Descrição dos experimentos realizados

Esta seção visa explicar como foram realizados os experimentos.

O trabalho seguiu dois autores importantes no assunto de Predição de Links.

Liben-Nowell que trouxe um estudo sobre a Predição de Links utilizando a abordagem de ranqueamento.

Munasinge que definiu uma métrica baseada no tempo.

Ambos os trabalhos utilizam bases de co-autoria em particular a Arxiv.

Liben-Nowell utilizou o período de 1994 a 1999 para realização de seus experimentos e as bases: gr-qc, hep-th, astro-ph, cond-mat e hep-ph

Munasige utilizou o período 1997 a 2005 e além da base cond-mat de coautoria ele também trabalho com uma base do facebook.

# Primeiro Experimento

**DataSets**: gr-qc, hep-th, astro-ph, cond-mat e hep-ph (todos cedidos pelo Liben-Nowell)

**Período**: 1994 - 1999

**Objetivo**: Validar a implementação do PredLig com base no ranqueamento.

**Procedimento aplicado**: O Liben-Nowell definiu em seu experimento o conceito de Core. Que é a quantidade de publicações feitas por um autor no período de treino e de teste. Embora ele tenha usado o Core =3, neste experimento foram utilizados o Core=1, e o Core=2. As métricas executas foram: Adamic-Adar, Common Neghbor, Preferential Attachment, Jaccard Coeficcient e Time Score (com as funções de decaimento 0.8, 0.5 e 0.2)

**Resultados**:

|  |  |  |
| --- | --- | --- |
| astro-ph | | |
| Core 1 | Core 2 | Core 3 |
| astroph1.png | astroph2.png | astroph3.png |
| cond-mat | | |
| Core 1 | Core 2 | Core 3 |
| astroph1.png | astroph2.png | astroph3.png |
| qr-qc | | |
| Core 1 | Core 2 | Core 3 |
| astroph1.png | astroph2.png | astroph3.png |
| hep-th | | |
| Core 1 | Core 2 | Core 3 |
| astroph1.png | astroph2.png | astroph3.png |
| hep-ph | | |
| Core 1 | Core 2 | Core 3 |
| astroph1.png | astroph2.png | astroph3.png |

**Análise**:

Observando apenas a métrica TimeScore é possível perceber a função de decaimento com valor igual a 0.2 é melhor que as experimentadas (0.5 e 0.8)

Para as bases cond-mat, gr-qc o TimeScore mostrou-se melhor. O munasinghe trabalhou apenas com cond-mat. Para as outras bases a métrica vencedora é a Adamic-Adar.

# Segundo Experimento

**DataSets**: gr-qc, hep-th, astro-ph, cond-mat e hep-ph (todos cedidos pelo Liben-Nowell e enriquecido com as palavras chaves dos artigos)

**Período**: 1994 - 1999

**Objetivo**: Avaliar a métrica Domain TimeScore.

**Procedimento aplicado**: os DataSets cedidos pelo Liben-Nowell possui o seguinte formato:

1994 M\_SILVA&C\_MUNIZ

1995 M\_OLIVEIRA&C\_MUNHOZ

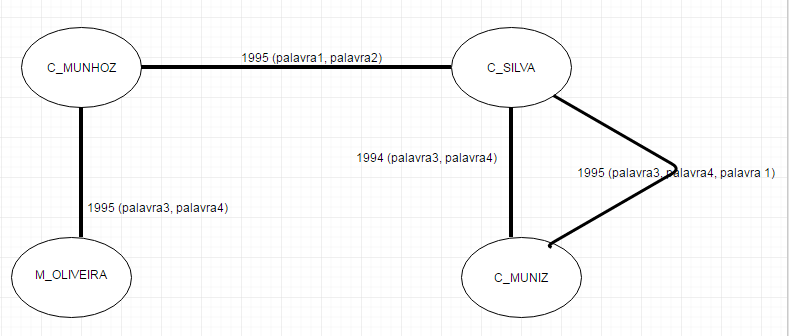
1995 M\_SILVA&C\_MUNHOZ

1995 M\_SILVA&C\_MUNIZ

O que significa que em 1994 existe uma publicação realizada por M\_SILVA E C\_MUNIZ e em 1995 M\_OLIVEIRA publicou com C\_MUNHOZ.

Como é possível observar não existe informações de palavras chaves no dataset. Por isso foi preciso realizar o Enriquecimento da base de Liben-Nowell resgatando do site www.arxiv.com as palavras chaves. Foi criado um algorítmo para resgatar para cada publicação dos arquivos do Nowell localize as palavras chaves.

O resultado foi um grafo similar a este:



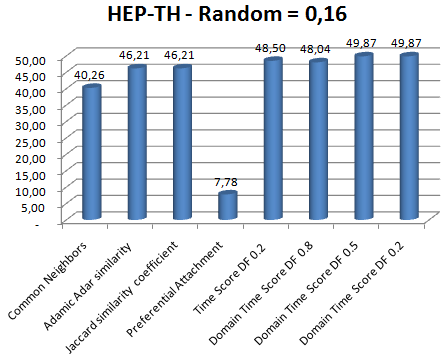
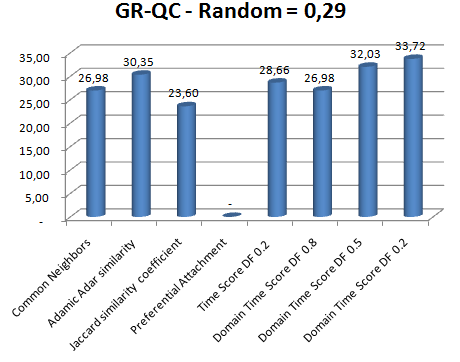
Houve uma mudança na base diante do fato de ter encontrado divergência entre os nomes dos autores encontrados na base do Nowell e o que o site do archive oferecia.

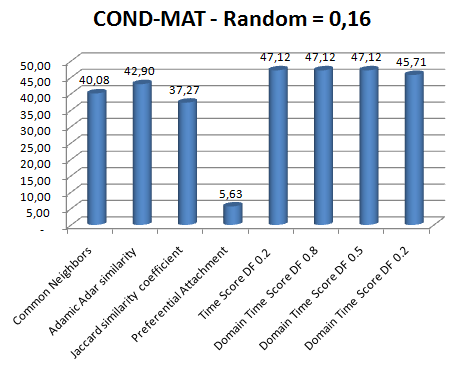
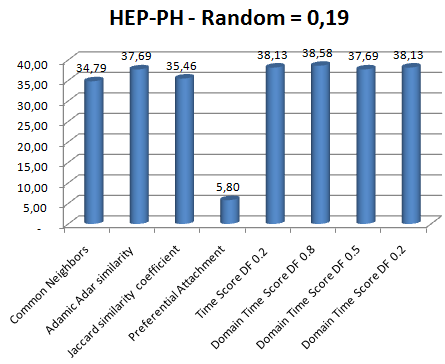
Um exemplo:

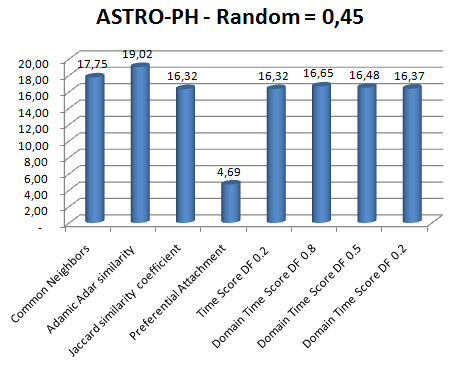
Autor com o nome: M\_C e M\_CEVIT. O algorítmo conseguiu encontrar o artigo escrito por M\_CEVIT, mas M\_C não, logo este artigo foi ignorado para fins de processamento. Por isso executamos com todas as métricas novamente. A suspeita é que no dataset do Nowell não conseguiu interpretar Mirjam Ceviti\xc4\x8d.

Os experimentos foram executados com o Core 3, A função de decaimento do TimeScore igual a 0.2 e para a função decaimento do Domain TimeScore foi utilizado o 0.2 para a função do Tempo e para a função do Domínio utilizamos 0.2, 0.5 e 0.8

**Resultados**:







**Análise**:

Não houve uma definição sobre qual seria a melhor função de decaimento para o Domain Time Score.

Para os datasets cond-mat, gr-qc, hep-th, hep-ph o Domain TimeScore mostrou igual ou superior ao TimeScore porém melhor que as outras métricas. Apenas um resultado ruim para a base astro-ph, no entanto o Time Score acompanhou este resultado ruim.

Independente de datasets, o Domain TimeScore se mostrou igual ou superior ao TimeScore mostrando que a adaptação trouxe melhoria.

# Terceiro Experimento

**DataSets**: gr-qc, hep-th, astro-ph, cond-mat e hep-ph (resgatado do site do Arxiv)

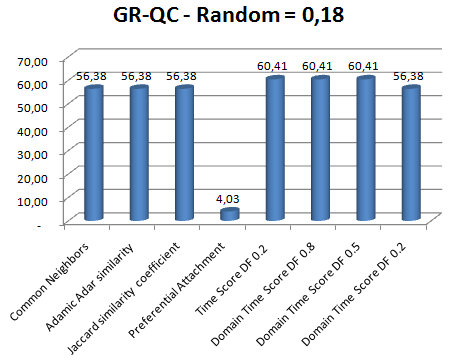
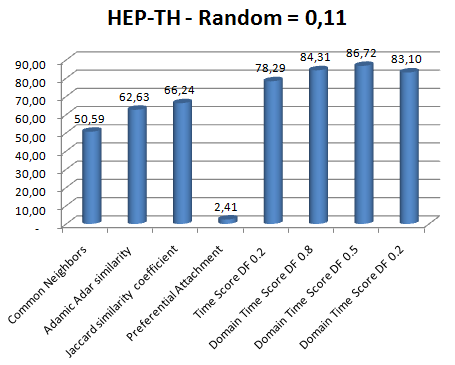
**Período**: 1994 - 1999

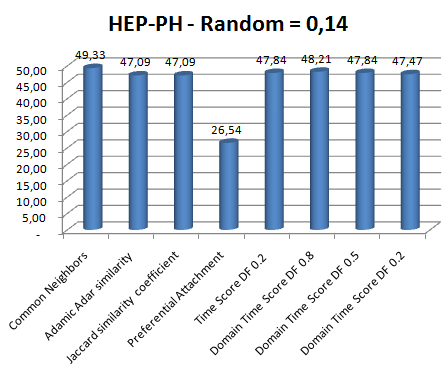
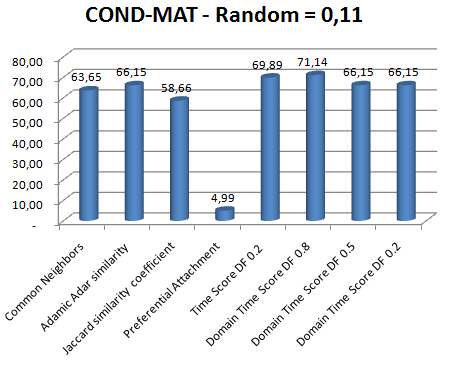
**Objetivo**: Avaliar a métrica Domain TimeScore.

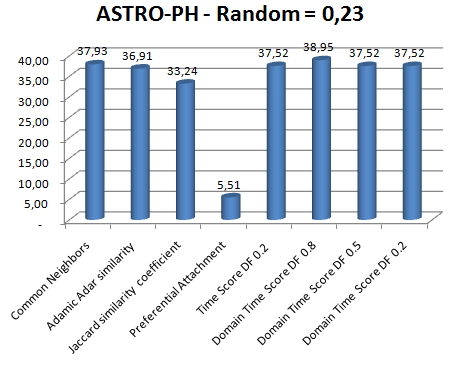
**Procedimento aplicado**: Como foi identificado problemas com relação as base de Liben Nowell. Foi realizado um experimento gerando os grafos diretamente do site Arxiv. A única distinção é que o nome dos autores foram mantidos retirando apenas os caracteres especiais. Por exemplo: Mirjam Ceviti\xc4\x8d foi substituído por MIRJAN\_CEVITIC e assim aconteceu com todos os autores com caracteres especiais no nome.

Os experimentos foram executados com o Core 3, A função de decaimento do TimeScore igual a 0.2 e para a função decaimento do Domain TimeScore foi utilizado o 0.2 para a função do Tempo e para a função do Domínio utilizamos 0.2, 0.5 e 0.8

**Resultados**:



**Análise**:

Não houve uma definição sobre qual seria a melhor função de decaimento para o Domain Time Score.

Diferente do segundo experimento, o único dataset cujo TimeScore e o Domain Time Score não venceram foi o hep-ph (Domain TimeScore ficou em 2 lugar).

# Quarto Experimento

**DataSets**: gr-qc, hep-th, astro-ph, cond-mat e hep-ph (resgatado do site do Arxiv)

**Período**: 2010 - 2015

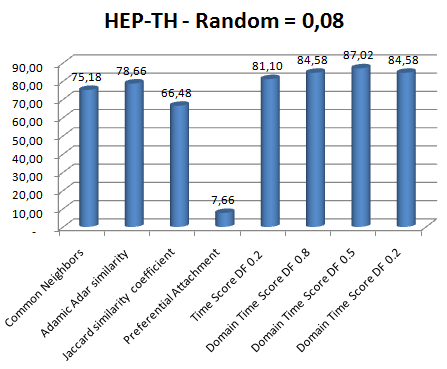
**Objetivo**: Avaliar a métrica Domain TimeScore num grafo com publicações mais recentes.

**Procedimento aplicado**: Foi utilizado as mesmas características utilizada no Terceiro experimento para resgatar do Arxiv os dados com anos mais atuais.

Foi resgatado os dados para as mesmas bases utilizadas nos outros experimentos porém a quantidade de publicações é muito maior gerando grafos cujo o tempo de processamento demoraria em torno de 100 dias para executar (isso observado para apenas uma base).

Com isso foi levantado um alerta sobre a maneira como foi implementado o PredLig.

**Resultados**:



**Análise:**

A única base que conseguimos processar foi o Hep-TH pelo fato de ser a menor. Qualquer outra base demoraria em torno de 100 dias para executar cada função de decaimento do Domain Time Score.

É preciso ser revisto a forma como foi implementado para ver se existe alguma melhoria que podemos utilizar.

# Quinto Experimento

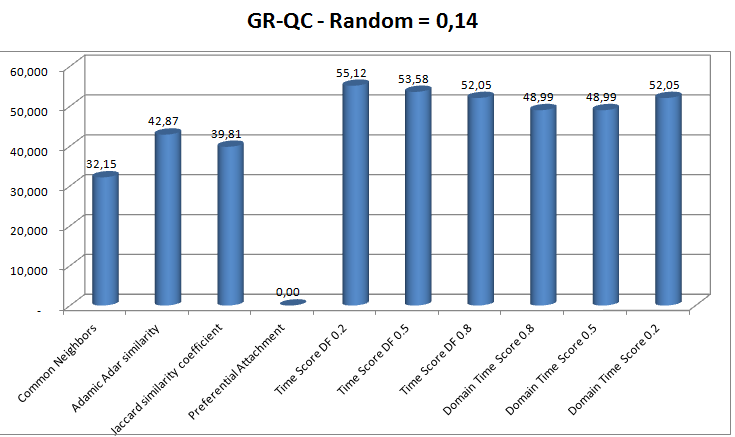
**DataSets**: gr-qc, hep-th, astro-ph, cond-mat e hep-ph (resgatado do site do Arxiv)

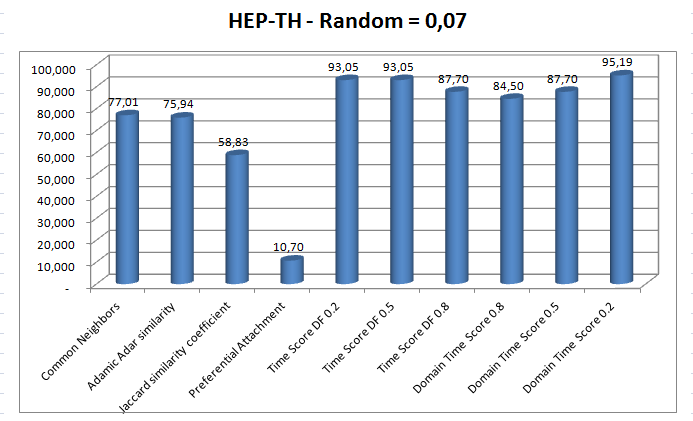
**Período**: 2000 - 2005

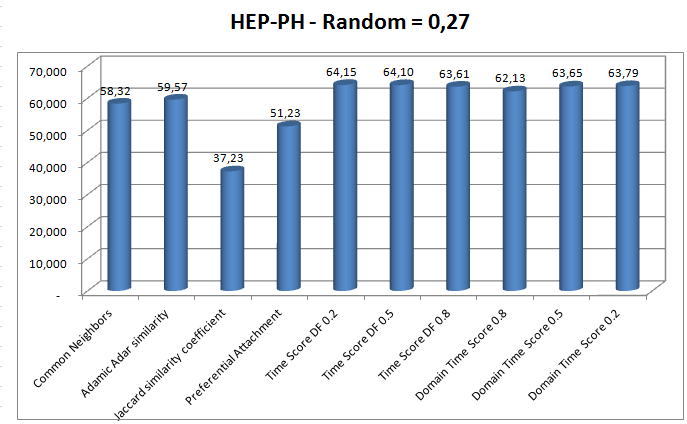
**Objetivo**: Avaliar a métrica Domain TimeScore num grafo com publicações mais recentes.

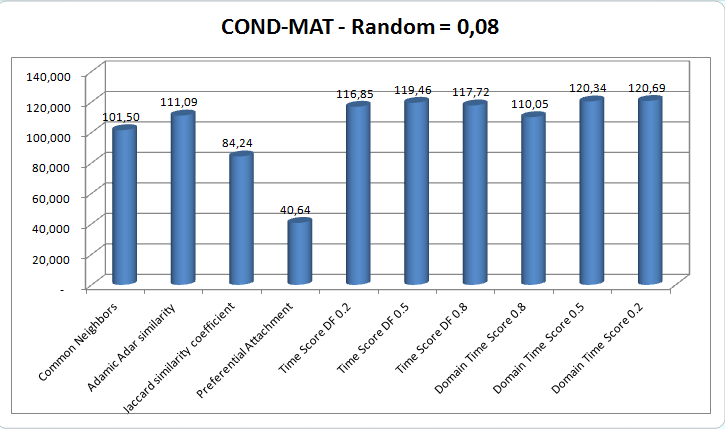
**Procedimento aplicado**: Foi utilizado as mesmas características utilizada no Terceiro experimento para resgatar do Arxiv os dados com os mesmos anos utilizados pelo Munasigne.

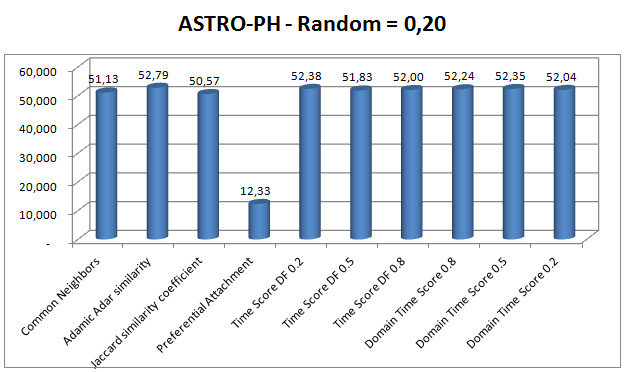
**Resultados:**











**Análise:** Domain based Time Score demonstrou-se ser uma métrica melhor que as demais.

# Sexto Experimento

**DataSets**: gr-qc, hep-th, astro-ph, cond-mat e hep-ph (todos cedidos pelo Liben-Nowell)

**Período**: 1994 - 1999

**Objetivo**: Avaliar grafos ponderados.

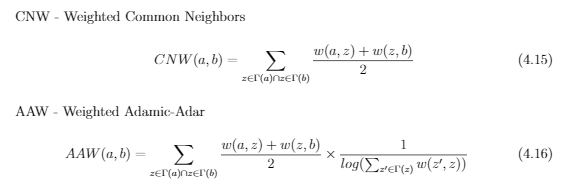
**Procedimento aplicado**: Foi feito a ponderação dos grafos com a métrica WeightedbyTime.



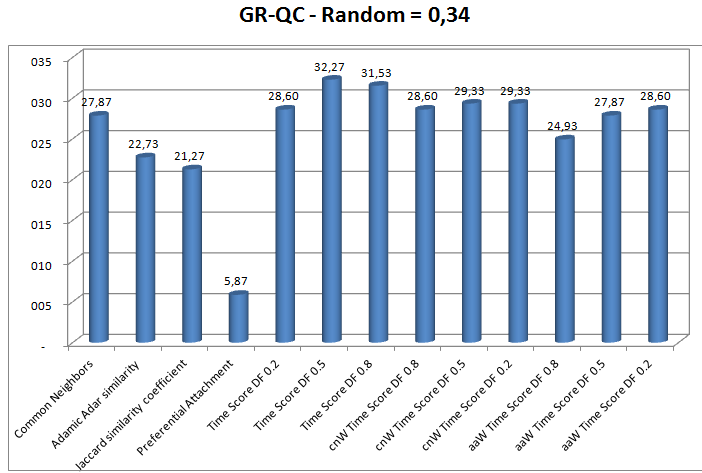
onde o alfa, assim como no Time Score varia de 0 a 1.

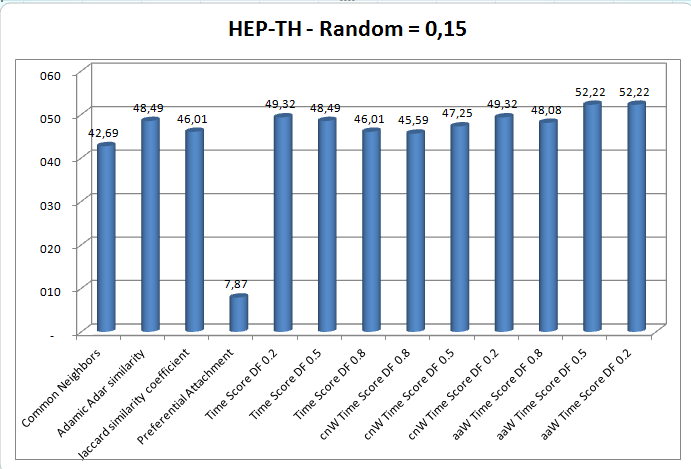
Foram utilizadas as métricas definidas pelo Murata e Moriassu (2007)

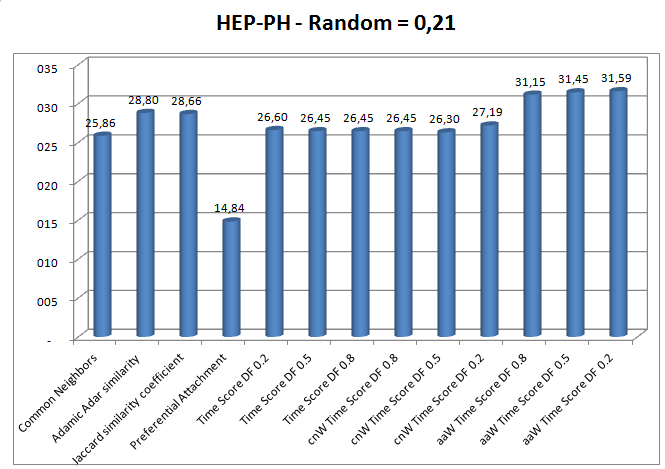
Elas foram:

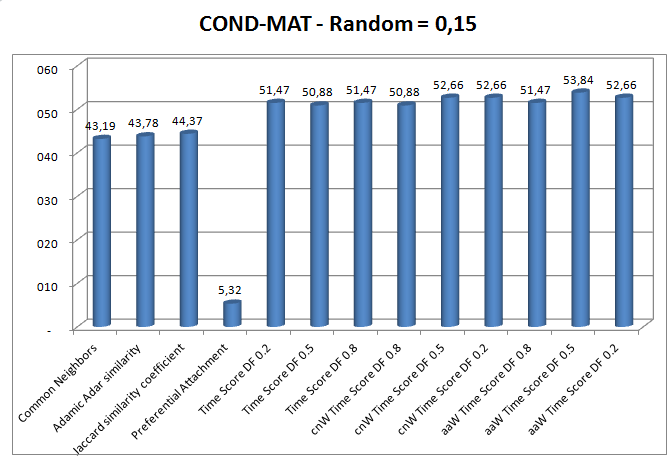


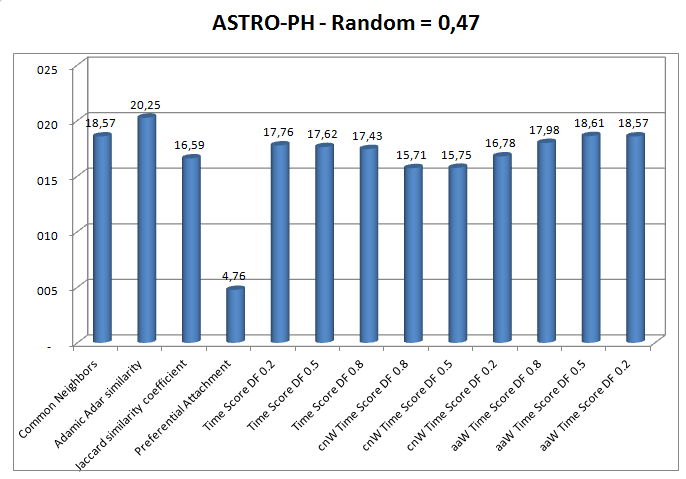
**Resultados:**











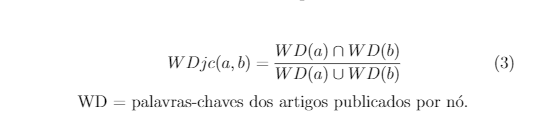
# Sétimo Experimento

**DataSets**: gr-qc, hep-th, astro-ph, cond-mat e hep-ph (todos cedidos pelo Liben-Nowell e enriquecido com as palavras chaves dos artigos)

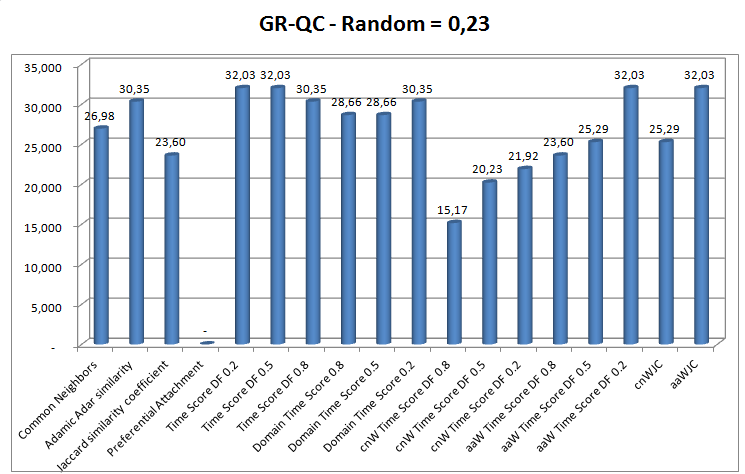
**Período**: 1994 - 1999

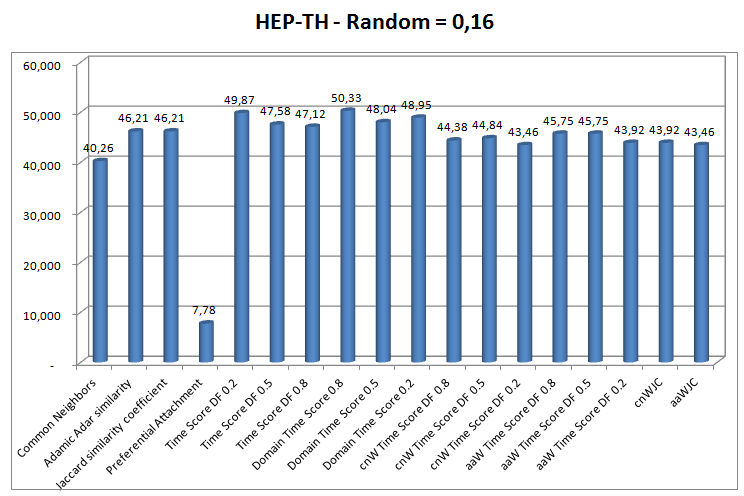
**Objetivo**: Avaliar grafos ponderados com palavras chaves

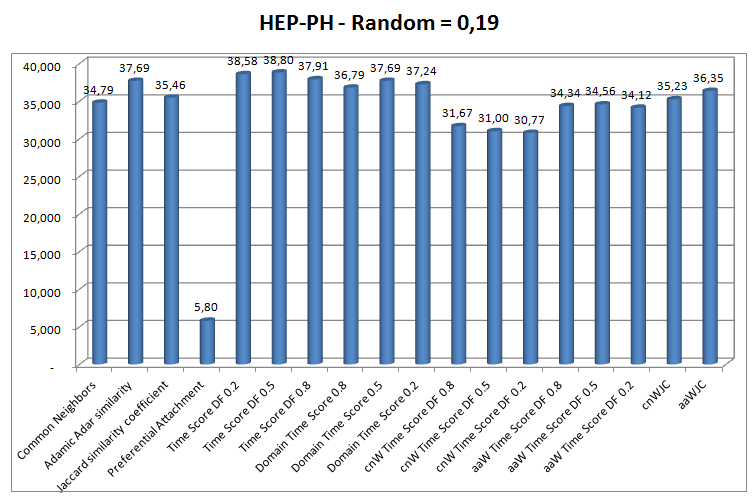
**Procedimento aplicado**: Foi criado mais uma métrica para a utilização das informações de domínio como ponderação do grafo.

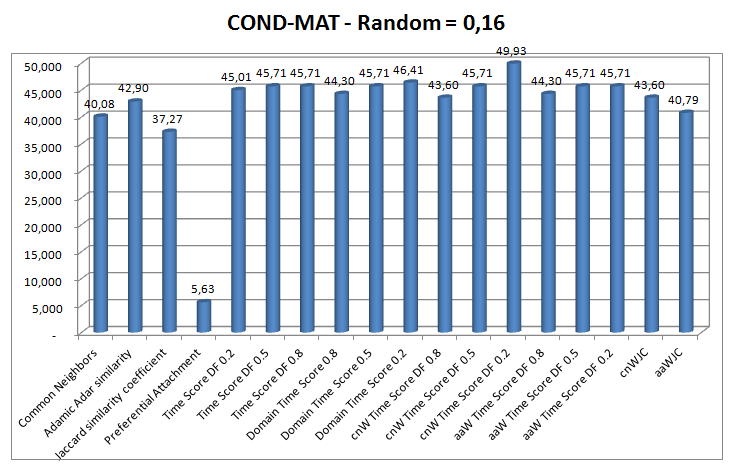


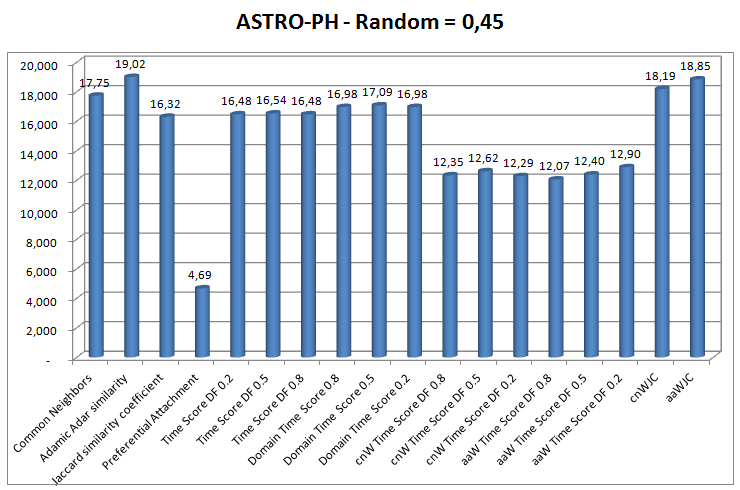
**Resultados:**











# Oitavo Experimento

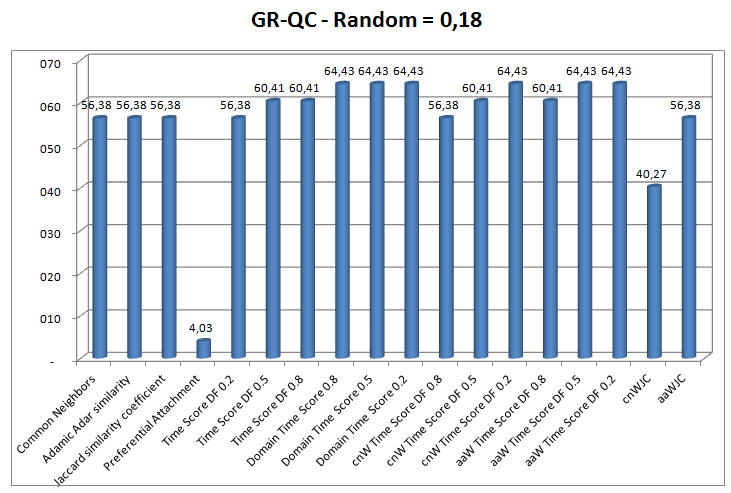
**DataSets**: gr-qc, hep-th, astro-ph, cond-mat e hep-ph (resgatado do site do Arxiv)

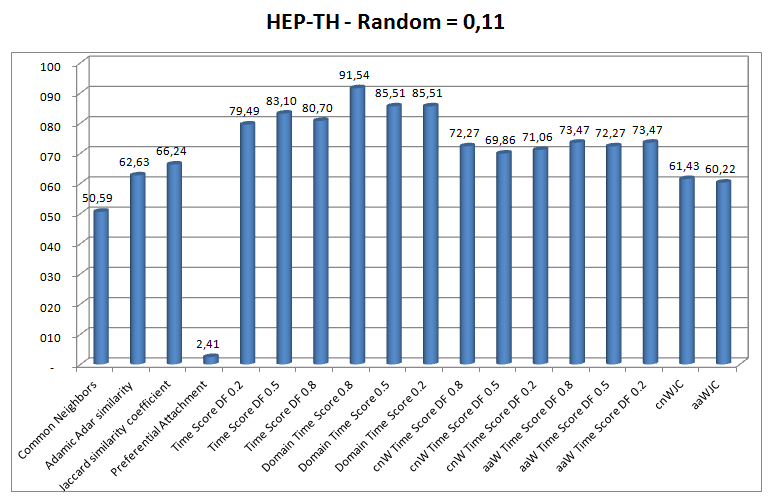
**Período**: 1994 - 1999

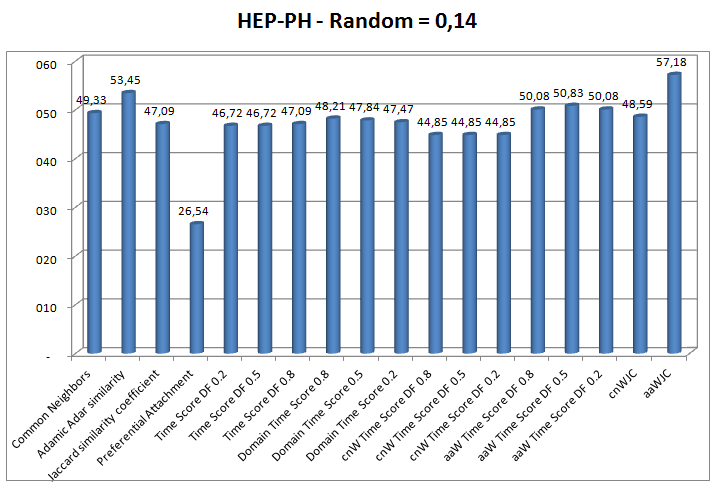
**Objetivo**: Avaliar grafo ponderados.

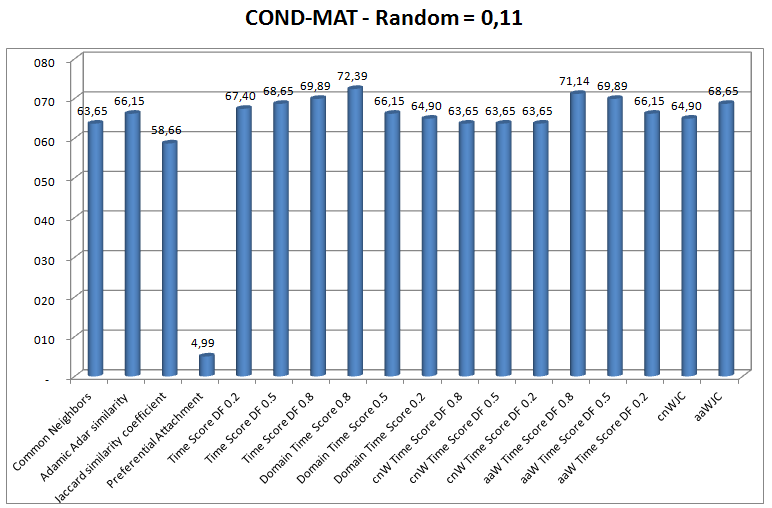
**Procedimento aplicado**: o mesmo procedimento do sétimo experimento foi aplicado neste.

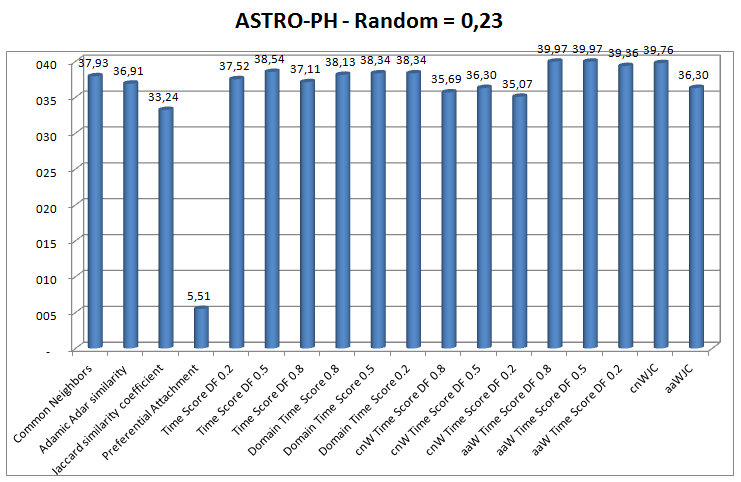
**Resultados:**











# Nono Experimento

**DataSets**: gr-qc, hep-th, astro-ph, cond-mat e hep-ph (resgatado do site do Arxiv)

**Período**: 2000 - 2005

**Objetivo**: Avaliar grafo ponderados.

**Procedimento aplicado**: o mesmo procedimento do sétimo experimento foi aplicado neste.

**Resultados:**

