	publication	authorship	citation

General principles:

- a paper is important if is cited by important papers, is co-authored by important authors, if is published in an important journal
- an author is important if he/she is co-author of important authors, is author of important papers, publishes in important journals
- a journal is important if it has citations from important journals, publishes important papers, contains paper of important authors

	J	A	P
J	citation	publication	publication
A	publication	co-authorship	authorship
P	publication	authorship	citation

General principles:

- a paper is important if is cited by important papers, is co-authored by important authors, if is published in an important journal
- an author is important if he/she is co-author of important authors, is author of important papers, publishes in important journals
- a journal is important if it has citations from important journals, publishes important papers, contains paper of important authors

	J	A	P
J	citation	publication	publication
A	publication	co-authorship	authorship
P	publication	authorship	citation

General principles:

- a paper is important if is cited by important papers, is co-authored by important authors, if is published in an important journal
- an author is important if he/she is co-author of important authors, is author of important papers, publishes in important journals
- a journal is important if it has citations from important journals, publishes important papers, contains paper of important authors

	J	A	P
J	citation	publication	publication
A	publication	co-authorship	authorship
P	publication	authorship	citation

We describe and analyze three models

- One-class model, made up by Papers only
- Two-class model, made up by Papers and Authors
- Three-class model, made up by Papers, Authors and Journals

One-class model

n: number of papers; $H = (h_{i,j})$ adjacency matrix $n \times n$

$$h_{i,j} = \begin{cases} 1 & \text{if paper } i \text{ cites paper } j \\ 0 & \text{otherwise} \end{cases}$$

$\mathbf{e} = (1, 1, \ldots, 1)^T$, $\mathbf{d} = H\mathbf{e}$, assume that $d_i \neq 0$, $i = 1 : n$

$$P = (p_{i,j}) := \mathrm{Diag}(\mathbf{d})^{-1} H$$

P is row stochastic, i.e., $P\mathbf{e} = \mathbf{e}$

If P is irreducible, by the Perron-Frobenius theorem it exists unique a vector $\pi > 0$ such that

$$\pi^T = \pi^T P, \qquad \mathbf{e}^T \pi = 1, \quad \pi_j = \sum_{i=1}^{n} \pi_i \frac{h_{i,j}}{d_i}$$

π_i is the rank of paper i

One-class model

n: number of papers; $H = (h_{i,j})$ adjacency matrix $n \times n$

$$h_{i,j} = \begin{cases} 1 & \text{if paper } i \text{ cites paper } j \\ 0 & \text{otherwise} \end{cases}$$

$\mathbf{e} = (1, 1, \ldots, 1)^T$, $\mathbf{d} = H\mathbf{e}$, assume that $d_i \neq 0$, $i = 1 : n$

$$P = (p_{i,j}) := \mathrm{Diag}(\mathbf{d})^{-1} H$$

P is row stochastic, i.e., $P\mathbf{e} = \mathbf{e}$

If P is irreducible, by the Perron-Frobenius theorem it exists unique a vector $\boldsymbol{\pi} > 0$ such that

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T P, \qquad \mathbf{e}^T \boldsymbol{\pi} = 1, \quad \pi_j = \sum_{i=1}^{n} \pi_i \frac{h_{i,j}}{d_i}$$

π_i is the rank of paper i

Problems

- In general P is not irreducible
- There may exist dangling nodes, i.e., papers which cite no papers, so that H may have null rows and P cannot be constructed
- Even though H is irreducible it may be periodic

Remedy:

We introduce a dummy paper which cites and is cited by all the existing papers except by itself

The dummy paper collects the importances of all the papers and redistributes them uniformly to all the subjects by creating no privileges

Problems

- In general P is not irreducible
- There may exist dangling nodes, i.e., papers which cite no papers, so that H may have null rows and P cannot be constructed
- Even though H is irreducible it may be periodic

Remedy:

We introduce a dummy paper which cites and is cited by all the existing papers except by itself

The dummy paper collects the importances of all the papers and redistributes them uniformly to all the subjects by creating no privileges

Example

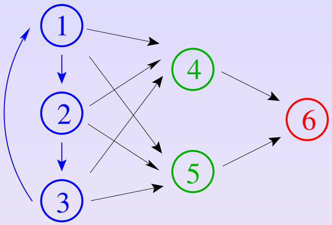
In matrix terms we add one row and one column to the matrix H made of all ones

Example:

$$H = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \longrightarrow \quad \left[\begin{array}{cccc|c} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ \hline 1 & 1 & 1 & 1 & 0 \end{array}\right]$$

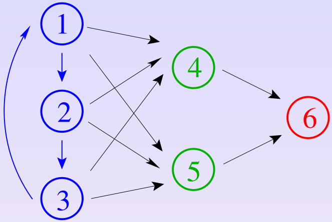
The new matrix is irreducible and aperiodic if $H \neq 0$ (there exist cycles of length 2 and 3)

Example



The adjacency matrix, including the dummy paper, is

$$H = \left[\begin{array}{cccccc|c} 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{array}\right].$$

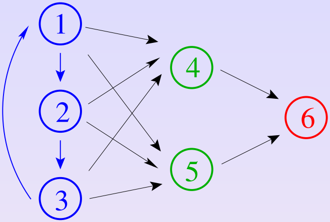


Papers 1,2, and 3 are on the same rank level: except for the dummy paper, they receive one citation each and are inside a cycle.

Papers 4 and 5 receive three citations by papers 1,2,3 and are on the same level but in a higher position with respect to papers 1,2, and 3.

Paper 6 receives only two citations by papers 4 and 5. Therefore, in a model based only on the number of citations, the rank of paper 6 should be inferior to the rank of papers 4 and 5. However, since paper 6 is cited by two papers which are more important than papers 1,2, and 3, one should expect that in our model its rank is higher.

Example

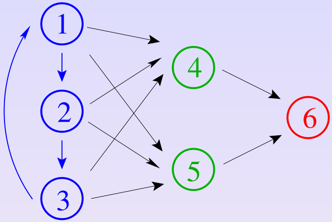


Papers 1,2, and 3 are on the same rank level: except for the dummy paper, they receive one citation each and are inside a cycle.

Papers 4 and 5 receive three citations by papers 1,2,3 and are on the same level but in a higher position with respect to papers 1,2, and 3.

Paper 6 receives only two citations by papers 4 and 5. Therefore, in a model based only on the number of citations, the rank of paper 6 should be inferior to the rank of papers 4 and 5. However, since paper 6 is cited by two papers which are more important than papers 1,2, and 3, one should expect that in our model its rank is higher.

Example



Papers 1,2, and 3 are on the same rank level: except for the dummy paper, they receive one citation each and are inside a cycle.

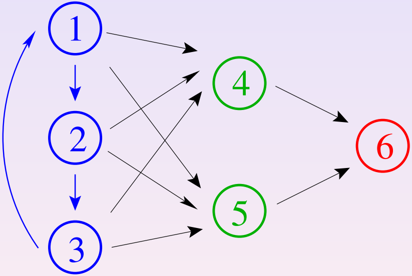
Papers 4 and 5 receive three citations by papers 1,2,3 and are on the same level but in a higher position with respect to papers 1,2, and 3.

Paper 6 receives only two citations by papers 4 and 5. Therefore, in a model based only on the number of citations, the rank of paper 6 should be inferior to the rank of papers 4 and 5. However, since paper 6 is cited by two papers which are more important than papers 1,2, and 3, one should expect that in our model its rank is higher.

In fact, the left eigenvector of $\mathrm{Diag}(H\mathbf{e})^{-1}H$ is

$$\mathbf{p}^T = (0.0784, 0.0784, 0.0784, 0.1176, 0.1176, 0.1764, 0.3529)$$

where $p_1 = p_2 = p_3 < p_4 = p_5 < p_6$ and paper 6 reaches the highest rank as expected.



Theoretical issues: Model validation

What happens about ranking if we add a new citation from a paper to another one?

What happens if we add a new paper which cites another paper?

One should expect that the cited paper increases its rank more than the other papers do.

Can we formalize and quantify this property?

Can we prove it?

Theoretical issues: Model validation

What happens about ranking if we add a new citation from a paper to another one?

What happens if we add a new paper which cites another paper?

One should expect that the cited paper increases its rank more than the other papers do.

Can we formalize and quantify this property?

Can we prove it?

Theoretical issues: Model validation

What happens about ranking if we add a new citation from a paper to another one?

What happens if we add a new paper which cites another paper?

One should expect that the cited paper increases its rank more than the other papers do.

Can we formalize and quantify this property?

Can we prove it?

Theoretical issues: Model validation

What happens about ranking if we add a new citation from a paper to another one?

What happens if we add a new paper which cites another paper?

One should expect that the cited paper increases its rank more than the other papers do.

Can we formalize and quantify this property?

Can we prove it?

Theoretical issues: Model validation

What happens about ranking if we add a new citation from a paper to another one?

What happens if we add a new paper which cites another paper?

One should expect that the cited paper increases its rank more than the other papers do.

Can we formalize and quantify this property?

Can we prove it?

Theoretical results

Let H be an irreducible adjacency matrix

(r,s) such that $h_{r,s} = h_{r,r} = 0$

define $\widehat{H} = (\widehat{h}_{i,j})$ such that $\widehat{h}_{r,s} = 1$, $\widehat{h}_{i,j} = h_{i,j}$ otherwise;

$$H = \begin{array}{cc} \begin{array}{ccccc} & r & & s & \end{array} & \\ \left[\begin{array}{ccccc} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{array}\right] & \begin{array}{c} \\ \leftarrow r \\ \\ \\ \\ \end{array} \end{array} \qquad \widehat{H} = \begin{array}{cc} \begin{array}{ccccc} & r & & s & \end{array} & \\ \left[\begin{array}{ccccc} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{array}\right] & \begin{array}{c} \\ \leftarrow r \\ \\ \\ \\ \end{array} \end{array}$$

Define $P = \mathrm{Diag}(H\mathbf{e})^{-1}H, \quad \widehat{P} = \mathrm{Diag}(\widehat{H}\mathbf{e})^{-1}\widehat{H}$

Theorem

For the left Perron vectors $\boldsymbol{\pi}$ and $\widehat{\boldsymbol{\pi}}$ of P and $\widehat{P}$ it holds

$$\sigma \frac{\widehat{\pi}_r}{\pi_r} \le \frac{\widehat{\pi}_j}{\pi_j} \le \frac{\widehat{\pi}_s}{\pi_s} \quad j = 1, \ldots, n, \qquad \sigma = q/(q+1) < 1$$

where q is the number of ones in the r-th row of H;

$$1 < \frac{\widehat{\pi}_s}{\pi_s} \quad \textit{and} \quad \frac{\widehat{\pi}_j}{\pi_j} < \frac{\widehat{\pi}_s}{\pi_s}, \quad \textit{if} \;\; h_{r,j} \ne 0$$

The paper which receives a new citation has an increase of rank larger than the increase of any other paper

Theorem

For the left Perron vectors $\boldsymbol{\pi}$ and $\widehat{\boldsymbol{\pi}}$ of P and $\widehat{P}$ it holds

$$\sigma\frac{\widehat{\pi}_r}{\pi_r} \le \frac{\widehat{\pi}_j}{\pi_j} \le \frac{\widehat{\pi}_s}{\pi_s} \quad j = 1,\ldots,n, \qquad \sigma = q/(q+1) < 1$$

where q is the number of ones in the r-th row of H;

$$1 < \frac{\widehat{\pi}_s}{\pi_s} \quad \textit{and} \quad \frac{\widehat{\pi}_j}{\pi_j} < \frac{\widehat{\pi}_s}{\pi_s}, \quad \textit{if} \;\; h_{r,j} \ne 0$$

The paper which receives a new citation has an increase of rank larger than the increase of any other paper

Theorem

For the left Perron vectors $\boldsymbol{\pi}$ and $\widehat{\boldsymbol{\pi}}$ of P and $\widehat{P}$ it holds

$$\sigma \frac{\widehat{\pi}_r}{\pi_r} \le \frac{\widehat{\pi}_j}{\pi_j} \le \frac{\widehat{\pi}_s}{\pi_s} \quad j = 1, \ldots, n, \qquad \sigma = q/(q+1) < 1$$

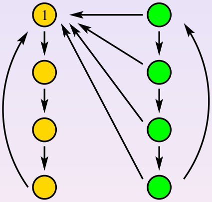
where q is the number of ones in the r-th row of H;

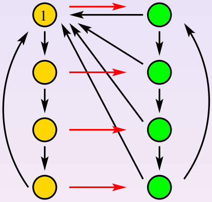
$$1 < \frac{\widehat{\pi}_s}{\pi_s} \quad \text{and} \quad \frac{\widehat{\pi}_j}{\pi_j} < \frac{\widehat{\pi}_s}{\pi_s}, \quad \text{if} \;\; h_{r,j} \ne 0$$

The paper which receives a new citation has an increase of rank larger than the increase of any other paper

One would expect that **all** the papers that receive new citations have an increase of rank which is larger than the increase received by the remaining papers

This is false in general





If we introduce a new paper which contains a citation to paper s, then paper s has an increase of importance strictly greater than the increase of importance obtained by the other papers

Theorem

Let H be an $n \times n$ irreducible adjacency matrix. Let $\widehat{H}$ the matrix obtained by adding a new row and and a new column to H such that $\widehat{h}_{n+1,s} = 1$, $\widehat{h}_{n+1,j} = 0$ for $j \neq s$, $\widehat{h}_{i,n+1} = 0$.
Denote with P and $\widehat{P}$ the stochastic matrices obtained by adding the dummy node to H and to $\widehat{H}$ and normalizing by rows. Then for the left Perron vectors $\boldsymbol{\pi}$ and $\widehat{\boldsymbol{\pi}}$ of P and $\widehat{P}$ it holds

$$\frac{2}{3}\frac{\widehat{\pi}_{n+1}}{\pi_{n+1}} \leq \frac{\widehat{\pi}_j}{\pi_j} < \frac{\widehat{\pi}_s}{\pi_s}, \quad j = 1:n, \quad j \neq s$$

$$\frac{\widehat{\pi}_s}{\pi_s} > 1 + \frac{2}{n}\pi_{n+1}$$

Two-class model

Besides n papers we consider the set of m authors

Consider the adjacency $m \times n$ matrix $K = (k_{i,j})$ such that
$k_{i,j} = \begin{cases} 1 & \text{if author } i \text{ is (co-)author of paper } j \\ 0 & \text{otherwise} \end{cases}$

The matrix $A = KK^T$ is such that $a_{i,j}$ is the number of papers co-authored by authors i and j.

Assume that the dummy paper is written by all the authors

The system is represented by the matrix

$$S = \begin{bmatrix} KK^T & K \\ K^T & H \end{bmatrix}$$

which captures the relationship of authorship and citation among the different subjects of the models.

Example

Consider three papers and three authors such that

$$P1 \to P2 \to P3 \to P1$$

$$A1 \to P1, P2, P3; \quad A2 \to P2; \quad A3 \to P3$$

Then

$$H = \left[\begin{array}{ccc|c} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ \hline 1 & 1 & 1 & 0 \end{array}\right], \quad K = \left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right]$$

$$A = KK^T = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{bmatrix}$$

$$S = \left[\begin{array}{ccc|cccc} 4 & 2 & 2 & 1 & 1 & 1 & 1 \\ 2 & 2 & 1 & 0 & 1 & 0 & 1 \\ 2 & 1 & 2 & 0 & 0 & 1 & 1 \\ \hline 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{array}\right] = \begin{bmatrix} KK^T & K \\ K^T & H \end{bmatrix}$$

In principle S can be scaled by rows in order to make it stochastic and then used as weight matrix in order to assign ranking

However, we prefer to scale the four blocks, $A = KK^T$, K, K^T, H separately in order to make them stochastic, by obtaining $A_{i,j}$, $i,j = 1,2$, and then combine them into a larger stochastic matrix with weights

$$\Gamma = \begin{bmatrix} \gamma_{1,1} & \gamma_{1,2} \\ \gamma_{2,1} & \gamma_{2,2} \end{bmatrix}, \quad \text{row stochastic}$$

$$P = \begin{bmatrix} \gamma_{1,1}A_{1,1} & \gamma_{1,2}A_{1,2} \\ \gamma_{2,1}A_{2,1} & \gamma_{2,2}A_{2,2} \end{bmatrix}$$

The rank vector is defined as $\boldsymbol{\pi}$ such that

$$\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$$

The parameters $\gamma_{i,j}$ determine the amount of importance that the class i transfer to the class j

Each block $A_{i,j}$ determines the distribution of importance from the entries of class i to the entries of class j

Remark: The previous perturbation theorems still hold in this model

Example

For the example shown before, where $P1 \to P2 \to P3 \to P1$ and $A1 \to P1, P2, P3, \quad A2 \to P2, \quad A3 \to P3$

choosing $\Gamma = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$ one gets

$$\begin{array}{ll} \text{author rank:} & (0.49, 0.25, 0.26) \\ \text{paper rank:} & (0.18, 0.23, 0.24) \end{array}$$

Observe that in this model, the importance that a paper receives from authors is proportional to the number of coauthors

This drawback can be overcome by means of a column normalization so that the importance received by the authors is the **average** of the importances of the authors

Example

For the example shown before, where $P1 \to P2 \to P3 \to P1$ and $A1 \to P1, P2, P3, \quad A2 \to P2, \quad A3 \to P3$

choosing $\Gamma = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$ one gets

$$\begin{array}{ll} \text{author rank:} & (0.49, 0.25, 0.26) \\ \text{paper rank:} & (0.18, 0.23, 0.24) \end{array}$$

Observe that in this model, the importance that a paper receives from authors is proportional to the number of coauthors

This drawback can be overcome by means of a column normalization so that the importance received by the authors is the **average** of the importances of the authors

Probabilistic interpretation

The rank obtained in this model is the probability that a virtual reader has to read a paper or to examine an author with the following rules:

- While examining an author A, the reader decides
  - with probability $\gamma_{1,1}$ to examine the co-authors of A; in this case he/she chooses any co-author of A with probability depending on the number of jointpapers
  - with probability $\gamma_{1,2} = 1 - \gamma_{1,1}$ to examine a paper of A; in this case he/she chooses a paper of A with uniform probability (scaled probability in the modified model)
- While examining a paper P, the reader decides
  - with probability $\gamma_{2,1}$ to examine the co-authors of P; in this case he/she chooses any co-author of P with uniform probability;
  - with probability $\gamma_{2,2} = 1 - \gamma_{2,1}$ to examine a paper in the reference list of P chosen with uniform probability (dummy paper included)

Probabilistic interpretation

The rank obtained in this model is the probability that a virtual reader has to read a paper or to examine an author with the following rules:

- While examining an author A, the reader decides
  - with probability $\gamma_{1,1}$ to examine the co-authors of A; in this case he/she chooses any co-author of A with probability depending on the number of joint papers
  - with probability $\gamma_{1,2} = 1 - \gamma_{1,1}$ to examine a paper of A; in this case he/she chooses a paper of A with uniform probability (scaled probability in the modified model)
- While examining a paper P, the reader decides
  - with probability $\gamma_{2,1}$ to examine the co-authors of P; in this case he/she chooses any co-author of P with uniform probability;
  - with probability $\gamma_{2,2} = 1 - \gamma_{2,1}$ to examine a paper in the reference list of P chosen with uniform probability (dummy paper included)

Probabilistic interpretation

The rank obtained in this model is the probability that a virtual reader has to read a paper or to examine an author with the following rules:

- While examining an author A, the reader decides
  - with probability $\gamma_{1,1}$ to examine the co-authors of A; in this case he/she chooses any co-author of A with probability depending on the number of jointpapers
  - with probability $\gamma_{1,2} = 1 - \gamma_{1,1}$ to examine a paper of A; in this case he/she chooses a paper of A with uniform probability (scaled probability in the modified model)
- While examining a paper P, the reader decides
  - with probability $\gamma_{2,1}$ to examine the co-authors of P; in this case he/she chooses any co-author of P with uniform probability;
  - with probability $\gamma_{2,2} = 1 - \gamma_{2,1}$ to examine a paper in the reference list of P chosen with uniform probability (dummy paper included)

Probabilistic interpretation

The rank obtained in this model is the probability that a virtual reader has to read a paper or to examine an author with the following rules:

- While examining an author A, the reader decides
  - with probability $\gamma_{1,1}$ to examine the co-authors of A; in this case he/she chooses any co-author of A with probability depending on the number of jointpapers
  - with probability $\gamma_{1,2} = 1 - \gamma_{1,1}$ to examine a paper of A; in this case he/she chooses a paper of A with uniform probability (scaled probability in the modified model)
- While examining a paper P, the reader decides
  - with probability $\gamma_{2,1}$ to examine the co-authors of P; in this case he/she chooses any co-author of P with uniform probability;
  - with probability $\gamma_{2,2} = 1 - \gamma_{2,1}$ to examine a paper in the reference list of P chosen with uniform probability (dummy paper included)

Probabilistic interpretation

The rank obtained in this model is the probability that a virtual reader has to read a paper or to examine an author with the following rules:

- While examining an author A, the reader decides
  - with probability $\gamma_{1,1}$ to examine the co-authors of A; in this case he/she chooses any co-author of A with probability depending on the number of jointpapers
  - with probability $\gamma_{1,2} = 1 - \gamma_{1,1}$ to examine a paper of A; in this case he/she chooses a paper of A with uniform probability (scaled probability in the modified model)
- While examining a paper P, the reader decides
  - with probability $\gamma_{2,1}$ to examine the co-authors of P; in this case he/she chooses any co-author of P with uniform probability;
  - with probability $\gamma_{2,2} = 1 - \gamma_{2,1}$ to examine a paper in the reference list of P chosen with uniform probability (dummy paper included)

Three-class model

Besides, Authors and Papers we introduce the class of Journals of cardinality q

Together with H and K we introduce the matrices

- $F = (f_{i,j})$, such that $f_{i,j} = 1$ if journal i publishes paper j, $f_{i,j} = 0$ elsewhere
- $G = (g_{i,j})$ such that $g_{i,j} = r$ if author i has published r papers in journal j
- $E = (e_{i,j})$ such that $e_{i,j}$ is the number of citations from papers published in journal i to papers published in journal j

It holds $E = FHF^T, \quad G = FK^T$
The complete adjacency matrix is

$$S = \begin{bmatrix} FHF^T & FK^T & F \\ KF^T & KK^T & K \\ F^T & K^T & H \end{bmatrix}$$

Normalization of blocks and the use of a 3×3 parameter matrix $\Gamma = (\gamma_{i,j})$ lead to a stochastic matrix P of which the left Perron vector $\boldsymbol{\pi}$ represents the ranking of the subjects

Column normalization: blocks (Authors,Papers), (Authors,Journals) and (Papers, Journals) need column normalization

Dummy journal: A dummy journal can be introduced which publishes only the dummy paper; alternatively, the dummy paper is published in all the journals

Probabilistic interpretation: It still holds for the three-class model

Perturbation theorems: The previous perturbation theorems hold valid for this model

Time

A more realistic model which includes time is based on the following modification:

Journals are replicated for each year. This means that, say, SIMAX-2008, SIMAX-2007, SIMAX-2006, ... are considered as **different journals**

This way, the journal evaluation is time dependent as well as the influence that it has on the published papers

We used the CiteSeer dataset, focused primarily on the literature in computer and information sciences, made up by 250,000 authors, 350,000 papers

One class model:

paper	pos.	cit.
Diffie, Hellman- New directions in Cryptography	31	553
Rivest, Shamir, Adleman - Public Key cryptography	3	1218
Bryant -Boolean Functions Manipulation, BDD	1	1636
Kirkpatrick, Gelatt, Vecchi- Simulated Annealing	2	1337
Floyd, Jacobson - TCP/IP Protocol	4	1125
Canny - Computational approach to Edge detection	10	834

Two-class model:

Author	cit	pap.	cit./pap.
Randal Bryant	2615	83	31.5
Sally Floyd	4950	91	54.4
John K. Ousterhout	2214	23	96.3
Luca Cardelli	2112	91	23.2
Van Jacobson	4719	40	118.0
Rakesh Agrawal	4745	83	57.2
Jack J. Dongarra	2799	291	9.6
Raj Jain	1038	116	8.9
Douglas C. Schmidt	2980	329	9.1
Vern Paxson	2735	66	41.4
John Mccarthy	911	41	22.2
Thomas A. Henzinger	3694	176	21.0