# Construção de Micro Rede Social Acadêmica para Análise a Influência dos Artigos e Autores

**Ícaro Araújo Dantas, Li Weigang, Ahmed Abdelfattah Saleh**

TransLab, Departamento de Ciência da Computação da Universidade de Brasília, Brasil

icaro.a.dantas@gmail.com, weigang@unb.br, ahmdsalh@yahoo.com

***Abstract –*** *Google Scholar and other research information network supplies vast information in Internet. In many cases, information that is more detailed is not always available for querying and analysing. This research propose to use Follow Model as engine to rank networks nodes. As the Follow Model has the potential presenting ability for the online social networks. It is adapted with the indexes and models of PageRank, and Inventor Rank to study MSSN in Air Traffic Management field. The case study shows that Follow Model is a robust model and how to ranking a heterogeneous academic network.*

***Resumo –*** *O Google Acadêmico e outras redes de pesquisa por materiais acadêmicos disponibilizam grande quantidade de informação na Internet. Em diversos casos não é possível fazer consulta e análise mais detalhada das informações. Esta pesquisa propõe utilizar Follow Model como ferramente para categorizar os nós de uma rede. Como o Follow Model apresenta grande potencial para trabalharmos com redes sociais online, ele foi adaptado para podermos fazer classificação de nós, assim como os modelos do PageRank, e InventorRank, no estudo da MRSA. O estudo demonstra como o Follow Model é robusto e como podemos melhorar consultas de material acadêmico.*

## 1. Introdução

Cerca de 114 milhões de documentos científicos como livros, artigos, teses, dissertações, documentos técnicos e artigos de trabalho, são hoje acessíveis pela Web [Khabsa e Giles, 2014]. Com números tão expressivos de dados, ferramentas como Google, Scopus e Microsoft Academic Search são frequentemente acessadas, e possuem como grande desafio fornecer a informação desejada.

Tendo em vista o problema de buscar, no meio de grandes bases de dados, a informação que o usuário deseja, nos últimos anos pesquisas foram feitas propondo soluções para estes problemas. Sandes et al. [2012] introduzem o conceito de Follow Model para o desenvolvimento de consultas avançadas em redes sociais. DU et al. [2015] demonstra como fazer a análise de importância de nós em redes heterogêneas.

Continuando com os esforços de buscar sempre a informação mais significativa, este artigo propõe a utilização do Follow Model para categorizar de forma eficiênte os dados.

Para experimentação foi criada uma rede que chamamos, Micro Rede Social Acadêmica (MRSA) com a finalidade de facilitar as demonstrações.

## 2. Micro Rede Social Acadêmica (MRSA)

A partir do artigo "The flow management problem in air traffic control" de Amadeo R. Odoni, foram coletados, manualmente, os dez artigos mais citados que o citaram, e todos os artigos que Amadeo R. Odoni citou em seu trabalho. A partir desse novo conjunto de artigos, novamente foram coletados os trabalhos citados por estes, e os dez mais citados que citaram cada um deles.

Neste artigo representaremos a rede criada como um grafo G = (V, E), onde V é um conjunto de nós, e E o conjunto de arestas. No caso da rede criada os nós do conjunto V serão artigos e/ou atores, e o conjunto E são os relacionamentos existentes

A partir dos dados coletados foram criadas três redes. As duas primeiras redes são redes homogêneas, onde os nós são compostos somente de autores ou artigos. Dentro dessas redes os relacionamentos existentes representam as citações de um nó para outro.

A terceira e última rede é heterogênea, formada pela junção das duas primeiras. O grafo que a representa é G = (V, $V', E, E', E'', C$), onde V é o conjunto de artigos e E as relações entre eles, $V'$ é o conjunto de autores e $E'$ a relação entre eles e $E''$ é o conjunto de relações entre autores e artigos (participação no artigo). No conjunto $E''$ as relações existentes são relações de participação. Por exemplo, se o autor $A'$, participou do artigo A, então existe uma relação ($A'$, A) $\in E''$. O conjunto C é o conjunto que representa relações de coautoria. Esta rede esta representada na Figura 1.

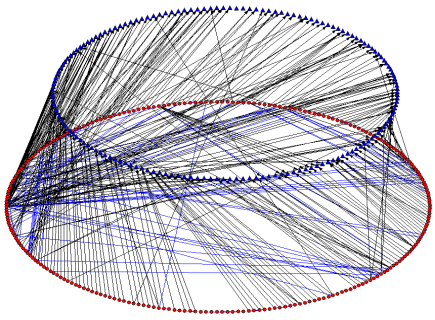

**Figura 1. Rede heterogênea. Triângulos azuis representam artigos. Círculos vermelhos representam autores.**

## 3. Medidas e modelos.

### 3.1. *Follow Model*

*Follow Model* foi criado por Sandes et al. [2012] com a intenção de representar relações em redes sociais online (*OSNs*). Para o modelo criado existem três possíveis relações: *a* é *follower* de *b, b* é *followee* (seguido por) de *a, e a* terceira e última relação quando ambos os nós são seguidores um do outro, chamados então de *r-friends*.

Essas relações são representadas por arestas de um grafo G = (V, E) da seguinte forma. Sendo *a, b* $\epsilon$ V nós de nossa rede, se (*a, b*) $\epsilon$ *E*, então *a* é *follower* de *b* e *b* é *followee* de *a*. Caso também exista uma relação (*b, a*) $\epsilon$ E, os nós também serão chamados de *r-friends*. As funções abaixo demonstram as relações aqui descritas.

$f_{in}(a) = \{v|(v, a)\ \epsilon\ E\}$, é o conjunto de *followers* de *a*, onde $v \epsilon V^*$, $V \rightarrow V^*$, $V^* \subset V$

$f_{out}(a) = \{v|(a, v)\ \epsilon\ E\}$, é o conjunto de *followee* de *a,* onde $v \epsilon V^*$, $V \rightarrow V^*$, $V^* \subset V$

$f_r(a) = f_{out}(a) \cap f_{in}(a)$, é o conjunto de todos *r-friends* de *a*, $V \rightarrow V^*$, $V^* \subset V$

Para cada função $f_{in}, f_{out}, f_r$, podemos também ter variações como, $f_{in}^{p}(a) = \{p(v)|(v,a)\epsilon E\}$, onde $p(v)$ é um atributo do nó $v$. Outra variação é $f_{in}^{w}(a) = \{w(v)|(v,a)\epsilon E\}$, onde $w(v)$ é um atributo do relacionamento entre $a$ e $v$. Outras propriedades desse modelo são a relação inversa, composição e extensão [Weigang et al., 2014, Weigang et al. 2015].

### 3.2. *PageRank*

Criado por Brin [1998], esse algoritmo é um dos motivos de sucesso do buscadores Google. Nesse modelo o peso de um nó é dado pela seguinte equação.

$$PR(A) = (1-d) + d\sum_{j=0}^{n} \frac{PR(T_j)}{C(T_j)} \quad (1)$$

### 3.3. *InventorRank*

Em seu trabalho Du et al. (2015) apresenta uma rede heterogênea de patentes e inventores, constituída de duas sub-redes. $G_I$ é a rede de co-inventores, onde os relacionamentos possuem peso, que é o número de vezes que os inventores trabalharam juntos. $G_{IP}$ é uma rede de patentes onde seus elementos se relacionam com os nós da rede $G_I$. Se um inventor da rede $G_I$ participou em alguma patente de $G_{IP}$ então um relacionamento é criado entre as duas redes.. Seguem as três regras para a classificação.

**Regra 1**: Autores bem classificados tendem a fazer artigos com outros autores bem classificados. Nessa função temos k como o autor em estudo, r é coautor de k, $M_{ii}$ uma matriz composta pelo número de coautorias que cada autor possui com outro autor.

$$R_i(k) = \alpha_{ii}\left[\sum_{r=1}^{n} M_{ii}(k,r)R_i(r)w_{ii} + R_i(k)(1-w_{ii})\right] \quad (2)$$

**Regra 2**: Autores bem classificados geralmente produzem artigos bem classificados. Aqui k é o autor do artigo j, e $M_{pi}$ é a matriz composta pela posição de cada autor na lista de autoria do artigo.

$$R_p(j) = \frac{\alpha_{pi}[\sum_{k=1}^{n} M_{pi}(j,k)R_i(k)}{R_{max}(M)} \quad (3)$$

**Regra 3**: Artigos bem classificadas são feitos por autores bem classificados.

$$R_i(r) = \alpha_{ip}\left[\sum_{j=1}^{n} M_{ip}(r,j)R_p(j)w_{ip} + R_i(r)\left(1-w_{ip}\right)\right] \quad (4)$$

Nessa última função $j$ é o artigo do autor $r$. $M_{ip}$ é a matriz simétrica de $M_{pi}$. Os fatores $\alpha$, presentes nas três funções são utilizados para denotar a importância da regra no momento do cálculo de classificação, e as constantes $w$ são constantes de atenuação.

## 4. Conversão dos modelos para o *Follow Model*

### 4.1. Conversão do *PageRank*

O *PageRank,* calculado de acordo com a Equação 1, pode ser interpretado conforme a equação abaixo:

$$PR(i) = (1-d) + d\left[\frac{s(f_{in}^{p}(i))}{|f_{out}(f_{in}(i))|}\right] \quad (5)$$

### 4.2. Conversão do *InventorRank*

A primeira regra fica da seguinte forma.

$$R_i(k) = \alpha_{ii}\{[s(f_r^w(i).f_r^p(i).w_{ii}) + R_i(k)(1 - w_{ii}).|f_r(k)|]\} \quad (6)$$

Devemos detalhar que a multiplicação entre $f_r^w(i)$ e $f_r^p(i)$ e $w_{ii}$, deve ocorrer elemento por elemento, ou seja, o elemento na posição 1 de $f_r^w(i)$, vezes o elemento na posição 1 de $f_r^p(i)$, vezes $w_{11}$ e assim por diante. As outras regras permanecem da mesma forma.

## 5. Refinamento nos Cálculos

González-Pereira [2010] propõe uma forma de classificar a influência de um periódico chamado *SCImago Journal Rank* (SJR) que utilizamos nesse trabalho para considerarmos a qualidade do periódico na classificação dos nós. Os valores do SCImago variam, em sua maioria, entre os valores 0 e 2, por isso sugerimos que o valor da classificação apresentada nas seções anteriores seja mutiplicada pelo SCImago, como uma espécie de ponderamento. Desse modo as equações ficam da seguinte forma.

- ***PageRank.*** $SR(i) = PR(i) * SCImago\_Rank$ (7)
- ***InventorRank.*** $SR_p(j) = R_p(j) * SCImago\_Rank$ (8)

As regras 1 e 3 do algoritmo *InventorRank não foram modificadas, pois elas não estão envolvidas com a classificação de artigos.*

## 6. Estudo de caso

O experimento demonstra como *Follow Model* pode ser utilizado em algoritmos de classificação, e como a adição do SJR afeta os resultados de classificação.

A MRSA foi utilizada para todos os testes. A constante $d$ do *PageRank* possui o valor 0.5. Para os parâmetros $\alpha_{ii}, \alpha_{ip}$ e $\alpha_{pi}$, foram usados os valores 0.4, 0.4 e 0.2, respectivamente. As constantes $w_{ii}$e $w_{ip}$ foram valoradas, ambas, com 0.5. Os valores da classificação SCImago foram todos buscados em http://www.scimagojr.com/. Periódicos não encontrados foram considerados como tendo a classificação nula e não influenciam no cálculo de posicionamento.

### 6.1. Classificação de autores e artigos sem considerar SJR

O autor AR Odoni aparece, na Tabela 1, como primeiro lugar no *PageRank*, isso ocorre devido ao fato que este é o autor com mais citações dentro da rede, com o total de 13 citações. E Ferons, autor vencedor no modelo *InventorRank,* alcançou o topo devido a sua coautoria com AR Odoni, B Delcairet, H Idris, JP Clarke , WD Hall e B Delcairet, autores que estão também no topo da classificação. Isso demonstra que apesar de AR Odoni ter muitas citações, 89 no total, esse não é um fator muito significativo, pois as pessoas com quem ele trabalha não estão entre os melhores classificados.

**Tabela 1. Top 5 autores, testados por todos os modelos**

| Posição | InventorRank | PageRank |
|---|---|---|
| 1 | E Ferons | AR Odoni |
| 2 | L Kang | D Trivizas |
| 3 | JP Clarke | HN Psaraftis |
| 4 | B Delcairet | EP Gilbo |
| 5 | WD Hall | Dear |

A Tabela 2 mostra como o *InventorRank* se diferencia dos outros modelos. Constata-se como a influência de um autor afeta a classificação de um artigo e vice-versa.

**Tabela 2. Top 5 Artigos classificados pelo *PageRank* e *InventorRank***

| Posição | InventorRank | PageRank |
|---|---|---|
| 1 | Queuing model for taxi-out time estimation. | A dynamic programming approach for sequencing groups of identical jobs. |
| 2 | Collaborative decision making in air traffic flow management | Traffic Control and Transport Planning: A Fuzzy Sets and Neural Networks Approach. |
| 3 | Observations of departure processes at Logan airport to support the development of departure planning tools. | The flow management problem in air traffic control |
| 4 | Input-output modeling and control of the departure process of congested airports. | The traffic flow management-rerouting problem in air traffic control: A dynamic network flow approach. |
| 5 | A comparison of formulations for the single-airport ground-holding problem with banking constraints | The dynamic scheduling of aircraft in the near terminal area. |

## 6.2. Classificação de autores e artigos com SJR.

Na Tabela 3 e 4 vemos que o *InventorRank* reclassificou alguns nós, mostrando que, o peso do SJR dos periódicos pode influênciar sim na classificação dos nós. Sendo que os nós que estão relacionados com bons periódicos subiram na classificação. Vemos também que o PageRank teve uma variação muito maior na sua classificação comparando com a classificação da seção 6.1.

**Tabela 3. Top 5 autores, nos periódicos importantes onde os artigos foram publicados**

| Posição | Nome |
|---|---|
| 1 | AR Odoni |
| 2 | MO Ball |
| 3 | WD Hall |
| 4 | DJ Bertsimas |
| 5 | E Ferons |

**Tabela 4. Top 5 artigos considerando valor de SJR**

| Posição | InventorRank | PageRank |
|---|---|---|
| 1 | Collaborative decision making in air traffic flow management | A dynamic programming approach for sequencing groups of identical jobs. |
| 2 | Observations of departure processes at Logan airport to support the development of departure planning tools. | Applications of operations research in the air transport industry |
| 3 | Queuing model for taxi-out time estimation. | Stochastic and dynamic networks and routing. |

| | | |
|---|---|---|
| 4 | The multi-airport ground-holding problem in air traffic control | The traffic flow management-rerouting problem in air traffic control: A dynamic network flow approach. |
| 5 | Collaborative Decision-Making in Air Transportation | The flow management problem in air traffic control |

## 7. Conclusão

Fazendo um comparativo entre os modelos perceber-se que *InventorRank* é um modelo muito mais robusto, comparado aos outros modelos, quando levamos em consideração o número de informações que ele utiliza para fazer sua classificação, a ponto de não alterar facilmente quando se é adicionado uma nova características a rede, e ao mesmo tempo é flexível, pois podemos definir a importância de cada regra para os cálculos. Além disso vimos como SCImago pode influênciar a classificação dos nós e por isso propomos que seja estudado maneiras de incluir esse fator nos calculos para uma melhor categorização dos nós.

Fica claro também como o *Follow Model* pode se ajustar as necessidades de representação de relações dos diversos modelos existentes [Weigang et al. 2015], de forma simples, possibilitando ao sistema tirar proveito de sua eficiência no momento de classificação.

## Referências


Khabsa, M. and Giles, C. L. (2014). The Number of Scholarly Documents on the Public Web. In *PloS one* 9, no. 5: e93949.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. In *Proceedings of the National academy of Sciences of the United States of America*, *102*(46), 16569-16572.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *In Computer networks and ISDN systems*, 30.1 (1998): 107-117.

Sandes, E. F. O., Weigang, L. and Melo, A. C. M. A. (2012). Logical Model of Relationship for Online Social Networks and Performance Optimizing of Queries. *Proceedings of Web Information Systems Engineering-WISE*, pages 726-736, Paphos, Cyprus, Nuvember 2012. Springer Berlin Heidelberg.

Du, Yong-ping, Yao, Chang-qing, Li, Nan (2015). Using Heterogeneous Patent Network Features to Rank and Discover Influential Inventors. To appear in Frontiers of Information Technology & Electronic Engineering, Doi:10.1631/FITEE.1400394

González-Pereira, B., Guerrero-Bote, V. P., and Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. In *Journal of informetrics*, 4(3), 379-391.

Weigang, L., Sandes, E. F. O., Zheng, J., de Melo, A. C. M. A, and Uden, L. (2014). Querying dynamic communities in online social networks. In *Journal of Zhejiang University – Science C*, 15(2):81–90.

Weigang, L., Dantas, I. A., Saleh, A. A., and Li, D. (2015) Influential Analysis in Micro Scholar Social Networks, to appear in the Proceedings of International Workshop on Social Influence Analysis (SocInf), IJCAI 2015, Buenos Aires.