Caracterizando as redes de coautoria de currículos Lattes

Jesús P. Mena-Chalco[1], Luciano A. Digiampietri[2], Roberto M. Cesar-Jr[2]

[1]Universidade Federal do ABC (UFABC)

[2] Universidade de São Paulo (USP)

jesus.mena@ufabc.edu.br, digiampietri@usp.br, cesar@ime.usp.br

Abstract. *The topological characterization of complex networks allows to analyze behaviors of natural and social systems. These analyzes facilitate the understanding of structure and dynamics of networks. In recent years, the (complex) networks of co-authorship have attracted considerable interest due to the fact that such networks represent social behavior among researchers. This paper describes a proposal for (i) the automatic identification of co-authorships using bibliographical productions, and (ii) the topological characterization of co-authorship networks obtained from research groups registered in the Lattes Platform. This proposal was evaluated considering 176,114 Lattes curricula of researchers associated with the Area of Exact and Earth Sciences.*

Resumo. *A caracterização topológica de redes complexas permite analisar comportamentos de sistemas naturais e sociais. Essas análises facilitam a compreensão da estrutura e dinâmica das redes. Nos últimos anos, as redes (complexas) de coautoria têm atraído considerável interesse devido ao fato que estas representam o comportamento social entre pesquisadores. Este trabalho descreve uma proposta para (i) a identificação automática de coautorias em produções bibliográficas, e (ii) a caracterização topológica de redes de coautoria de grupos de pesquisadores cadastrados na Plataforma Lattes. A proposta foi avaliada considerando 176 114 currículos Lattes de pesquisadores associados à Grande Área de Ciências Exatas e da Terra.*

1. Introdução

As colaborações científicas, na forma de coautoria de produções bibliográficas de grupos de pesquisadores, vem recebendo especial interesse de entidades avaliadoras e de fomento em ciência e tecnologia [Balancieri et al. 2005] pois, ao contrário de se concentrar apenas na lista e qualidade de produções bibliográficas, as coautorias acadêmicas fornecem uma visão sobre a estrutura e dinâmica inerentes das colaborações entre os pesquisadores. Comumente, as coautorias acadêmicas entre pesquisadores são representadas por meio de redes (grafos) de colaboração, em que os atores (pesquisadores) são representados por nós, e as participações em coautoria conjunta entre estes são representados por arestas.

A estrutura e dinâmica de colaboração em redes de coautoria acadêmica de grupos de pequeno e médio porte têm sido amplamente estudadas nas áreas de Ciência da Informação, Bibliometria/Cientometria [Liu et al. 2005]. Neste trabalho descrevemos uma proposta para estudar a estrutura e dinâmica de colaborações de grupos de grande porte, de pessoas cadastradas na Plataforma Lattes. A caracterização de extensos grupos acadêmicos é realizada através de métricas globais que exploram a topologia de redes complexas.

A Plataforma Lattes oferece um interessante estudo de caso por vários motivos: (i) os currículos Lattes tornaram-se um padrão nacional utilizado na avaliação individual das atividades científicas, acadêmicas e profissionais; (ii) a grande maioria dos pesquisadores brasileiros, de todas as áreas do conhecimento, está cadastrada na Plataforma Lattes, sendo que atualmente o número de currículos Lattes ultrapassa a marca de um milhão; e (iii) nos últimos anos, a ciência brasileira vem apresentando um rápido crescimento de produção acadêmica, impulsionada pelas políticas de ciência e tecnologia. Os motivos acima mencionados tornam os dados da Plataforma Lattes uma fonte extremamente rica (mas pouco explorada) para investigar e compreender o comportamento de diversos grupos de pesquisa de grande porte, sejam estes relacionados a áreas do conhecimento, grandes áreas do conhecimento, ou até o banco de dados inteiro de currículos Lattes. Nesse contexto, um dos trabalhos relacionados com a identificação de redes de coautoria para Grandes Áreas do conhecimento foi proposto em [Mena-Chalco and Cesar-Jr. 2011]. Nesse trabalho foram utilizadas apenas duas métricas individuais (próprias para cada pesquisador) referentes ao número de colaboradores e à medida que estima a colaboração no grupo (*Author Rank*).

Contribuição: Neste trabalho propomos a identificação rápida de coautorias, e a utilização de métricas globais sobre redes de coautoria extraídas de publicações em diferentes períodos de tempo (triênios). Esta proposta representa um avanço na direção da *caracterização massiva de extensos grupos acadêmicos*, cadastrados na Plataforma Lattes, através de métricas. A relevância deste trabalho recai sobre as vantagens das análises que são realizadas considerando as métricas topológicas de grafos correspondentes a Grandes Áreas do Conhecimento do Brasil. Essas análises são muito importantes para compreender/interpretar a Ciência Brasileira no âmbito das colaborações científicas.

O restante do artigo está organizado da seguinte maneira. Na Seção 2 são descritas as estratégias usadas para a obtenção dos currículos Lattes. Nas Seções 3 e 4 são apresentados o método usado para a identificação, e as métricas consideradas na caracterização de redes de coautoria, respectivamente. Finalmente, na Seção 5 são apresentados resultados relacionados com diferentes redes de coautoria pertencentes 176 114 pesquisadores associados à Grande Área de Ciências Exatas e da Terra.

2. Banco de dados de currículos

Os currículos Lattes foram obtidos pela internet e baixados localmente utilizando-se duas estratégias para a identificação automática da maior quantidade possível dos códigos de identificação (IDs) Lattes. Na primeira abordagem, foram feitas 80 consultas na interface de busca provida pela Plataforma Lattes, cada uma utilizando como palavras-chave o nome de uma Área do Conhecimento. Foi desenvolvido um *parser* para encontrar os IDs Lattes em cada uma das respostas. Para a segunda abordagem foi utilizado um método recursivo, similar ao de busca em grafos, de identificação de colaboradores com ID cadastrado na Plataforma Lattes (ex. coautores, professores orientadores, alunos orientados). Na primeira iteração, dado um ID Lattes, foram extraídos todos os IDs dos colaboradores identificados no currículo em formato HTML. Nas iterações seguintes foram utilizados os novos IDs na busca por novos IDs Lattes[1].

[1]Um caso curioso é que com apenas um único ID Lattes pertencente a um pesquisador com Bolsa de Produtividade 1A, foram identificados automaticamente mais de 340 000 IDs Lattes

As duas estratégias de identificação de IDs Lattes e armazenamento de currículos Lattes foram executadas em maio de 2011. Ao todo foram baixados 1 236 548 currículos totalizando pouco mais de 16 GB em arquivos HTML. É importante frisar que as estratégias adotadas não visavam a obtenção de toda a base de currículos Lattes, mas sim um conjunto significativo de currículos para serem processados e servirem de base para a criação e análise de redes sociais de pesquisadores. Levando-se em conta que o CNPq anunciou que em agosto de 2007 [CNPq 2007] a base de currículos atingiu um milhão de currículos, considera-se que foi obtida uma quantidade representativa de currículos do banco de dados de currículos do CNPq.

Cada currículo Lattes obtido foi processado a fim de identificar e popular um banco de dados com todos os elementos de produção bibliográfica, técnica, artística, projetos, orientações, bancas, dentre outros. Vale ressaltar que um desafio computacional no processamento dos currículos em formato HTML é a identificação das partes constituintes de cada produção acadêmica. Assim, foi desenvolvido um *parser* para que, através de expressões regulares, sejam identificadas todas as partes constituintes de cada produção.

3. Identificação de redes de coautoria

Uma rede de coautoria acadêmica mostra atividades acadêmicas, na forma de produção bibliográfica (e.g. artigos publicados em congressos), que são realizadas de forma conjunta por um determinado grupo de pesquisadores [Maia and Caregnato 2008]. Na rede de coautoria, geralmente, cada pesquisador é representado por um nó. A ligação (aresta) entre dois nós representa pelo menos uma produção feita em coautoria.

Do banco de dados criado temos apenas a lista de produções associadas a cada currículo Lattes de cada pesquisador. Assim, foi necessário o desenvolvimento de um procedimento de identificação automática de produções feitas em coautoria. A deteção de produções bibliográficas iguais entre pesquisadores é realizada através de comparações dois a dois entre todas as produções de conjuntos de dados discretizados por ano e tipo de produção (e.g. artigo publicado em periódico ou capítulo de livro). As produções bibliográficas com anos de publicação diferentes não foram usadas em nenhuma comparação.

Devido a inconsistências (erros de digitação ou falta de padronização na escrita dos nomes dos coautores [Kang et al. 2009]) no preenchimento das informações nos currículos Lattes, a comparação de duas produções quaisquer é realizada através de um casamento aproximado entre os títulos associados a cada produção. Duas publicações são consideradas iguais se a porcentagem de similaridade entre os títulos for maior do que uma determinado limiar. A similaridade entre duas cadeias baseia-se na distância proposta por Levenshtein [Navarro 2001]. A distância Levenshtein é obtida através do número mínimo de inserções, eliminações ou substituições de caracteres necessárias para transformar um texto em outro. Consideramos dois títulos iguais/equivalentes se ambos são pelo menos 90% similares.

O método de identificação de coautoria baseia-se em uma melhora do algoritmo utilizado no módulo de tratamento de redundâncias do scriptLattes [Mena-Chalco and Cesar-Jr. 2009]. Com a versão do algoritmo apresentado no scriptLattes, o número de comparações requeridas para identificar coautorias em uma lista de n produções bibliográficas está no ordem de $\Theta(\frac{1}{2}(n^2 + n)) = O(n^2)$. Na prática, se n=10 000, e cada

comparação for realizada em 0.001 segundos, então serão requeridas 13.89 horas para processar todas as comparações do grupo. Certamente esta implementação limitaria o processamento de listas de produções de algumas dezenas de milhares de elementos.

Nesse sentido, criamos uma leve modificação deste algoritmo de tal forma que o custo computacional seja o menor possível. Através de uma análise empírica, partimos do suposto de que os erros na digitação ou preenchimento dos dados das publicações acontecem no meio ou no fim do título da publicação, sendo que raramente isso acontece na primeira letra do título. Assim, antes da realização das comparações, é realizada uma classificação das publicações considerando a primeira letra contida no seu título, criando tantas listas quanto letras contidas na primeira letra dos títulos das produções bibliográficas, e.g. publicações que começam com a letra 'A' estarão na lista `A`. Com essa modificação, e assumindo a existência de no mínimo 5 letras iniciais diferentes nos títulos das produções, o número de comparações requeridas para identificar coautorias em uma lista de n produções bibliográficas será $\Theta(n + \frac{1}{2}(\frac{n}{5}^2 + \frac{n}{5})) = O(n^2)$. Na prática, se n=10 000, e cada comparação for realizada em 0.001 segundos, então serão requeridas 00.55 horas. Observe que a complexidade computacional é a mesma para ambas as implementações ($O(n^2)$), entretanto a constante embutida é menor na última modificação implementada, permitindo assim uma diminuição substancial de tempo de processamento.

Neste trabalho as produções bibliográficas que consideramos na identificação das redes de coautoria são: (1) artigos publicados em periódicos, (2) livros publicados/organizados ou edições, (3) capítulos de livros, (4) textos em jornais de notícias/revistas, (5) trabalhos completos publicados em anais de congressos, (6) resumos expandidos publicados em anais de congressos, e (7) resumos publicados em anais de congressos. Todas as coautorias são representadas através de listas de adjacência. Essa forma de representação, frente às matrizes de adjacência, é a mais adequada para o tratamento de extensas listas de produções correspondentes a grandes grupos.

4. Caracterização de redes de coautoria

A análise de redes sociais está baseada na premissa de que relações entre os atores sociais podem ser descritas mediante um *grafo* (direcionado ou não-direcionado) [Liu et al. 2005]. Em redes de coautoria acadêmica os pesquisadores são considerados os atores/entidades e as colaborações (participação em forma conjunta na elaboração de uma produção bibliográfica) são consideradas relações/ligações entre pesquisadores.

A vantagem da representação das redes de coautoria por meio de grafos é que esta última permite a utilização da Teoria dos Grafos para analisar comportamentos de relacionamento social, padrões e implicações destes relacionamentos, que de outra forma podem ser pobremente interpretadas [Wasserman and Faust 1994]. Neste trabalho as relações entre os pesquisadores somente são referidas a coautorias provenientes de produções acadêmicas. Assim, conjuntamente com a utilização da Teoria dos Grafos, podem ser investigadas diferentes métricas estruturais que possibilitem caracterizar as coautoria.

As métricas podem ser divididas em globais (características estimadas sobre todo o grafo) e individuais (características estimadas sobre os atores individuais). Neste trabalho, ainda em andamento, consideramos apenas métricas globais sobre grafos obtidas em diferentes períodos de tempo. Nas nossas análises, ao todo foram 12 as métricas consideradas sobre as redes de coautoria:

- **Arestas**. Para o caso das redes de coautoria uma medida simples que reflita o número de ligações entre os pesquisadores é o número de arestas em um determinado grafo. Esta métrica é comumente utilizada para avaliar a evolução temporal do número de coautorias em redes obtidas em diferentes períodos de tempo.
- **Nós participantes em coautoria**. Refere-se ao número de pesquisadores que, em um determinado período, participaram pelo menos em uma coautoria, i.e. o número de nós cujo grau seja maior ou igual a 1.
- **Diâmetro**: Refere-se ao tamanho da maior distância geodésica[2] entre qualquer par de pesquisadores. O diâmetro de um grafo pode variar de um mínimo de 1 (se o grafo for completo) a um máximo de V-1, onde V é o número de nós no grafo. Para o caso de grafos não-conexos o diâmetro um grafo refere-se ao maior diâmetro entre os diâmetros das componentes conexas.
- **Grau médio**: O grau de um nó refere-se ao número de arestas incidentes nele. O grau de um nó pode variar de um mínimo de 0 (se não existem arestas adjacentes no nó), a um máximo de V-1, onde V é o número de nós no grafo (se o nó é adjacente a todos os outros). Um nó com grau zero é denominado *isolado*. O grau médio refere-se à média dos graus de todos os nós existentes no grafo.
- **Densidade**: Refere-se à razão/proporção entre o número de arestas e o número de arestas possíveis. A densidade de um grafo pode variar de um mínimo de 0 (se não existem arestas no grafo) a um máximo de 1 (se cada nó for adjacente a todos os outros).
- ***Coreness***: O k-*core* de um grafo é um subgrafo obtido da eliminação (do grafo original) de todos os nós com grau menor ou igual a k. O *coreness* de um dado nó, é o máximo k tal que o nó ainda esteja presente no k-*core*, mas eliminado do $(k-1)$-*core* [Borgatti and Everett 2000]. O *coreness* de um grafo é o valor médio dos *coreness* de todos os nós existentes no grafo.
- ***Rich club***: O clube rico de um grafo refere-se a um subgrafo que contém os nós com maior grau. Geralmente o tamanho (número de nós) do clube rico é limitado por uma porcentagem (e.g. 20%) dos nós com maior grau. O coeficiente de clube rico quantifica quão perto um subgrafo, que contém os nós com maior grau, está de formar um *clique* [Colizza et al. 2006]. Este coeficiente é a proporção do número de arestas no subgrafo, dividido pelo número máximo de arestas possíveis no subgrafo (i.e. $k(k-1)/2$, onde k o número de nós pertencentes ao clube rico). O coeficiente de clube rico pode variar de um mínimo de 0 (se não existem arestas) a um máximo de 1 (se o subgrafo for completo).
- **Tamanho da maior componente conexa**: Em um grafo desconexo, uma componente conexa é um subgrafo fechado no qual existe um caminho entre quaisquer dois nós pertencentes ao subgrafo, e não existe caminho entre um nó pertencente ao subgrafo com outro nó não-pertencente ao subgrafo. Nesse contexto, o tamanho da maior componente conexa refere-se à componente conexa com maior número de nós.
- **Porcentagem da maior componente conexa**: Uma rede de coautoria não é representada por uma componente conexa, pois comumente existem determinados grupos que trabalham de forma isolada, entretanto caracterizar um grupo através de sua maior componente conexa (e sua porcentagem) permite investigar sobre o comportamento do maior grupo colaborativo.

[2] A distância geodésica entre dois nós é o caminho mínimo existente entre eles.

- **Transitividade**: Refere-se à probabilidade de que dois nós adjacentes a um nó estejam ligados. Comumente, este valor também é denominado de coeficiente de *clustering*, e para o nó i este valor representa a proporção de arestas entre nós dentro da vizinhança do nó i, dividido pelo número máximo de arestas que poderiam existir entre todos os vizinhos. Isto é, o coeficiente de *clustering* é a razão entre o número de triângulos que contêm o nó i e o número de triângulos que poderia existir se todos os vizinhos do nó i forem interligados. O coeficiente de *clustering* do grafo é o valor médio dos coeficientes de *clustering* de todos os nós [Wasserman and Faust 1994].
- **Assortatividade**: Para o nó i refere-se à preferência do nó i ter arestas a outros nós que mantenham grau do nó similares (ou diferentes) ao grau do nó i. O coeficiente de assortatividade é o coeficiente de correlação de Pearson dos graus entre os pares de nós ligados [Newman 2002]. Valores positivos do coeficiente de Pearson indicam uma correlação entre nós de grau similar, entretanto valores negativos indicam uma correlação entre nós de grau diferente. Este coeficiente pode variar de um mínimo de -1 a um máximo de 1 (rede com assortatividade máxima).
- **Caminho médio**: Um caminho entre os nós i e j é referido a um passeio sem nós repetidos para chegar em j, partindo de i. O caminho mínimo entre os nós i e j corresponde ao caminho de comprimento mínimo entre i e j. O caminho médio de um grafo refere-se ao valor médio dos caminhos mínimos entre todos os possíveis pares de nós existentes no grafo.

5. Resultados experimentais

A proposta de caracterização de redes de coautoria acadêmica entre diferentes pesquisadores cadastrados na Plataforma Lattes foi testada considerando extensos grupos de pesquisa. Em particular, nesta seção mostramos os resultados correspondentes a grafos de coautoria obtidos do processamento de currículos Lattes de pesquisadores associados à Grande Área de "Ciências Exatas e da Terra". Ao todo, foram identificados 176 114 pesquisadores associados com esta Grande Área do conhecimento. Segundo o CNPq, esta grande área agrupa as áreas relacionadas com: Astronomia, Ciência da computação, Probabilidade e estatística, Física, Geociências Matemática, e Química.

Definida a Grande Área, dois diferentes experimentos foram realizados. Para o primeiro experimento foram extraídas do banco de dados (Seção 2) todas as produções bibliográficas publicadas durante o período de 1990-2010. Para o segundo experimento

Tabela 1. Períodos considerados na caracterização de redes de coautoria do 176 114 pesquisadores associados à Grande Área de Ciências Exatas e da Terra.

	Períodos	Publicações	Coautorias
t1	1990-1992	91 270	11 157
t2	1993-1995	160 152	19 892
t3	1996-1998	297 255	36 330
t4	1999-2001	495 958	61 996
t5	2002-2004	776 959	98 688
t6	2005-2007	992 295	128 808
t7	2008-2010	910 647	124 506
t8	**1990-2010**	**3 724 536**	**303 886**

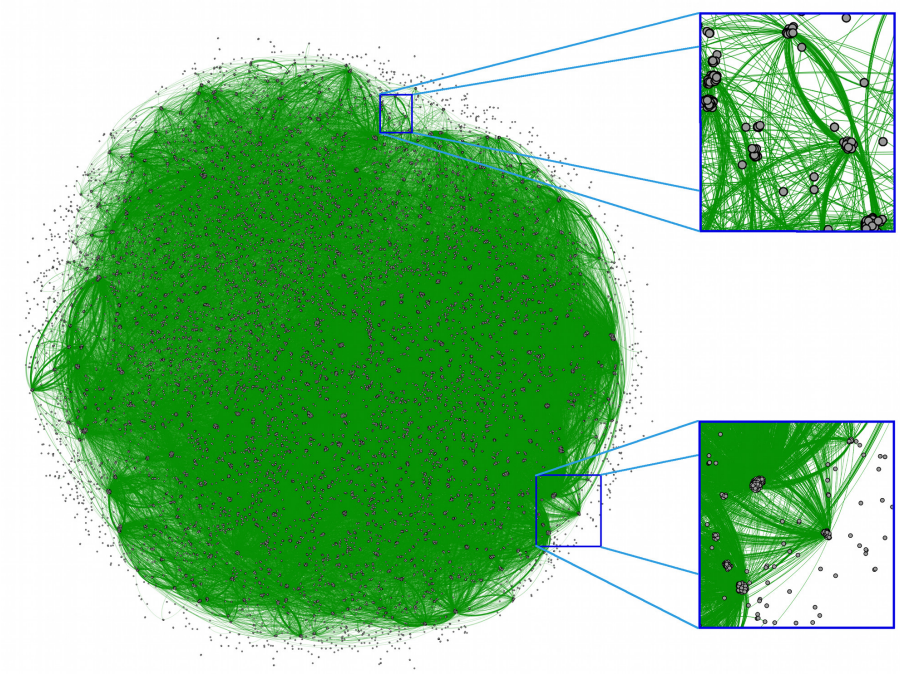

Figura 1. Rede de coautoria pertencente a 176 114 pesquisadores associados à Grande Área de Ciências Exatas e da Terra. As coautorias entre pesquisadores estão representadas por arestas (na cor verde).

foram criadas 7 listas de produções bibliográficas, cada uma relacionada com um período trienal. Na Tabela 1 é apresentado um quadro resumo dos períodos, publicações e coautorias identificadas por nossa proposta (Seção 3).

5.1. Rede de coautoria no período 1990-2010

Para este experimento foram extraídas todas as produções bibliográficas dos 176 114 pesquisadores associados à Grande Área de "Ciências Exatas e da Terra" e cadastradas na Plataforma Lattes. Ao todo foram extraídas 3 724 536 publicações. Em seguida foi identificada a rede de coautoria considerando somente a lista de publicações.

Na Figura 1 apresenta-se a rede de coautoria obtida para o grupo em análise. Foram identificadas 303 886 coautorias diferentes pertencentes a 65 221 (aprox. 37%) pesquisadores. Os nós isolados (i.e., pesquisadores sem coautoria com outros da mesma Grande Área) não foram diagramados. No grafo, os nós que visualmente aparecem isolados correspondem a um subconjunto de pesquisadores (geralmente formado por dois ou três pesquisadores) desconexos da maior componente conexa.

Em média cada pesquisador mantém colaboração com 3,45 pesquisadores, e a transitividade (coeficiente de *clustering*) é de 0,16. Observa-se que no grafo existem diversos subgrupos (agrupamentos de nós bem próximos um do outro) de pesquisadores que formam *clusters*. Visualmente observa-se que os pesquisadores desses subgrupos tendem a colaborar mais entre si. A maior componente conexa é composta de 63 066 (aprox. 36%) pesquisadores, e o diâmetro do grafo é 1 037. Esta última medida quantifica quão longe estão os dois nós mais distantes no grafo. Finalmente, uma característica importante do grafo é que o caminho médio é de 5,69. Este último valor é bem próximo ao fenômeno descrito por [S. Milgram 1967].

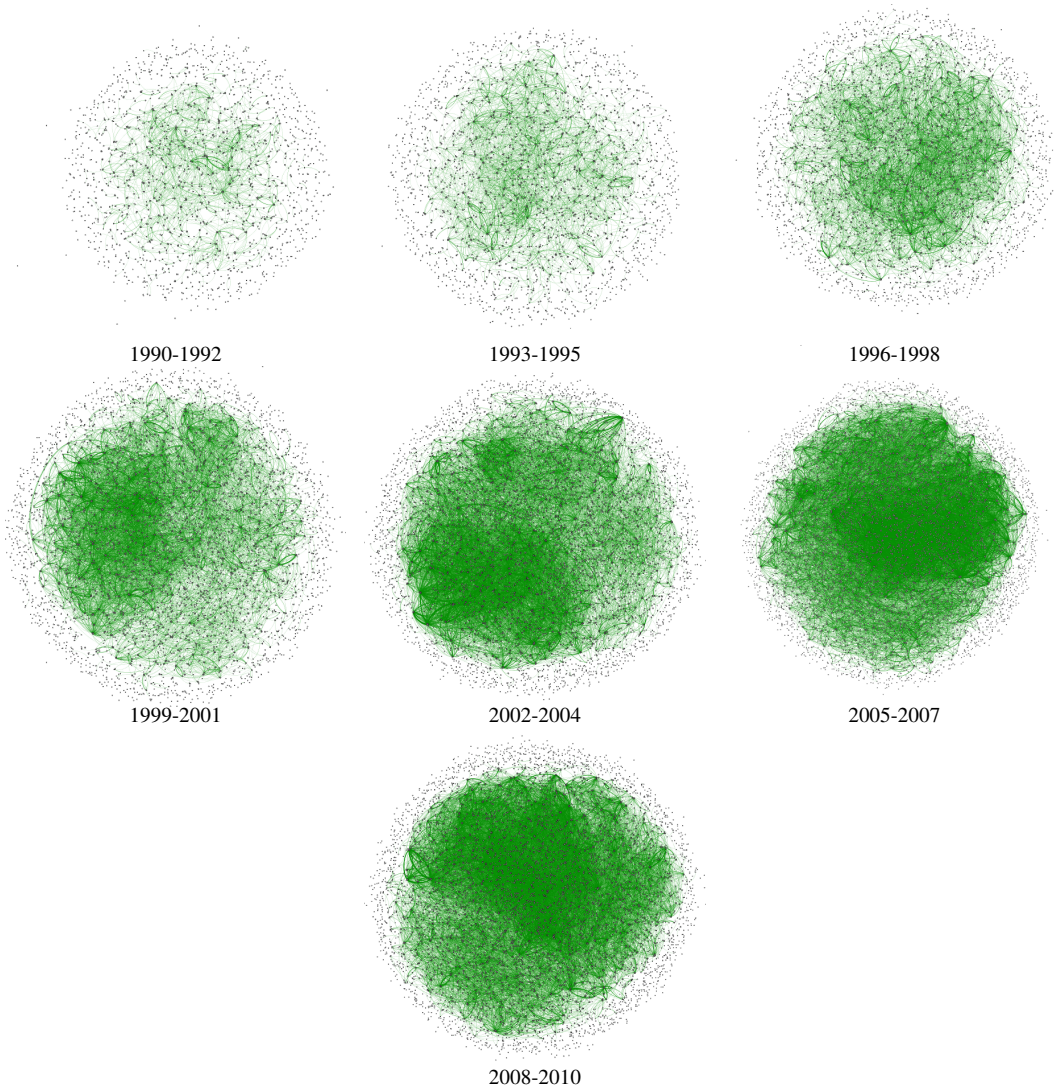

Figura 2. Redes de coautoria, discretizadas por triênios, pertencentes a 176 114 pesquisadores associados à Grande Área de Ciências Exatas e da Terra.

5.2. Redes de coautoria discretizadas por triênios

Para este experimento foram extraídas todas as produções bibliográficas dos 176 114 pesquisadores associados à Grande Área de "Ciências Exatas e da Terra" e discretizados por triênios. Para cada período foram obtidas as listas de produções bibliográficas e em seguida foram identificadas as redes de coautoria correspondentes. Na Figura 2 apresentam-se as redes de coautoria obtidas para os 7 períodos. Uma aresta (na cor verde) indica pelo menos uma publicação realizada em colaboração entre dois pesquisadores. Os nós isolados não foram diagramados. Nos grafos, os nós que visualmente aparecem isolados correspondem a colaborações entre poucos pesquisadores (geralmente colaborações entre dois ou três pesquisadores). Para cada período foram identificados diferentes subgrupos de pesquisadores que formam *clusters*. Observa-se que ao longo dos triênios a 'complexidade' do grafo se incrementa. Essa simples representação visual permite inspecionar o incremento na colaboração acadêmica ao longo do tempo.

Na Figura 3 apresentam-se as 12 métricas obtidas para cada período analisado. Adicionalmente são apresentadas as métricas correspondentes a rede de coautoria obtida

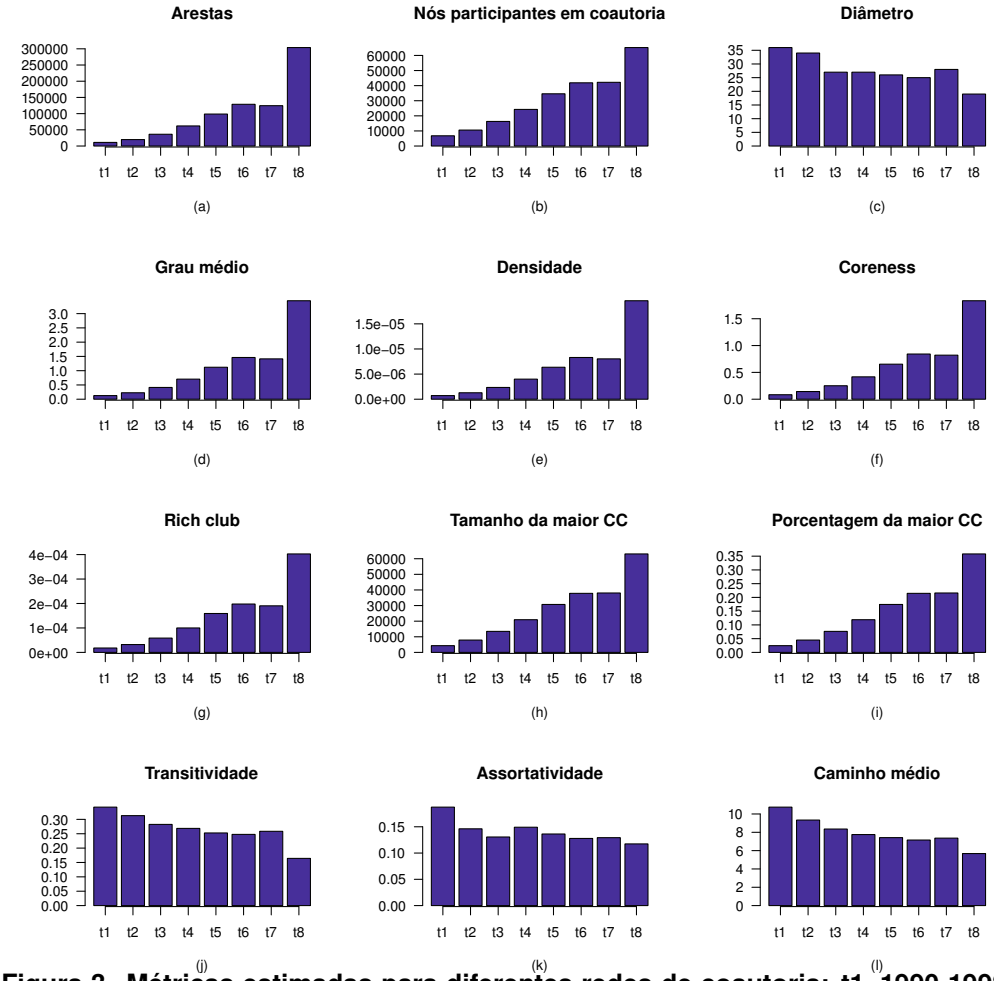

Figura 3. Métricas estimadas para diferentes redes de coautoria: t1=1990-1992, t2=1993-1995, t3=1996-1998, t4=1999-2001, t5=2002-2004, t6=2005-2007, t7=2008-2010, t8=1990-2010.

para o período 1990-2010. As métricas correspondentes a número de arestas, nós participantes em coautoria, grau médio, densidade, *coreness*, *rich club*, tamanho e porcentagem da maior componente conexa apresentam um comportamento ascendente ao longo dos períodos (Figuras 3a-i). Claramente as métricas obtidas para os períodos t1 até t6 mostram um crescimento exponencial, entretanto para o período t7 mostram uma aparente desaceleração de crescimento. Um dos motivos aparentes poderia ser a possível falta de atualização das produções bibliográficas (publicadas desde 2008) nos currículos obtidos automaticamente da Plataforma Lattes. Este comportamento ainda precisa ser investigado mas acreditamos que também está relacionado com a aparente desaceleração do crescimento no ensino superior [Brito-Cruz 2012].

Por outro lado, as métricas correspondentes ao diâmetro, transitividade, assortatividade e caminho médio (Figuras 3j-l) apresentam um comportamento descendente ao longo dos períodos. A métrica de transitividade em um grafo é o valor médio dos coeficientes de *clustering* de todos os nós. Vemos que, ao longo dos períodos novas coautorias foram estabelecidas, formando agrupamentos ou até incrementando agrupamentos já existentes em períodos anteriores, tornando assim menor a transitividade no grafo. Isto é, ao longo dos triênios, o coeficiente de *clustering* do grafo como um todo, diminue pois a tendência de todos os nós se agruparem também diminue.

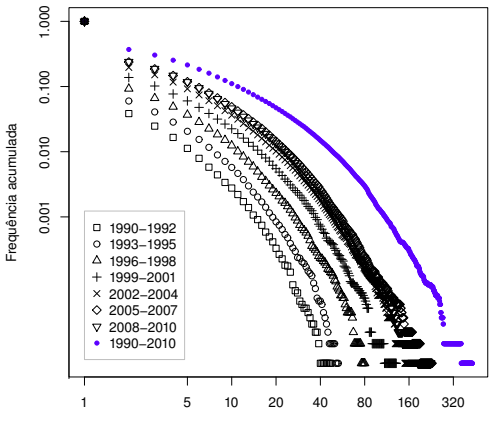

Figura 4. Distribuições cumulativas de graus das redes de coautoria.

Já a assortatividade intuitivamente indica quão homogêneos são os graus dos nós. Os valores positivos indicam uma correlação positiva entre nós de grau similar. Ao longo dos períodos esta métrica tem um comportamento descendente pois a homogeneidade nos graus dos nós diminui. Finalmente, como esperado, o caminho médio diminui ao longo dos períodos pois, em cada novo período, existe uma maior conetividade entre os nós.

Na Figura 4 apresentam-se as distribuições cumulativas de grau para todas as redes de coautoria tratadas neste trabalho. Observe que essas distribuições cumulativas explicitam o comportamento crescente identificado na métrica de grau dos nós. Existe uma pequena fração de pesquisadores que são coautores com uma grande quantidade de colaboradores. Por outro lado, existe uma grande fração de pesquisadores que são coautores com uma pequena quantidade de colaboradores. Observe também que ao longo dos triênios o grau dos nós (o número de coautorias) vêm se incrementando.

A fim de avaliar, ao longo dos períodos, a evolução da coautoria nos pesquisadores da Grande Área de Ciências Exatas e da Terra, foi elaborado um *ranking* dos 100 pesquisadores com maior grau em cada período (Figura 5). Neste diagrama, os pesquisadores estão representados por caixas e posicionados na forma vertical (coluna). Para cada período, o pesquisador de maior grau (maior número de coautores) está posicionado no topo da coluna. O pesquisador com menor grau (menor número de coautores) está posicionado na parte inferior da coluna. As linhas entre as colunas ti (período i) e t$i+1$ (período $i+1$) representam as mudanças de posição no *ranking* de um pesquisador. Essa diagrama permite observar a dinâmica no *ranking* da métrica relacionada ao grau de cada pesquisador. Muitas vezes estamos interessados em identificar os pesquisadores que mantenham de certa forma a regularidade no posicionamento do *ranking* do período i ao período $i+1$. Neste trabalho consideramos um comportamento regular se a mudança de posição no *ranking* entre os períodos for menor ou igual a 5 (linhas na cor vermelha na Figura 5). Observe que, na média, os pesquisadores com maior número de colaboradores tendem a manter essa regularidade no posicionamento, período após período.

Um caso particular de regularidade no posicionamento do *ranking* pertence ao pesquisador com maior número de colaboradores no período t7 (a caixa associada ao pesquisador está representada na cor azul e a mudança de posicionamento, com linhas grossas na cor vermelha). Os posicionamentos no *ranking* para os sete períodos foram de: t1=9, t2=4, t3=2, t4=1, t5=1, t6=1, e t7=1. Atualmente este pesquisador mantém uma bolsa de produtividade 1A do CNPq. Finalmente, é importante destacar que esta forma de visualização permite por em evidência a evolução no *ranking* considerando diferentes métricas individuais (métricas associadas a nós individuais).

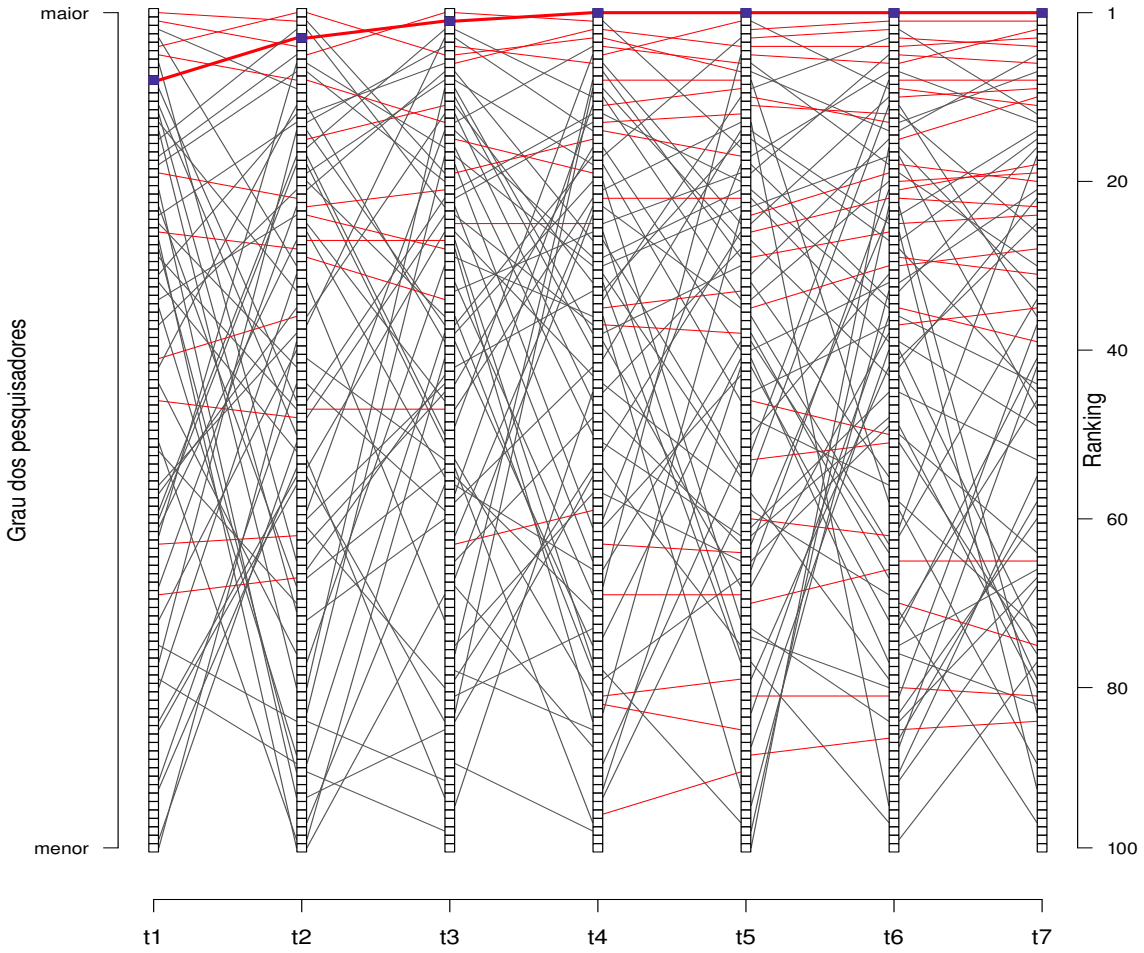

Figura 5. *Ranking* dos 100 pesquisadores com maior grau obtido para cada período: t1=1990-1992, t2=1993-1995, t3=1996-1998, t4=1999-2001, t5=2002-2004, t6=2005-2007, t7=2008-2010.

6. Conclusões

Neste trabalho foi apresentada uma proposta para a identificação e caracterização de redes de coautoria de grupos de grande porte cadastrados na Plataforma Lattes. As métricas consideradas podem elucidar a compreensão da estrutura e dinâmica de colaboração em redes de coautoria acadêmica. Adicionalmente foi apresentada uma forma de análise da evolução, ao longo de períodos igualmente espaçados, de redes de coautoria.

Embora seja louvável a tarefa de processamento automático de grupos de grande porte, deve-se perceber que as análises estão condicionadas/limitadas aos dados contidos no banco de dados gerado localmente (Seção 2). Tanto atualizações de currículos Lattes, quanto o registro de novos currículos Lattes após a data de obtenção de todos os dados da Plataforma Lattes (maio de 2011) não são considerados. Para contornar essa limitação, o banco de dados local de currículos Lattes deverá ser atualizado com versões atualizadas e inserções dos novos currículos. Por outro lado, também deve-se perceber que, o banco de dados local é uma fonte rica para análises bibliométricas/cientométricas que se concentrem na produção acadêmica histórica de pesquisadores cadastrados na Plataforma Lattes.

Como trabalhos futuros destacamos os seguintes direcionamentos: (i) a caracterização das redes de coautoria de pesquisadores associados às oito 'Grandes Áreas do Conhecimento' do CNPq; e (ii) a caracterização e avaliação da evolução temporal de redes de coautoria, discretizadas por triênios, considerando todo o banco de dados de currículos Lattes. Certamente, estes trabalhos permitirão descobrir a presença (ou ausência) de padrões da inerente interação acadêmica entre pesquisadores atuantes na Ciência Brasileira.

Referências

Balancieri, R., Bovo, A. B., Medina, V., Pacheco, R., and Barcia, R. M. (2005). A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na plataforma lattes. *Ciência da Informação*, 34:64–77.

Borgatti, S. P. and Everett, M. G. (2000). Models of core/periphery structures. *Social Networks*, 21(4):375 – 395.

Brito-Cruz, C. H. (2012). A parada no crescimento do ensino superior. http://www.jornaldaciencia.org.br/Detalhe.jsp?id=81267. Último acesso em 09/04/2012.

CNPq (2007). Plataforma lattes alcança 1 milhão de currículos. http://www.cnpq.br/saladeimprensa/noticias/2007/0820c.htm. Último acesso em 02/04/2012.

Colizza, V., Flammini, A., Serrano, M. A., and Vespignani, A. (2006). Detecting rich-club ordering in complex networks. *Nature Physics*, 2(2):110–115.

Kang, I. S., Na, S. H., Lee, S., Jung, H., Kim, P., Sung, W. K., and Lee, J. H. (2009). On co-authorship for author disambiguation. *Information Processing and Management*, 45(1):84–97.

Liu, X., Bollen, J., Nelson, M., and de Sompel, H. V. (2005). Co-authorship networks in the digital library research community. *Informations Processing and Management*, 41(6):1462–1480.

Maia, M. F. and Caregnato, S. E. (2008). Co-autoria como indicador de redes de colaboração científica. *Perspectivas em Ciência da Informação*, 13(2):18–31.

Mena-Chalco, J. P. and Cesar-Jr., R. M. (2009). scriptLattes: An open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society*, 15(4):31–39.

Mena-Chalco, J. P. and Cesar-Jr., R. M. (2011). Towards automatic discovery of co-authorship networks in the brazilian academic areas. In *IEEE Seventh International Conference on e-Science Workshops 2011 (eScienceW)*, pages 53–60. IEEE.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88.

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, 89:208701+.

S. Milgram, S. (1967). The small world problem. *Psychology today*, 2(1):60–67.

Wasserman, S. and Faust, K. (1994). *Social network analysis: methods and applications*. Cambridge University Press.