Análise da Produtividade da Rede Social de Computação do Brasil

Jonice O. Sampaio[1], Fabrício F. Faria[1], Ruben A. Perorazio[1], Evelyn C. de Aquino[2],

[1]Programa de Pós-Graduação em Informática (PPGI) – Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brasil

[2]Departamento dos Métodos Estatísticos (DME) – Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brasil

jonice@dcc.ufrj.br,{firminodefaria,rrpero}@ppgi.ufrj.br,
evelyn.atuaria@gmail.com

Abstract. *Scientific social networks help to understand the dynamics of production of scientific knowledge. The present work aims to analyze the social network computing in Brazil in relation to productivity parameters and metrics of social networks, making a comparison between both. For this purpose we propose a model of a data warehouse able to pass the information space, time and research groups in relation to their productivity and relative metrics social measures.*

Resumo. *Redes sociais científicas ajudam a entender a dinâmica de produção de conhecimento científico. O presente trabalho tem como meta analisar a rede social de computação no Brasil em relação a parâmetros de produtividade e métricas de redes socias, fazendo um comparativo entre ambos. Para este objetivo é proposto um modelo de um Data Warehouse capaz de cruzar as informações espaciais, temporais e dos grupos de pesquisa em relação a sua produtividade e em relação as métricas sociais medidas.*

1. Introdução

Redes sociais científicas tem sido utilizadas como objeto de estudo para compreensão da dinâmica da ciência e para auxiliar a produtividade dos indivíduos que a compõe (NEWMAN, 2001). Como uma rede, ela representa um conjunto de vértices (pessoas) unidos aos pares através de arestas que denotam uma relação que pode ser a produção científica desenvolvida em parceria entre estas pessoas ou a citação de trabalhos de outros cientistas. A relação de amizade também pode ser utilizada, entretanto esta informação pode ser de difícil acesso por não estar explícita em bases científicas.

O presente trabalho se propõe a analisar características sociais e de produtividade científica, e a relação entre ambas, na rede de Ciência da Computação do Brasil no período de 2000 até 2010. Esta rede compreende 961 pesquisadores dos 45 programas de pós-graduação em computação reconhecidos pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) .

Como fonte de dados foi utilizada a plataforma Lattes. Os currículos dos pesquisadores foram extraídos manualmente, dado que esta plataforma não possui nenhum mecanismo automático para captura de dados. Também foram utilizadas informações sobre a avaliação dos programas de pós-graduação feita pela CAPES. Todos estes dados foram extraídos, tratados e integrados em um Data Warehouse (DW). Métricas sociais foram calculadas e adicionados ao DW, permitindo o cruzamento de informações espaciais, temporais e dos programas em relação a sua produtividade e em relação as métricas sociais medidas. O objetivo deste trabalho é analisar a correlação entre as interações científicas e a maturidade dos programas, tentando achar correlações ou padrões entre estes dois fatores.

Na próxima seção será feito um levantamento de trabalhos que avaliaram redes sociais científicas no mundo e no Brasil, na seção três será descrito o processo de captura de dados, tratamento e integração além de detalhar o DW criado. Na seção quatro será feita a análise dos grupos de pesquisa e suas interações sociais além da correlação dos parâmetros de produtividade com métricas sociais. Finalmente, na seção cinco, será feita a discussão e conclusão das informações obtidas.

2. Trabalhos Relacionados

Redes sociais científicas vêm sendo estudadas há pelo menos 40 anos (NEWMAN, 2001). Em (SOLLA PRICE, 1965) o autor construiu uma base de dados manualmente, coletando as citações de 100 artigos de revistas científicas em um intervalo de sete anos. Foram totalizadas 91 interações neste período. Mesmo com este pequeno conjunto de dados o autor mostra que a rede de citação tem uma distribuição de cauda longa em relação ao grau (com ocorre com a maioria das redes sociais por se tratarem de redes complexas). Dados sobre a produtividade dos cientistas também são apresentados, porém não são correlacionados com métricas da rede social.

No trabalho de (NEWMAN, 2001) foi feita uma avaliação extensa sobre características sociais das redes de co-autoria em redes científicas de computação, biologia, física e medicina em um período de 5 anos (1995-1999). Ele apresenta uma série de propriedades comuns nestas redes, como o alto índice de clusterização, que indica que provavelmente a rede tenha sido forma por cientistas apresentando seus colaboradores uns aos outros. Em (NEWMAN, 2004) o autor volta a analisar redes de co-autoria para identificar uma série de propriedades estatísticas, procurando identificar padrões que sejam comuns entre estas redes.

As redes do trabalho de Newman também se comportam-se como um "small world" em que a distância média entre os cientistas através de uma linha de colaboradores intermédios varia logaritmicamente com o tamanho da comunidade. A distância é de apenas cinco ou seis passos para ir de um cientista escolhido

aleatoriamente em um grupo para outro cientista qualquer. Apesar da semelhança, as redes apresentam alguma diferenças entre si, como o grau que é maior na rede de física do que nas outras redes ou o coeficiente de clusterização que é menor nas redes biológicas. Estas diferenças justificam o presente trabalho, uma vez que não podemos generalizar as redes científicas.

A evolução temporal das redes é apresentada no trabalho de (BARABÁSI *et al.*, 2002). Os autores criaram redes de co-autoria a partir de artigos de matemática e neurociências em um período de oito anos. Para o entendimento temporal destas redes eles utilizaram três pilares principais: o primeiro baseado em medidas empíricas para definir a topologia da própria rede bem como sua evolução temporal; o segundo pilar é um modelo proposto para explicar a evolução temporal das redes; o terceiro pilar é baseado em simulação numérica para detectar o comportamento das métricas que não puderam ser determinadas analiticamente. A principal conclusão dos autores é que as métricas utilizadas para caracterizar uma rede (como diâmetro, coeficiente de clusterização e grau) são totalmente dependentes do tempo.

Alguns trabalhos com análises de rede sociais científicas brasileiras podem ser vistas em (STRÖELE *et al.*, 2011) e (MENEZES *et al.*, 2008). STRÖELE et al apresenta uma abordagem baseada em técnicas de mineração de dados para identificar ligações intra e inter grupos de pessoas com perfis semelhantes em Ciência da Computação. Já Menezes et al, faz um estudo comparativo com a rede social de Computação do Brasil com outras redes sociais de computação de outros países através da comparação de métricas.

No trabalho de (MONCLAR, R.S., *et al.*, 2011) é apresentado um framework para análises de redes sociais baseado em técnicas de Data Warehouse para redes de co-autoria e participação em projetos do Instituto Nacional de Ciência e Tecnologia (INCT) do câncer. Este framework é utilizado em (MONCLAR, RAFAEL STUDART; OLIVEIRA, JONICE; *et al.*, 2011), onde é feito a análise do fluxo de informação ao longo do tempo entre os integrantes do INCT do câncer. Na tese de (FREIRE, VINICIUS P., 2010) é feito um proposta para comparação entre programas de Ciência da Computação, entretanto, focado em características individuais dos pesquisadores e não colaboração entre eles.

3. Metodologia

Como o objetivo do presente trabalho é analisar características de produtividade em redes sociais científicas e compará-las com métricas sociais para descrever o comportamento da rede, é fundamental a construção de um mecanismo que possa armazenar os dados coletados no experimento e que possibilite a combinação e filtragem para análise. Estes mecanismos será descrito na sub-seção 3.3. As fontes de dados e as métricas sociais que foram utilizas serão descritas nas sub-seções 3.1 e 3.2 respectivamente.

3.1 Fonte de dados

Foram utilizadas duas fontes de dados para construção do experimento. Do site da CAPES (2012) foi obtido as informações sobre os 45 programas de pós-graduação em informática *strictu sensu* .Também foram extraídos destes programas, os nomes dos pesquisadores que fazem parte de cada grupo e a classificação atribuída aos grupos (esta classificação é a avaliação trienal dada pela CAPES, que pode ir de 3 a 7).

Uma vez com as informações dos pesquisadores fazem parte de quais grupos, através da plataforma Lattes (2012) foi feito o download dos currículos Lattes de 961 pesquisadores. De cada currículo Lattes foram obtidos os dados pessoais, como sexo, nome de citação e nível (caso seja pesquisador do CNPq, podendo este nível ser 2, 1D, 1C, 1B ou 1A) bem como as suas produções bibliográficas e seus projetos de pesquisas. É importante ressaltar que não existem padrões ou recomendação de como os projetos devem ser preenchidos no Lattes, tornado-o suscetível a erros.

3.2 Métricas coletadas

Para a análise do perfil da rede social de computação no Brasil, foram escolhidas algumas métricas que são de dois tipos: i) relativas a produção científica e ii) relativas a redes sociais. Abaixo segue um resumo de cada uma destas métricas.

- **Métricas de Produção:**

Número de artigos publicados: o número de artigos em congresso ou periódicos que o pesquisador possui. Nos limitamos a fazer uma contagem e não levar em consideração a qualificação atribuída a cada trabalho pois a correlação entre estas duas entidades não é trivial, uma vez que o currículo Lattes não faz referência a nota Qualis (WebQualis, 2012) do congresso ou da revista em que o artigo foi publicado.

  - **Número de projetos de pesquisa:** o número de projetos de pesquisa que um pesquisador possui. Neste caso também foi feita uma contagem simples pois não existem um sistema de qualificação de projetos propriamente dito.

- **Métricas de Redes Sociais:**

  - **Grau**: O número de conexões que um pesquisador tem com outros pesquisadores. Estas conexões aparecem quando dois ou mais pesquisadores possuem um relacionamento de co-autoria ou participam de um projeto em conjunto.

  - **Betweenness**: é uma medida da centralidade de um nó em uma rede. Ele é calculado como a fração de caminhos mais curtos entre pares de nós que passam através do nó de interesse. Betweenness é, em certo sentido, uma medida da influência de um nó em relação a propagação de informações dentro da rede, tendo em vista que vários caminhos na rede possuem este nó em comum.

- **Coeficiente de clusterização**: é uma relação v/w, onde v é o número de arestas (conexões) entre os vizinhos, e w é o número máximo de arestas que poderiam existir entre vizinhos.
- **Closeness**: é um índice que indica a centralidade global de um nó. Nós que são capazes de alcançar outros nós em comprimentos mais curtos de caminho. Essa vantagem estrutural pode ser traduzida em importância. O distanciamento de um nó é definido como a soma das suas distâncias para todos os outros nós, e seu closeness é definido como o inverso do distanciamento. Quanto menor o valor do closeness, mais perto um nó está dos demais no grafo.

3.3 Data Warehouse

Um Data Warehouse (DW) é uma base de dados usada para criação de relatórios ou para realização de análises uma vez que sua estrutura é pensada para este tipo de tarefa (INMON, W. H, 1996). Basicamente um DW é composto de dois tipos de entidades: fatos e dimensões. Dimensões podem ser entendidas como uma coleção de atributos altamente relacionados entre si, já fatos são as medidas propriamente ditas. O modelo utilizado na construção do DW pode ser visto na (Figura 1). Ele compreende seis dimensões e três fatos.

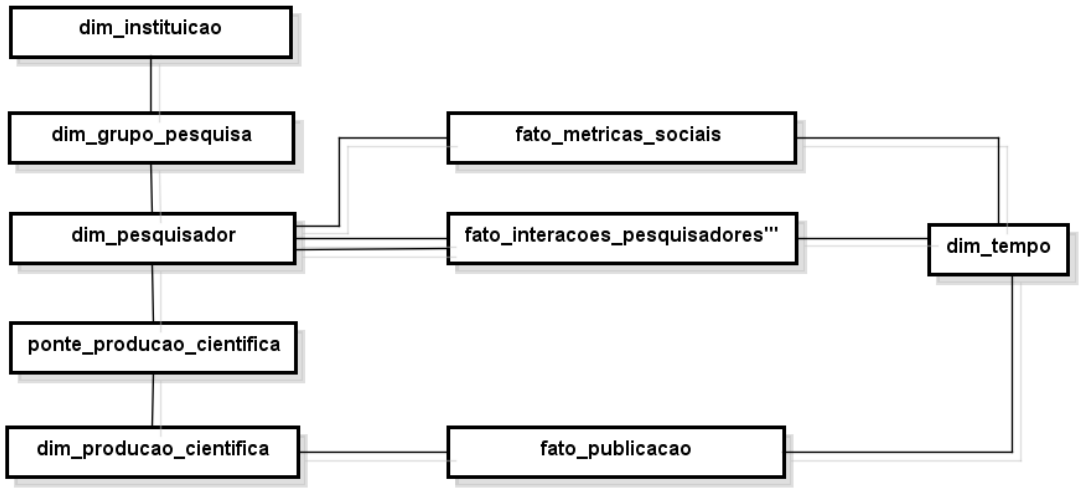

Figura 1. Modelo do Data Warehouse

Abaixo segue uma descrição simplificada sobre cada um dos elementos que compõe o DW utilizado nesta abordagem:

- **Dimensões**:

- **dim_instituicao**: armazena dados sobre as instituições (universidades ou centros de pesquisa) como nome e sua localidade.
- **dim_grupo_pesquisa**: dados sobre os grupos de pesquisa, como seu nome e a avaliação Capes atribuída.
- **dim_pesquisador**: dados específicos sobre um pesquisador, como nome, sexo, área de trabalho.
- **ponte_producao_cientifica**: pontes são dimensões especiais criadas para agrupar dados de outras dimensões, neste caso esta ponte permite definir quais pesquisadores participaram da criação de um trabalho de produção bibliográfica ou em um projeto de pesquisa.
- **dim_producao_cientifica**: a produção científica em si além do tipo atribuído a ela (se é uma produção bibliográfica ou um projeto).
- **dim_tempo**: O menor grão escolhido para esta dimensão foi o de ano, já que nem todos os artigos publicados ou projetos de pesquisa tem informações mais precisas do que o ano em que foram feito.

- **Fatos**:
  - **fato_metricas_sociais**: medidas com todas as métricas sociais calculadas. Estas métricas se relacionam a um pesquisador e a um instante de tempo em específico, logo tem relação com a dimensão dim_pesquisador e com a dim_tempo.
  - **fato_interacoes_pesquisadores**:Entendemos interação como uma relação dois a dois. Por exemplo, se quatro pesquisadores escreveram um artigo em conjunto teremos seis interações (a combinação de quatro tomado dois a dois). Este fato é uma relação entre as dimensões pesquisador e a dim_tempo.
  - **fato_produção**: este fato é uma medida direta do número de produções (artigos ou projetos de pesquisa) ao longo do tempo (dim_tempo). A partir dele também é possível obter os grupos que participaram do desenvolvimento de cada trabalho através da ponte_producao_cientifica.

Uma vez criado o DW, ferramentas capazes de trabalhar com dados neste formato (bases multidimensionais), chamadas ferramenta OLAP, foram utilizadas para construção das análises que serão apresentadas na próxima seção.

4. Análises

Uma das principais questões que devem ser levantadas quando se utiliza qualquer fonte de dados para análises de redes sociais é se a rede gerada a partir destas fontes obedecem determinadas propriedades. Algumas destas propriedades são: i)se a rede social obtida possui distribuição de grau segundo lei de potência e ii)se ela se comporta como um Small-World (mundo pequeno). A rede estudada neste trabalho possui ambas características. Para a rede de computação temos uma distância de 6.54, valor próximo a seis, como apontado pela literatura. Na (Figura 2) podemos ver o gráfico da distribuição de graus. Esta rede possui comportamente segundo lei de potências, onde muitos pesquisadores possuem grau-baixo (poucos relacionamentos), enquanto existem

pesquisadores com graus muito altos (maiores que 50). Estas análises foram construídas com todos os artigos/projetos no intervalo de 2000-2010.

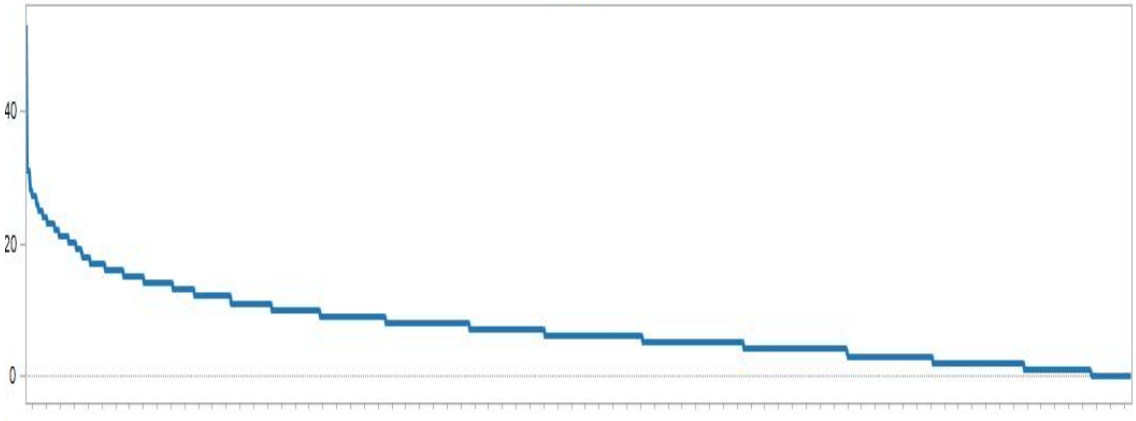

Figura 2. Grau x Número de pesquisadores para rede socia de computação

A distribuição do número de grupos por avaliação Capes pode ser vista na (Figura 3). A maior parte dos programas de pós-graduação é classificada como três e quatro, grupos de pesquisa de nível cinco são grupos intermediários. Grupos de nível seis e sete são considerados grupos de excelência. Apesar desta classificação mudar a cada três anos (de acordo com parâmetros estipulados pela CAPES) não foi possível obter os dados dos triênios anteriores. Desta maneira foram utilizados os dados do ultimo triênio (2009-2012).

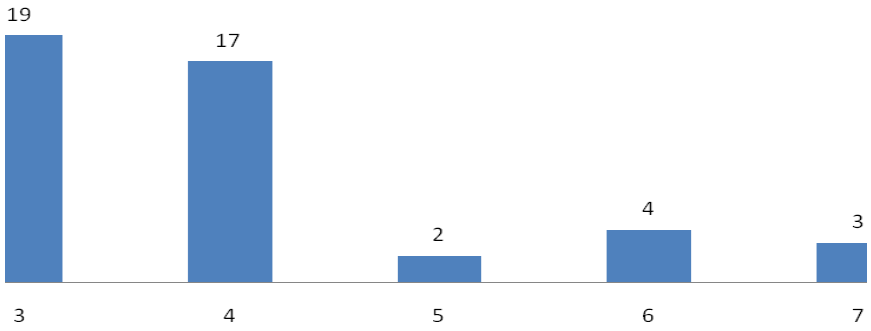

Figura 3. Avaliação do programa x quantidade de programas

A produtividade bibliográfica destes grupos ao longo dos anos pode ser visto no gráfico da (Figura 4), já a produtividade em relação ao número de projetos pode ser vista na (Figura 5). Apesar do número de programas seis e sete ser inferior ao número de programas três e quatro, a produção destes grupos é bem próxima, uma vez que os grupos de pesquisa para serem considerados grupos de excelência necessitam, entre outros critérios, produzir uma grande quantidade de trabalhos científicos segundo parâmetros estipulados pela Capes. Como o número de programas nível cinco é pequena

(apenas 2) este é o grupo com menor produtividade, tanto a nível de produção bibliográfica quanto por projetos de pesquisa.

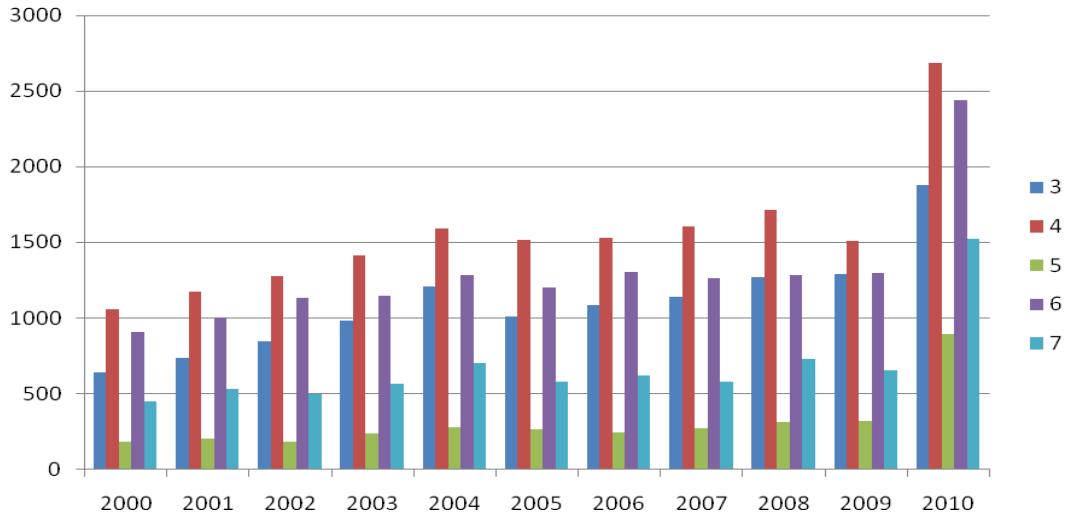

Figura 4. Tempo x número de produções bibliográficas por nível do grupo de pesquisa

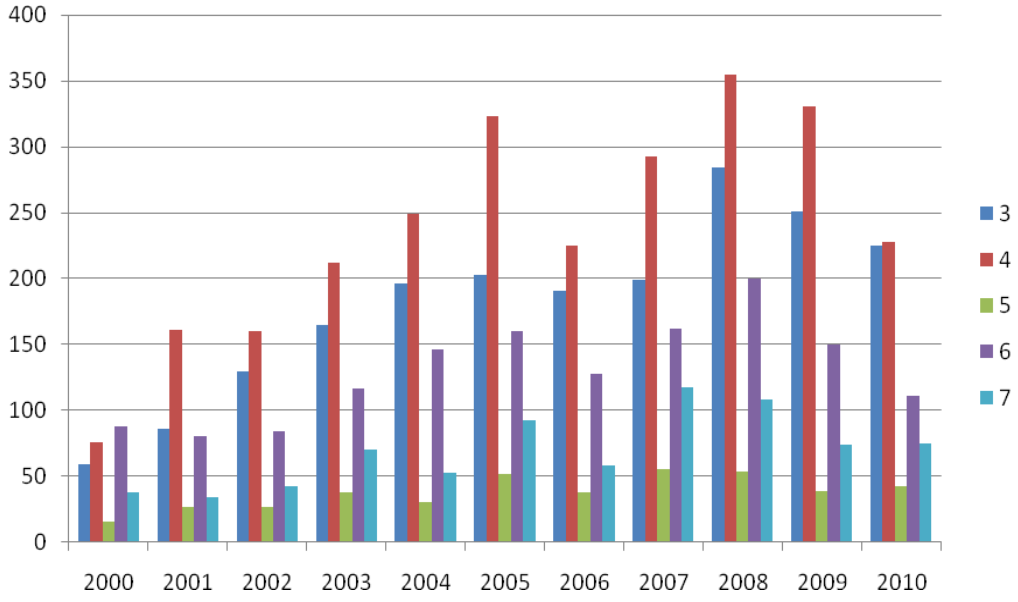

Figura 5. Tempo x número de projetos de pesquisa por nível do grupo de pesquisa

A evolução dos graus médios na rede está representada pela (Figura 6). Pode-se perceber um aumento considerável no último ano (2010). Como o número de publicações no mesmo período também aumentou (Figura 4), isso indica que mais

artigos em parceria foram produzidos. Não podemos dizer o mesmo dos projetos (Figura 5), uma vez que tivemos uma redução em seu número.

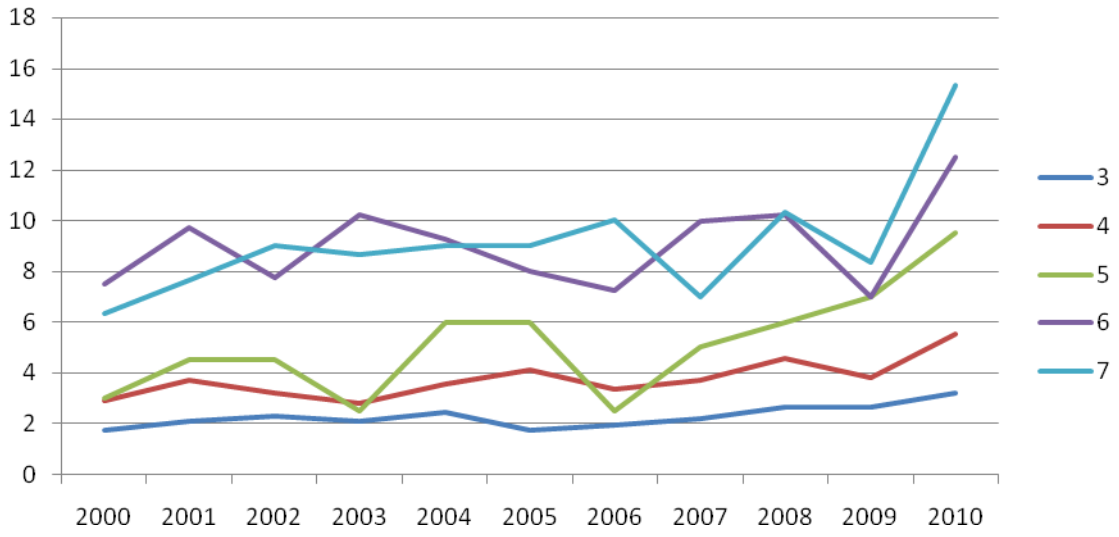

Figura 6. Tempo x Graus por nível dos programas

O Betweenness nos últimos anos vem decrescendo (Figura 7) . Isto indica que o número de menores caminhos passando por nós específicos vem diminuindo, ou seja, o fluxo de informações entre os pesquisadores está fluindo por caminhos mais curtos. Isto também implicará no aumento do coeficiente de clusterização (Figura 8) uma vez que mais relações surgiram na rede social. Outro aspecto interessante é que programas de nível três e quatro possuem valores de betweenness bem inferiores aos valores apresentados pelos programas de nível superior, indicando uma certa homogeneidade em relação a importância dos pesquisadores na rede de nível 3 e 4.

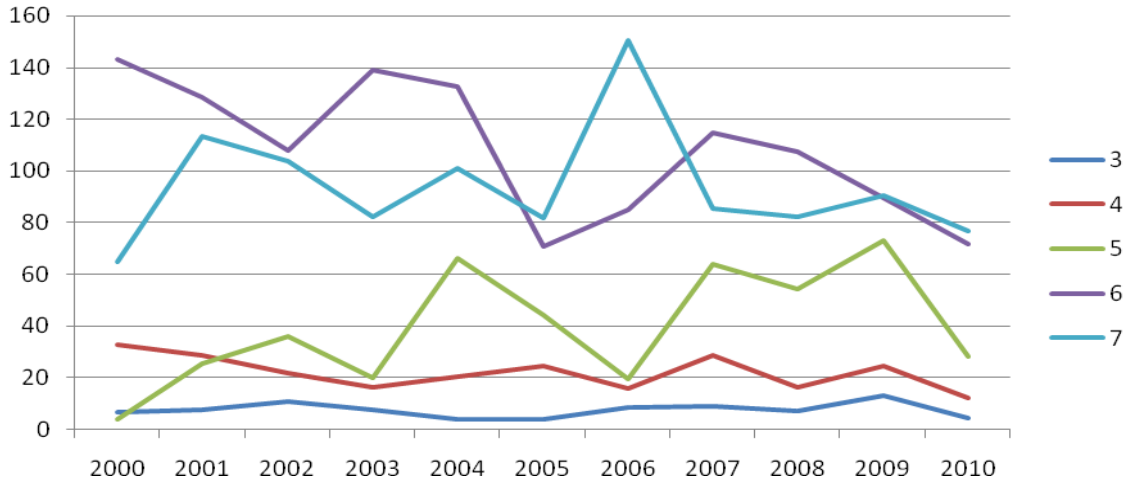

Figura 7. Tempo x Betweenness por nível dos programas

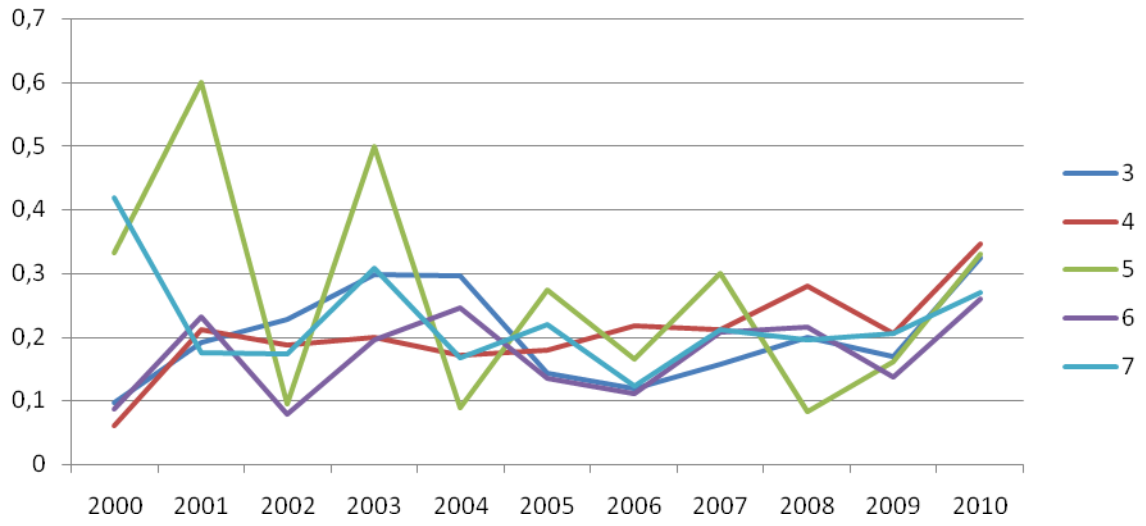

Figura 8. Tempo x Clustering por nível dos programas

O closeness indica o quão central um nó é em relação a rede. Em redes com poucas conexões alguns nós são os responsáveis por interconectar componentes conexas. Como foi percebido o aumento do grau, o closeness por conseqüência irá diminuir (Figura 9). Este fato ocorreu para todos os níveis dos grupos. A redução do closeness é algo desejável e indica uma uniformidade entre os vértices dentro da rede.

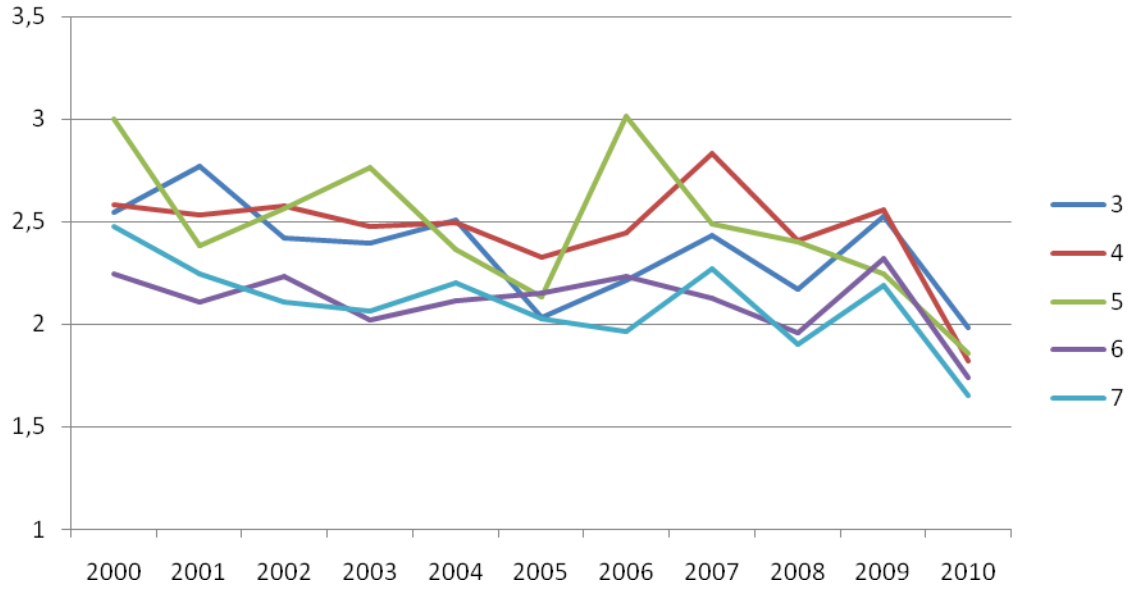

Figura 9. Closeness por nível dos programas ao longo dos anos

5. Conclusão

De acordo com as análises apresentadas, é possível traçar um padrão no comportamento da rede social de computação no Brasil, principalmente nos últimos anos, onde tivemos uma correlação direta entre o aumento da produção com as métricas avaliadas. Entretanto, é difícil definir a relação de causa/efeito, ou seja, não podemos afirmar que a maior interação entre os pesquisadores tenha gerado um maior número de trabalhos ou se o maior número de trabalhos implica em um maior número de interações.

Outro fator relevante é que a maioria dos grupos tiveram comportamento semelhante. O que indica que, apesar da diferença de produção entre os grupos, estes apresentam características semelhantes. Entretanto, programas considerados de excelência, apresentaram vértices diferenciados, fato corroborado pelo maior betweeness (Figura 7) em contraste com o Betweeness de grupos com classificação menor.

Como trabalho futuro propõe-se que a qualificação de cada trabalho seja levada em consideração, isto pode indicar um fator de diferenciação entre os grupos de excelência que para terem atingido este nível precisam obedecer a regras mais restritivas de produção acadêmica. Outro trabalho interessante seria a aplicação de técnicas de mineração para estabelecer as correlações mais relevantes. Estas relações não puderam ser totalmente exploradas devido as inúmeras combinações possíveis, que tornam a exploração manual difícil. Apesar do modelo de DW proposto permitir análises espaciais, apenas análises temporais foram exploradas, como trabalho futuro propõe-se a correlação espacial e temporal na produtividade dos grupos de pesquisa.

Referencias

BARABÁSI, A. .; JEONG, H.; NÉDA, Z. *et al.* Evolution of the social network of scientific collaborations. **Physica A: Statistical Mechanics and its Applications**, v. 311, n. 3-4, p. 590-614, ago 2002.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. Gephi : An Open Source Software for Exploring and Manipulating Networks. **American Journal of Sociology**, v. 2, n. 2, p. 361-362, 2009.

MENEZES, G. V.; ZIVIANI, N.; LAENDER, A. H. F. Um Estudo Comparativo de Redes Sociais em Ciência da Computação. **Workshop on Information Visualization and Analysis in Social Networks - WIVA**. 2008.

MONCLAR, R.S.; FARIA, F. F.; OLIVEIRA, J. *et al.* The Analysis and Balancing of Scientific Social Networks in Cancer Control. **Handbook of Research on Business Social Networking: Organizational, Managerial, and Technological Dimensions**. [S.l: s.n.], 2011. .

MONCLAR, RAFAEL STUDART; OLIVEIRA, JONICE; FIRMINO DE FARIA, F. *et al.* **Using social networks analysis for collaboration and team formation identification**. Proceedings of the 2011 15th International Conference on Computer

Supported Cooperative Work in Design (CSCWD). **Anais**... [S.l.]: IEEE. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5960128>. Acesso em: 24 abr. 2012. , jun 2011

NEWMAN, M. E. The structure of scientific collaboration networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 98, n. 2, p. 404-9, 16 jan 2001.

SOLLA PRICE, D. J. DE. Networks of Scientific Papers. **Science**, v. 149, n. 3683, p. 510-515, 30 jul 1965.

STRÖELE, V.; SILVA, R.; SOUZA, M. F. DE; *et al.* Identifying Workgroups in Brazilian Scientific Social Networks. **Journal of Universal Computer Science**, v. 17, p. 1951-1970, 2011.

Grupos de Pesquisa em Pós-graduação Capes. Disponível em: < http://conteudoweb.capes.gov.br/conteudoweb/CadernoAvaliacaoServlet?acao=filtraArquivo&ano=2009&codigo_ies=&area=2 >.Acesso em: 1 abr. 2012.

Plataforma de Currículos Lattes. Disponível em: < http://lattes.cnpq.br/>. Acesso em 2 abr. 2012.

WebQualis. Disponível em: < http://qualis.capes.gov.br/webqualis/>. Acesso em 2 abr. 2012.

INMON, W. H.; The Data Warehouse and Data Mining. **Journal of Communications of the ACM**. v. 39, 49-50, 1996.

NEWMAN, M. E. Coauthorship networks and patterns of scientific collaboration. **Proceedings of the National Academy of Sciences of the United States of America**, v. 101, p. 5200-5205, 2004.

MOURA, VINICIUS F., Uma Métrica Para Ranqueamento Em Redes De Colaboração Baseada Em Intensidade De Relacionamento. 2010. 63 folhas.Tese de Doutorado-Programa de Engenharia de Sistemas e Computação