Análise da Rede de Coautoria do Simpósio Brasileiro de Bancos de Dados

Péterson S. Procópio Jr, Alberto H. F. Laender, Mirella M. Moro

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
{peterson, laender, mirella}@dcc.ufmg.br

Abstract. Este artigo apresenta uma análise da rede de coautoria do Simpósio Brasileiro de Banco de Dados (SBBD) que em 2010 completou 25 anos de existência, consolidando-se como o maior e mais importante evento da América Latina para apresentação e discussão de resultados de pesquisa relacionados à área de bancos de dados. Para isso, coletamos dados bibliográficos de todas as suas 25 edições e apresentamos uma série de estatísticas, tais como média de artigos por autor, média de artigos por edição, média de coautores por artigo, entre outras. Além disso, construímos e analisamos a rede de coautoria do SBBD, fazendo uma análise tanto de suas características estruturais quanto de sua evolução temporal. Mostramos, ainda, que a rede em questão segue um fenômeno típico de diversas outras redes sociais, conhecido como mundo pequeno.

Categories and Subject Descriptors: H.Information Systems [**H.m. Miscellaneous**]: Databases

General Terms: Human Factors, Languages

Keywords: Redes Sociais, Redes de Coautoria, Simpósio Brasileiro de Bancos de Dados, SBBD

1. INTRODUÇÃO

Uma rede social é um conjunto de indivíduos ou grupos que possuem relacionamentos de algum tipo entre si [Newman 2001]. Redes de amizade entre pessoas, de contatos profissionais ou de afiliação institucional são exemplos de redes sociais. No estudo das redes sociais, indivíduos ou grupos são chamados de atores (*actors*) e os relacionamentos entre eles de laços (*ties*). A análise dessas redes nos permite entender o comportamento social de seus atores.

Uma rede de coautoria entre pesquisadores, também chamada de rede de colaboração científica, é um exemplo particular de uma rede social. Neste caso, os atores são autores de artigos e um laço existe entre dois autores se eles já escreveram algum artigo juntos. A produção de um artigo pode ser considerada como uma forma de documentação da colaboração entre dois ou mais autores [Newman 2004]. Assim, a rede de coautoria de uma comunidade pode revelar fatos interessantes a seu respeito, como, por exemplo, a existência de grupos que colaboram mais densamente, a identificação de relacionamentos mais intensos entre determinados autores ou ainda autores com maior grau de colaboração. O estudo de redes de coautoria pode ser utilizado também para se comparar os padrões de colaboração entre diferentes comunidades científicas.

Este artigo apresenta uma análise da rede de coautoria do Simpósio Brasileiro de Bancos de Dados (SBBD). O SBBD é o maior e mais importante evento da América Latina para apresentação e discussão de resultados de pesquisa relacionados à área de bancos de dados. Em 2010, o SBBD completou 25

Este trabalho é parcialmente financiado pelo InWeb (processo MCT/CNPq/FAPEMIG no. 573871/2008-6), pelo projeto Amanajé (processo MCT/CNPq no. 479541/2008-6) e por auxílios individuais a pesquisa do CNPq e FAPEMIG.

anos de existência, sendo este um momento oportuno para conhecermos um pouco mais a respeito dessa comunidade científica. Para isso, coletamos dados bibliográficos de todas as 25 edições do SBBD e apresentamos uma série de estatísticas, tais como média de artigos por autor, média de artigos por edição, média de coautores por artigo, entre outras. Além disso, construímos e analisamos a rede de coautoria do SBBD, fazendo uma análise tanto de suas características estruturais quanto de sua evolução temporal. Mostramos, ainda, que a rede em questão segue um fenômeno típico de diversas outras redes sociais, conhecido como mundo pequeno [Milgram 1967].

O artigo está organizado da seguinte forma. A Seção 2 descreve trabalhos relacionados ao estudo de redes de coautoria. A Seção 3 introduz alguns conceitos fundamentais necessários para se efetuar a análise de uma rede de coautoria. A Seção 4 apresenta estatísticas sobre os dados bibliográficos coletados e a análise da rede de coautoria do SBBD. Finalmente, a Seção 5 conclui este trabalho.

2. TRABALHOS RELACIONADOS

Nesta seção, apresentamos uma descrição sucinta de alguns trabalhos relacionados ao estudo de redes de coautoria. Newman [2004] apresenta um estudo das redes de coautoria de diferentes áreas do conhecimento, dentre elas Biomedicina, Física e Matemática. Nas redes estudadas, a grande maioria dos autores possui poucos colaboradores, enquanto uma pequena minoria possui muitos. De modo geral, o coeficiente de agrupamento das redes é alto e o caminho mínimo médio é baixo. Além disso, pesquisadores da área de Biomedicina, em geral, possuem mais colaboradores que os da área de Física e Matemática e, curiosamente, é menos provável que dois colaboradores de um autor da área de Biomedicina colaborem entre si do que na área de Matemática.

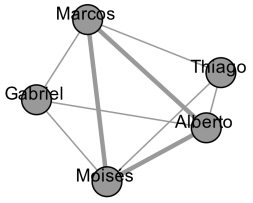
Liu et al. [2005] apresentam um estudo sobre área de Bibliotecas Digitais a partir da rede de coautoria derivada de suas três mais importantes conferências. Nesse estudo, os autores analisam diversos aspectos dessa rede, dentre eles os principais componentes conectados e o coeficiente de agrupamento. Trabalhos semelhantes foram realizados para diversas outras comunidades científicas, dentre elas as comunidades da *ACM SIGMOD International Conference on Management of Data* [Nascimento et al. 2003], da área de Engenharia Reversa de Software [Hassan and Holt 2004], do periódico *Data and Knowledge Engineering* [Huang et al. 2008].

Barabâsi et al. [2002] analisam a rede de coautoria obtida a partir dos dados de periódicos relevantes das áreas de Matemática e Neurociências relativos a um período de oito anos $1991 - 1998$. Os autores destacam a importância de se estudar a rede como um sistema dinâmico, que evolui com o tempo. Eles apresentam e explicam a evolução temporal de diversas medidas, como diâmetro, coeficiente de agrupamento, tamanho do maior componente conectado e distribuição do grau dos nós. Börner et al. [2005] analisam o impacto de grupos de pesquisa com base no número de publicações de seus integrantes e de suas respectivas citações em escala local e global. O estudo caracteriza as propriedades e a evolução de um novo campo científico, denominado Visualização da Informação.

Menezes et al. [2009] apresentam uma análise geográfica da produção científica em Ciência da Computação a partir das redes de coautoria de três diferentes regiões: Brasil, América do Norte (Estados Unidos e Canadá) e Europa (França, Inglaterra e Suiça). O estudo caracteriza as diferenças entre os padrões de colaboração nas três regiões, mostrando, entre outros resultados, a existência de grupos de colaboração isolados na rede de colaboração Européia, em contraste com o grau de conectividade encontrado nas redes do Brasil e da América do Norte.

3. CONCEITOS FUNDAMENTAIS

Existem diversas abordagens para se modelar redes de coautoria. Liu et al. [2005] apresentam três abordagens distintas e discutem métricas para cada uma delas. Neste artigo, modelamos uma rede de coautoria como um grafo não direcionado, ponderado, em que cada nó representa um autor e existe



—Gabriel Silva Goncalves, Moisés G. de Carvalho, Alberto H. F. Laender, Marcos André Gonçalves: Seleção Automática de Exemplos de Treino para um Método de Deduplicação de Registros baseado em Programação Genética

—Moisés G. de Carvalho, Alberto H. F. Laender, Marcos André Gonçalves, Thiago C. Porto: The Impact of Parameter Setup on a Genetic Programming Approach to Record Deduplication

Fig. 1. Exemplo de formação de uma rede de coautoria.

uma aresta entre dois nós se eles representam autores que já publicaram algum artigo em coautoria. O peso das arestas é relativo ao número de artigos publicados em coautoria por dois autores e representa a intensidade da colaboração entre eles. A Figura 1 ilustra a formação de uma rede de coautoria considerando apenas dois artigos.

Diversas medidas interessantes podem ser obtidas ao se modelar uma rede de coautoria como um grafo. A seguir, são apresentadas as principais medidas consideradas neste artigo.

Um *componente conectado* de uma rede é um subconjunto de nós em que existe um caminho entre quaisquer dois nós desse subconjunto. A análise do componente conectado geralmente consiste em medir o tamanho do *componente gigante* (ou *maior componente conectado*). O tamanho de um componente é dado pela fração de nós da rede que fazem parte desse componente. Uma rede de coautoria geralmente é formada por vários componentes conectados. A análise desses componentes pode ajudar a entender, por exemplo, como a rede foi formada [Menezes et al. 2009]. Além disso, diversas medidas para análise da rede somente são aplicadas a componentes conectados.

A *distância entre dois nós* de uma rede é o comprimento do menor caminho entre eles. O *diâmetro* de uma rede é a maior distância entre quaisquer dois de seus nós. Ao se pensar em uma aresta como um meio de conduzir informação entre os nós, pode-se dizer que a informação pode se espalhar mais rapidamente em redes com diâmetro menor. O *caminho mínimo médio* de uma rede é obtido determinando-se o menor caminho entre todos os pares de nós da rede e depois calculando-se a média de seu comprimento. Redes com um caminho mínimo médio pequeno também conduzem informação mais rapidamente.

O *grau de um nó* é obtido calculando-se o número de arestas adjacentes a ele, no caso de uma rede de coautoria representada por um grafo não direcionado. Autores com mais colaborações correspondem a nós de alto grau. O *grau médio* de uma rede é a média do grau de todos os seus nós, sendo um indicativo da quantidade média de colaborações de cada autor em um determinado período de tempo. Como reportado por Newman [2004], o grau médio de uma rede de coautoria pode variar bastante conforme a comunidade científica sendo analizada.

O *coeficiente de agrupamento* (*clustering coefficient*) de um nó representa o quanto os seus vizinhos estão conectados entre si. Mais especificamente, esse coeficiente é dado pela fração entre o número de arestas existentes entre os vizinhos diretos de um nó e o número de arestas que poderiam existir entre eles [Nascimento et al. 2003]. O coeficiente de agrupamento de um nó i pode ser definido como $C_i = \frac{e}{(v*(v-1)/2}$, onde v é o número de vizinhos de i e e é o número de arestas existentes entre todos os vizinhos de i. Para um autor, o coeficiente de agrupamento indica o quanto seus colaboradores também colaboram entre si. O *coeficiente de agrupamento de uma rede* é definido como a média do coeficiente de agrupamento de todos os seus nós.

Finalmente, o conceito de mundo pequeno (*small world*) tem recebido muita atenção recentemente. Redes com um pequeno caminho mínimo médio e com um alto coeficiente de agrupamento são comumente classificadas como redes de mundo pequeno. Esse fenômeno é típico das redes sociais e foi identificado pela primeira vez no famoso estudo de Milgram [1967]. Watts e Strogatz [1998] definem

Total de artigos	550
Total de autores	821
Média de artigos por edição	22
Média de autores por edição	56
Média de artigos por autor	1,88
Média de coautores por artigo	2,80

Table I. Sumário dos dados.

uma rede como sendo um *mundo pequeno* se, dado um grafo randômico de tamanho similar, ela possui um caminho mínimo médio um pouco maior e um coeficiente de agrupamento muito maior do que o desse grafo.

4. REDE DE COAUTORIA DO SBBD

4.1 Dados Coletados

Para construir a rede de coautoria do SBBD, coletamos dados dos artigos publicados nos anais de todas as suas edições, abrangendo os anos de 1986 a 2010. Palestras e tutoriais não foram computados, uma vez que estávamos interessados em estudar a colaboração entre os pesquisadores dessa comunidade por meio de suas coautorias em artigos científicos. Como as primeiras edições do SBBD foram realizadas em uma época anterior à popularização da Web, elas não possuem os dados de seus artigos disponíveis em formato digital. Desta forma, a coleta de dados envolveu fontes digitais e não digitais. Os dados das edições de 1986 a 1998 foram coletados manualmente, diretamente dos anais dessas edições, enquanto que os das edições de 1999 a 2010 foram extraídos automaticamente da DBLP [Ley 2002]. Por fim, os dados oriundos das duas fontes foram integrados e armazenados em um banco de dados relacional. É importante ressaltar que durante esse processo foi realizada a deduplicação dos nomes dos autores, de forma semi-automática, a fim de garantir a consistência dos dados para o estudo.[1]

A coleta de dados resultou em um total de 550 artigos e 821 autores. Em média, em cada edição do SBBD foram aceitos 22 artigos envolvendo 56 autores. Cada autor publicou, em média, 1,88 artigo, sendo de 2,80 a média de coautores por artigo. A Tabela I sumariza esses dados.

4.2 Estatísticas

A Figura 2 ilustra a rede de coautoria construída a partir dos dados coletados sobre os artigos publicados nos anais do SBBD. Cada nó representa um autor e cada aresta representa a colaboração entre dois autores. O tamanho do nó é proporcional ao número de artigos publicados pelo autor. A espessura da aresta é proporcional ao número de artigos que dois autores já publicaram juntos. A cor dos nós identifica o componente conectado a que pertencem. Como podemos observar, o maior componente conectado da rede é grande, abrangendo, como veremos na Seção 4.3, cerca de 70% de seus nós, o que significa tratar-se uma comunidade bastante coesa.

A Figura 3 mostra a evolução do número de artigos e do número de autores ao longo dos 25 anos do SBBD. É possível perceber que o número de artigos e o número de autores estão altamente correlacionados. O número de autores cresceu mais rapidamente que o número de artigos, o que mostra que o SBBD tem atraído novos autores ao longo de sua existência. Entretanto, isso não significa que a sua comunidade colabora pouco entre si, como será visto mais adiante.

A Figura 4 apresenta a distribuição do número de coautores por artigo. Observa-se que 43% dos artigos têm dois coautores e cerca de 28% têm três coautores. Curiosamente, foi verificada a existência de um artigo com 18 coautores, fato pouco comum em se tratando de um evento da área de Ciência

[1]Os dados coletados estão disponíveis em `http://homepages.dcc.ufmg.br/~peterson/sbbd_data.tar.gz`

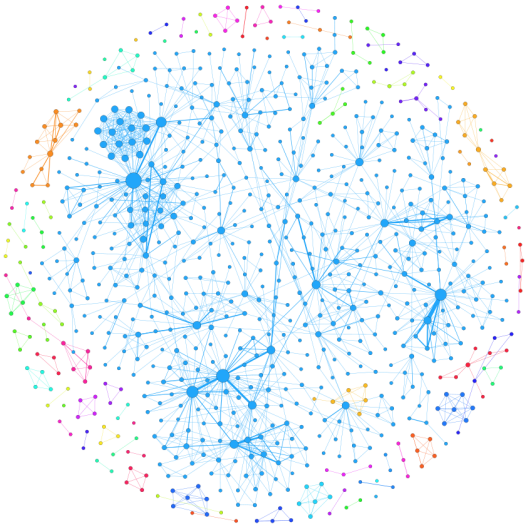


Fig. 2. Rede de coautoria do SBBD.

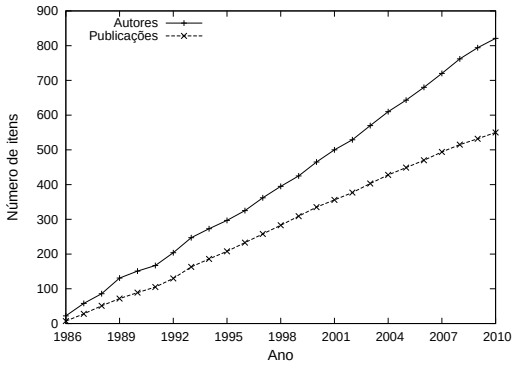


Fig. 3. Evolução do número de autores e artigos ao longo do tempo.

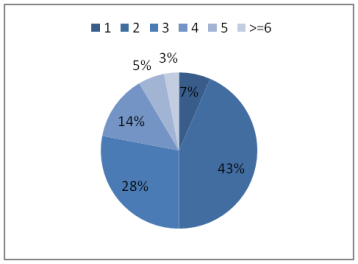


Fig. 4. Número de coautores por artigo.

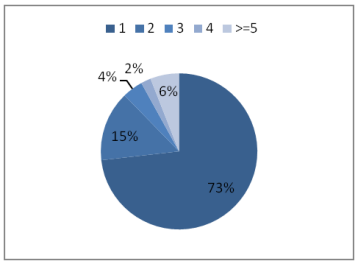


Fig. 5. Número de artigos por autor.

da Computação. Uma análise semelhante pode ser feita verificando-se a média de artigos por autor. A Figura 5 apresenta esses dados. Como é possível observar, a grande maioria dos autores (cerca de 73%) publicou apenas uma única vez no SBBD, aproximadamente 15% publicaram duas vezes e menos de 6% publicaram cinco vezes ou mais.

A Tabela II complementa esses dados apresentando a lista dos 10 autores mais prolíficos e dos 10 autores com mais colaboradores. Como podemos observar, os autores mais prolíficos tendem a ser

10 Autores Mais Prolíficos		10 Autores com Mais Colaboradores	
Alberto H. F. Laender	33	Marta Lima de Queirós Mattoso	58
Caetano Traina Jr.	29	Alberto H. F. Laender	46
Marta Lima de Queirós Mattoso	28	Caetano Traina Jr.	39
Marco Antônio Casanova	18	Marcos André Gonçalves	39
Rubens Nascimento Melo	17	Jano Moreira Souza	33
Claudia Bauzer Medeiros	16	Marco Antônio Casanova	25
Marcos André Gonçalves	16	Wagner Meira Jr.	25
Agma Juci Machado Traina	15	Altigran S. da Silva	24
Berthier A. Ribeiro-Neto	15	Rubens Nascimento Melo	23
Wagner Meira Jr.	15	Berthier A. Ribeiro-Neto	23

Table II. 10 autores mais prolíficos e com mais colaboradores.

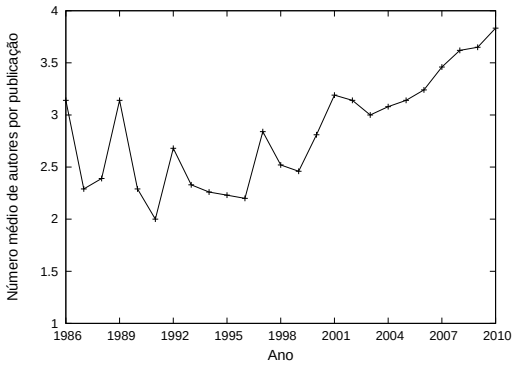


Fig. 6. Média de autores por artigo ao longo do tempo.

aqueles com maior grau de colaboração, ou seja, que possuem mais colaboradores. Dos 10 autores mais prolíficos do SBBD, oito também se encontram na lista dos autores com mais colaborações.

Finalmente, para determinar se a rede de coautoria do SBBD poderia ser classificada como um mundo pequeno, geramos um grafo randômico de tamanho similar e calculamos para ambos os respectivos coeficientes de agrupamento e o caminho médio. O coeficiente de agrupamento calculado para a rede do SBBD foi 0,61, enquanto que o do grafo randômico similar foi 0,07. Já o caminho mínimo médio calculado para a rede do SBBD é de aproximadamente 6,31, enquanto que o do grafo de tamanho similar é de 5,42. Assim, como era de se esperar, a rede de coautoria do SBBD, como outras redes de colaboração científica [Hassan and Holt 2004; Menezes et al. 2009; Nascimento et al. 2003], segue o fenômeno do mundo pequeno.

4.3 Análise Temporal

Uma primeira análise a ser feita relativa ao aspecto temporal da rede de coautoria do SBBD é a média de autores por artigo ao longo do tempo. Observando-se a Figura 6, é possível perceber que a média de autores por artigo vem aumentando ao longo das últimas edições do SBBD. Tal fato indica que, nos últimos anos, pesquisadores têm participado de equipes de tamanhos crescentes.

A Figura 7 nos permite observar a evolução temporal dos dois maiores componentes conectados da rede. É possível verificar que no último ano o maior componente conectado (componente gigante) representa cerca de 70% da rede. Esse componente também cresce ao longo do tempo à medida que outros componentes vão se conectando a ele. Tal fato é coerente com a evolução do diâmetro da rede, observado na Figura 8. À medida que o tamanho do maior componente conectado aumenta, o diâmetro da rede também aumenta. Posteriormente, à medida que o maior componente vai se tornando mais conectado, o diâmetro tende a diminuir.

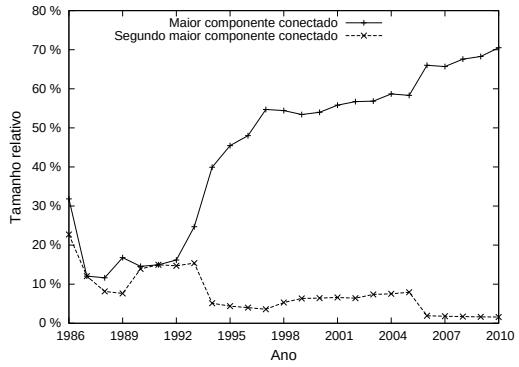


Fig. 7. Tamanho relativo dos componentes conectados.

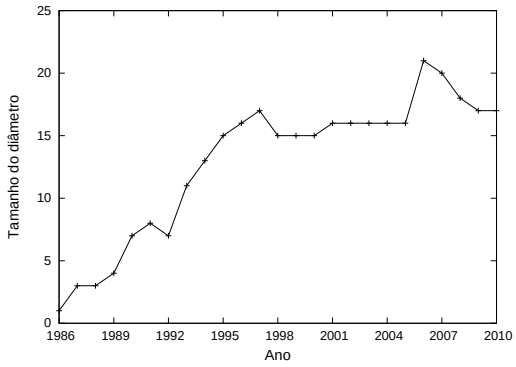


Fig. 8. Evolução do diâmetro da rede.

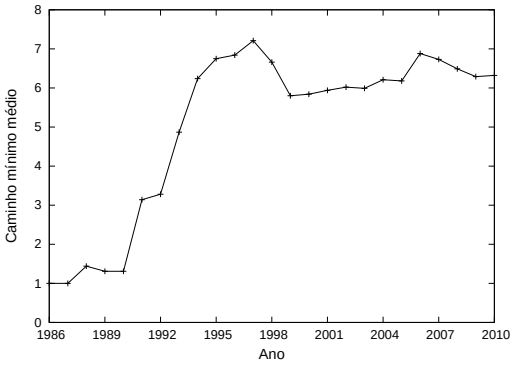


Fig. 9. Evolução do caminho mínimo médio da rede.

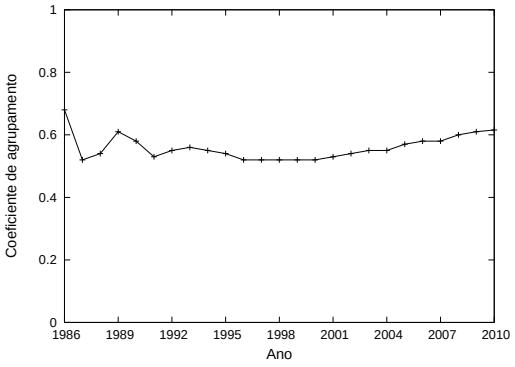


Fig. 10. Evolução do coeficiente de agrupamento da rede.

Outro fator interessante de ser observado é o tamanho do segundo maior componente conectado. Nota-se na Figura 7 que o seu tamanho relativo ao longo dos quatro últimos anos é menor que 1%. Isto mostra que a rede é formada por um maior componente conectado de tamanho expressivo e por diversos componentes de pequeno tamanho.

A evolução do caminho mínimo médio da rede é mostrada na Figura 9. Observa-se que esta medida segue um padrão semelhante ao da evolução do diâmetro. Já o coeficiente de agrupamento da rede, analisado na Figura 10, possui um crescimento mais estável. No primeiro ano ele atinge o seu valor mais alto, de 0,68. Isto se deve ao maior componente conectado nesse ano ser fortemente conectado. No ano de 1989 esse coeficiente sofre um rápido aumento, justificado pela publicação de um artigo com 18 coautores. Dezesseis desses coautores publicaram no SBBD pela primeira vez naquele mesmo ano. Como consequência, seu grau de agrupamento na rede no ano é igual a 1(todos os seus vizinhos colaboram entre si). Isto provocou o aumento do grau de agrupamento da rede como um todo. A publicação desse artigo também fez com que o maior componente conectado da rede aumentasse, como pode ser visto na Figura 7. Nos anos seguintes, o coeficiente de agrupamento decresce um pouco apresentando um ligeiro crescimento a partir de 1991. No ano de 2010, o seu valor é de 0,61, similar ao de outras redes de coautoria estudadas [Hassan and Holt 2004].

5. CONCLUSÃO E TRABALHOS FUTUROS

Este artigo apresentou uma análise da rede de coautoria do Simpósio Brasileiro de Bancos de Dados. Foi mostrado que essa rede, assim como outras redes sociais, pode ser considerada um mundo pequeno. O coeficiente de agrupamento da rede é de 0,61, enquanto que o tamanho do caminho mínimo médio

é de aproximadamente 6,32. O maior componente conectado da rede tem tamanho relativo de 70%. Essas medidas se assemelham às de outras redes de colaboração encontradas na literatura. A rede SIGMOD [Nascimento et al. 2003], por exemplo, apresenta um coeficiente de agrupamento de 0,70, um tamanho relativo do maior componente conectado de 59% e um caminho mínimo médio de 5,65. Já a rede de colaboração da área de Ciência da Computação no Brasil, estudada em [Menezes et al. 2009], apresenta para as mesmas medidas, respectivamente, os valores de 0,30, 78,15% e 6,47.

A média de coautores por artigo do SBBD tem aumentado ao longo dos anos, o que mostra que os pesquisadores da comunidade brasileira de Bancos de Dados tendem a atuar em grupos de tamanho crescente. O diâmetro da rede de coautoria também cresceu ao longo do tempo, mas sofreu uma diminuição nos últimos cinco anos. O mesmo fenômeno foi observado no estudo de Menezes et al. [2009], de modo que essa diminuição do diâmetro da rede pode ser explicada pela existência de novas colaborações entre os pesquisadores do seu maior componente conectado, fazendo com que a distância entre dois pesquisadores quaisquer nesse componente tenda a diminuir. O aumento inicial do diâmetro da rede pode ser visto como natural e se explica devido ao aumento do tamanho do seu maior componente conectado.

O tamanho do caminho mínimo médio e do maior componente conectado da rede de coautoria do SBBD seguem um padrão de crescimento semelhante ao do diâmetro, mostrando uma correlação entre essas três medidas. Foi mostrado ainda que, no caso dessa rede, existe uma grande correlação entre os pesquidores mais prolíficos e aqueles com com alto grau de colaboração.

Como trabalho futuro, podemos considerar uma análise mais detalhada do comportamento da rede de coautoria do SBBD ao longo dos anos, buscando-se identificar a causa de variações em algumas medidas, como o tamanho do seu maior componente conectado e do coeficiente de agrupamento. Outra análise interessante a ser feita teria como objetivo determinar a existência de alguma correlação entre as colaborações existentes e aspectos específicos relacionados aos autores, como, por exemplo, localização geográfica e instituição de origem.

REFERENCES

Barabâsi, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications* 311 (3-4): 590–614, 2002.

Börner, K., Dall'Asta, L., Ke, W., and Vespignani, A. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams: Research Articles. *Complexity - Understanding Complex Systems* 10 (4): 57–67, March, 2005.

Hassan, A. E. and Holt, R. C. The Small World of Software Reverse Engineering. In *Proceedings of the Working Conference on Reverse Engineering*. Delft, the Netherlands, pp. 278–283, 2004.

Huang, J., Zhuang, Z., Li, J., and Giles, C. L. Collaboration over time: characterizing and modeling network evolution. In *Proceedings of the International Conference on Web Search and Web Data Mining*. Stanford, CA, USA, pp. 107–116, 2008.

Ley, M. The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In *Proceedings of the International Symposium on String Processing and Information Retrieval*. Lisbon, Portugal, pp. 1–10, 2002.

Liu, X., Bollen, J., Nelson, M. L., and Van de Sompel, H. Co-authorship networks in the digital library research community. *Inf. Process. Manage.* 41 (6): 1462–1480, December, 2005.

Menezes, G. V., Ziviani, N., Laender, A. H., and Almeida, V. A geographical analysis of knowledge production in computer science. In *Proceedings of the International Conference on World Wide Web*. Madrid, Spain, pp. 1041–1050, 2009.

Milgram, S. The Small World Problem. *Psychology Today* 1 (1): 60–67, 1967.

Nascimento, M. A., Sander, J., and Pound, J. Analysis of SIGMOD's co-authorship graph. *SIGMOD Rec.* 32 (3): 8–10, 2003.

Newman, M. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences* 101 (1): 5200–5205, 2004.

Newman, M. E. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America* 98 (2): 404, 2001.