Escola Regional de Banco de Dados em 15 edições: *Um Levantamento Bibliométrico*

Fernanda M. de Souza[1], **Vinícius Gasparini**[1], **Denio Duarte**[2] **e Rebeca Schroeder**[1]

[1]Universidade do Estado de Santa Catarina (UDESC)
Departamento de Ciências da Computação
Joinville - SC - Brazil

[2]Universidade Federal da Fronteira Sul (UFFS)
Chapecó - SC - Brazil
{fmd.souza,v.gasparini}@edu.udesc.br,
duarte@uffs.edu.br, rebeca.schroeder@udesc.br

Abstract. *This paper presents an analysis of the bibliographic production of the Escola Regional de Banco de Dados (ERBD), which in 2019 completed 15 editions. To this end, the main metadata of the papers published until now were collected. Given the unavailability of a repository with all this data, this paper contributes to the construction of a database with information from all editions of the event. In addition, the statistics of papers, authors and collaborations provided by this work can contribute to the self-knowledge of this community and encourage reflection on its evolution.*

Resumo. *Este artigo apresenta uma análise da produção bibliográfica da Escola Regional de Banco de Dados (ERBD), que em 2019 completou 15 edições. Para tanto, foram coletados os principais metadados dos artigos publicados nestes 15 anos de evento. Dada a inexistência de um repositório que reúna todos estes dados, este trabalho contribui com a construção de um banco de dados com informações de todas as edições do evento. As estatísticas de artigos, autores e colaborações fornecidas por este artigo podem contribuir para o auto-conhecimento desta comunidade e favorecer a reflexão de sua evolução.*

1. Introdução

Desde 2005, a Escola Regional de Banco de Dados (ERBD) tem promovido anualmente a integração de alunos, professores e profissionais da área. Um dos objetivos do evento é incentivar estudantes a terem seus primeiros artigos publicados e apresentados. Apesar de ocorrer de maneira itinerante pelos estados da região Sul do Brasil, o evento recebe palestrantes e participantes de outras regiões. Em 15 edições completadas em abril de 2019, a ERBD vem se consolidando como o segundo maior evento da área de banco de dados no Brasil, perdendo apenas para o Simpósio Brasileiro de Bancos de Dados (SBBD)[SBC 2020].

Para celebrar seus 15 primeiros anos, este artigo examina a história da ERBD contada a partir da análise de dados extraídos dos anais do evento. Para isto, foram coletados dados bibliográficos de todas as 15 edições para apresentar uma série de indicadores relacionados à produção. Este é o primeiro trabalho a levantar e analisar dados da ERBD. Uma importante contribuição deste artigo está na construção de um banco de dados para

abrigar os dados coletados e permitir registrar a história do evento, dada a inexistência de um repositório que reúna todas estas informações. Além disto, são apresentados diversos indicadores por edição, como número de artigos, categorias e quantitativos de autores, bem como temas mais frequentes e autores mais produtivos. Por fim, a rede de coautoria da ERBD é apresentada, evidenciando a formação de grupos de pesquisadores.

O restante do artigo é organizado em mais 5 seções. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 apresenta detalhes da coleta de dados bibliográficos e do banco de dados criado. A Seção 4 apresenta e analisa dados do evento. Por fim, as conclusões e trabalhos futuros são apresentados na Seção 5.

2. Trabalhos Relacionados

A análise de comunidades científicas, e das redes sociais formadas por elas, é um tópico de pesquisa em evidência. [Brandão et al. 2017] analisaram a força das relações de quatro redes de coautoria extraídas de bases de dados como DBLP, PubMed e APS em alguns meses dos anos 2015 e 2016. Utilizando fast-RECAST*, as relações da rede foram separadas em quatro categorias: forte, ponte, fraco e aleatório. Os resultados mostraram que as relações das redes de coautoria são predominantemente fracas ou aleatórias, ou seja, as relações tendem a não permanecer ao longo do tempo, e mesmo as classificadas como forte tendem a perecer. Conclui-se que as relações fortes tenham esse comportamento devido à relação de professores e alunos, que tendem a parar de publicar juntos após o aluno se formar.

[Amblard et al. 2011] apresentaram a análise da rede formada por dados extraídos do hep-th (High Energy Physics Theory) do repositório arXiv no período de janeiro de 1992 a maio de 2003. Os dados foram estruturados tanto de forma estática quanto dinâmica, em duas categorias: coautoria e citação. Os dados estáticos mostram redes formadas por múltiplas ilhas conectadas por poucas pontes, com autores distantes entre si. Os dados dinâmicos têm um comportamento semelhante, entretanto, o coeficiente de agrupamento da rede de citações tem uma curva crescente e estabilizada, enquanto a rede de coautoria tem uma curva com grande oscilação e com valores maiores do que a rede de citações.

[Júnior et al. 2011] tiveram como foco a rede de coautoria do SBBD, coletando dados de 1986 até 2010, resultando em um total de 550 autores e 821 artigos. Foram analisados os autores mais prolíficos, aqueles que mais publicaram artigos, e constatou-se que, em sua maioria, também são os que tem o maior número de colaboradores. Outras medidas como diâmetro, coeficiente de agrupamento e caminho mínimo médio também foram aplicados, mostrando uma rede que tende a crescer no diâmetro, mas que diminui a partir do ano de 2005 devido a união de grupos de pesquisa.

[Kazi et al. 2017] analisaram 629814 artigos e 595774 autores na área de ciências da computação durante o período de 1955 a 2015. Sobre a rede, foi analisado a distância entre autores, e se a rede segue o padrão de mundo pequeno. O coeficiente de agrupamento tem um valor alto, mostrando que muitos autores estão conectados, mesmo assim, o maior componente tem 56.8% da rede. O fenômeno do mundo pequeno ocorre na rede, porém nos anos de 1985 a 1995 há uma diferença, provavelmente envolvendo os rápidos avanços da ciência da computação no período. Foram analisados autores mais detalhadamente, mostrando quais são os mais ativos e centrais da rede, além de aplicar o método

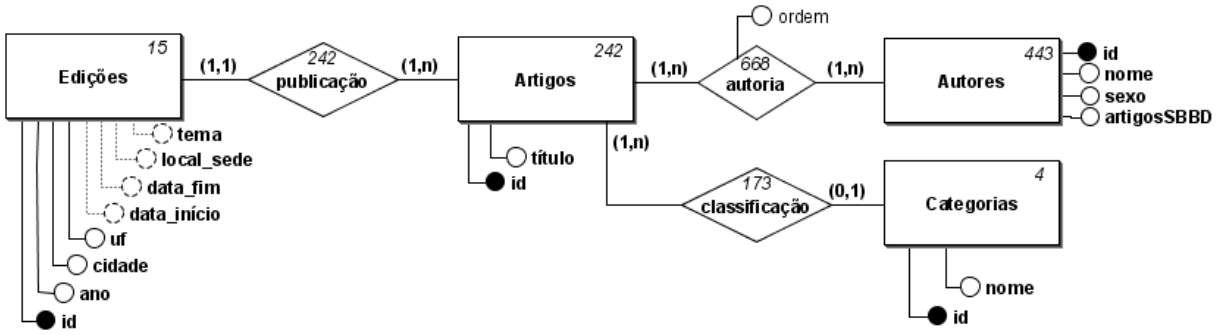

Figura 1. Esquema Conceitual dos dados coletados.

de *PageRank* para mostrar os autores mais citados.

Assim como os trabalhos levantados, este artigo pretende analisar uma comunidade científica extraindo alguns dados e características da mesma. Neste artigo, o foco está em levantar indicadores de artigos e autores, bem como qualificar colaborações e envolvimento da comunidade. Um importante diferencial deste trabalho está na construção de um banco de dados para reunir as informações essenciais da comunidade da ERBD. Trata-se de uma contribuição relevante pela inexistência de uma biblioteca digital disponível que contenha todas estas informações. Ressalta-se que os dados de algumas edições do evento precisaram ser recuperadas diretamente dos anais impressos ou em CDs, conforme apresentado pela próxima seção.

3. Coleta de Dados

A produção bibliográfica analisada por este trabalho foi extraída majoritariamente da Biblioteca Digital Brasileira de Computação (BDBComp)[1], na qual se encontram disponíveis os artigos de algumas edições da ERBD[2]. Os dados das edições de 2011, 2017, 2018 e 2019 foram obtidos diretamente dos respectivos *sites* ainda disponíveis no momento da coleta. Os dados da edição de 2008 foram obtidos dos anais impressos de posse de um dos autores deste artigo. A Figura 1 apresenta o esquema conceitual do banco de dados relacional produzido a partir desta extração que compreende os anos de 2005-2019.

Como pode ser observado na Figura 1, basicamente os metadados dos artigos publicados foram coletados. O número total de instâncias coletadas foi agregado às respectivas entidades e relacionamentos na figura, seguindo a notação de [Batini et al. 1992]. Em relação à entidade `Autores`, o atributo `sexo` foi manualmente informado para cada um dos autores. Na relação de `autoria`, o campo `ordem` atribui a posição sequencial na lista de autores de cada artigo. Esta informação é importante para identificar os autores principais (primeiro autor) dos artigos.

Quanto a entidade `Categorias`, é importante observar que artigos da ERBD eram originalmente classificados em `Artigos de Graduação` e `Artigos de Pós-graduação`. Com o passar das edições, o evento optou por classificá-los em `Artigos de Pesquisa` e `Artigos de Aplicações/Experiências`. Infelizmente na BDBComp os artigos não se encontram classificados, por esta razão não foi

[1]http://www.lbd.dcc.ufmg.br/bdbcomp

[2]Os dados foram extraídos em Abril de 2019, para os anos 2005-2007, 2009, 2010, 2012-2016

possível obter a classe de 69 artigos do evento. Por fim, a entidade `Edições` registra detalhes de cada uma das 15 edições da ERBD. Os atributos data (início e fim), tema e local (instituição sede) foram obtidos pela coleta de informações diretamente dos sites ainda disponíveis ou dos anais, e foram definidos como opcionais, pois não foi possível recuperá-los de todas as edições.

Um banco de dados relacional foi construído utilizando o PostgreSQL [PostgreSQL 1996] para o armazenamento e gerenciamento dos dados. Além disto, um banco de dados em grafo utilizando o Neo4J [Neo4J 2010] foi produzido a partir das relações entre artigos e autores, para facilitar a análise e visualização da rede de coautoria da ERBD. No grafo, os nós representam os autores e as arestas coautorias entre eles em artigos.

Uma coleta complementar buscou identificar autores da ERBD que também publicaram entre 2005 e 2019 no fórum nacional, o Simpósio Brasileiro de Bancos de Dados (SBBD). Para este fim, foi desenvolvido um extrator para buscar na base DBLP[3] os nomes de autores de artigos da trilha principal do SBBD, e comparar com a lista de autores da ERBD. Foram contabilizados o total de artigos por autor, representado pelo atributo `artigosSBBD` na entidade `Autores`. Eventuais variações nos nomes de autores foram avaliados utilizando uma combinação dos algoritmos de *Levenshtein* e *Jaro* para identificar nomes similares, os quais foram também conferidos de forma manual.

Para desambiguar os nomes dos autores da ERBD, no caso de várias publicações para um mesmo autor, foi desenvolvido um procedimento que age de forma semi-automática em duas etapas. A implementação considera que o formato do nome é sempre o primeiro nome por extenso seguido de um sobrenome. Na primeira etapa ocorre a retirada de toda ocorrência idêntica de um nome, bem como o agrupamento de todos os nomes em sublistas de seus primeiros nomes. A segunda etapa tem o objetivo de selecionar sobrenomes que estejam contidos em outros pertencentes à mesma sublista, considerando abreviações, como, por exemplo Abreu e A. ou Júnior e Jr. Em caso afirmativo, esses nomes são considerados pertencentes à mesma pessoa. Este método teve uma assertividade de 99,3% na identificação de autores duplicados, sendo que os erros de identificação encontrados foram corrigidos manualmente.

4. Estatísticas e Análise de Dados

Para a análise e interpretação dos resultados foram realizadas consultas à base de dados construída, de acordo com as especificações de cada análise. A seguir são apresentados os resultados produzidos.

4.1. Estatísticas Gerais de Artigos

A Figura 2 apresenta a quantidade total de artigos publicados em cada edição. Para as edições em que foi possível obter a classificação de artigos, optou-se por representar o quantitativo parcial de cada categoria. Como mencionado na Seção 3, a classificação de artigos mudou comparado ao que era praticado nos primeiros anos do evento. Atualmente, e pelo menos nos últimos 10 anos, a ERBD aceita artigos em duas categorias: `Pesquisa` (com artigos de até 10 páginas), e `Aplicações/Experiências` (com artigos de até 4 páginas).

[3]https://dblp.uni-trier.de/

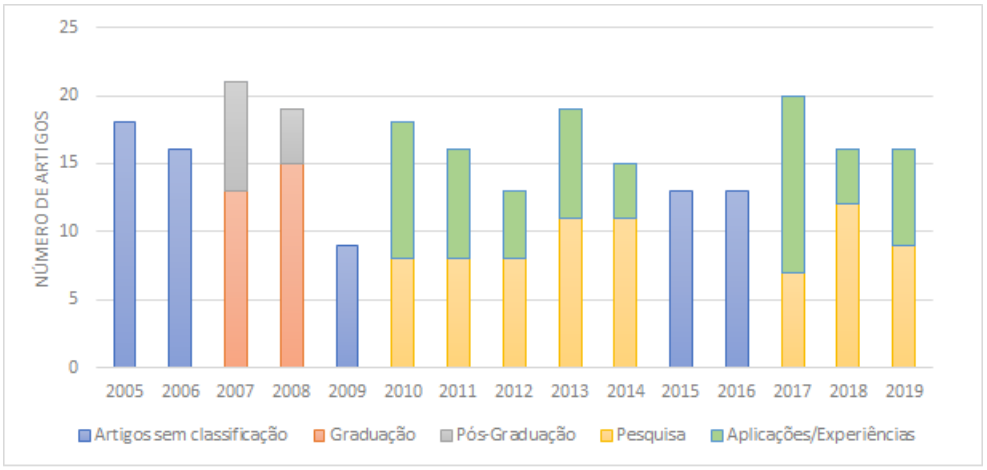

Figura 2. Número de Artigos por Edição e respectivas Classificações.

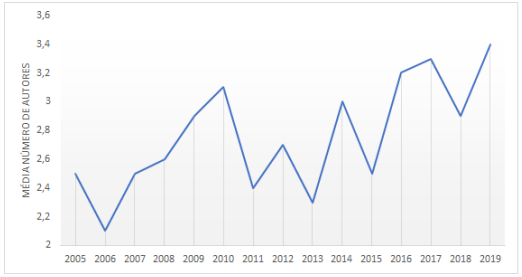

Figura 3. Média de Autores por Artigo.

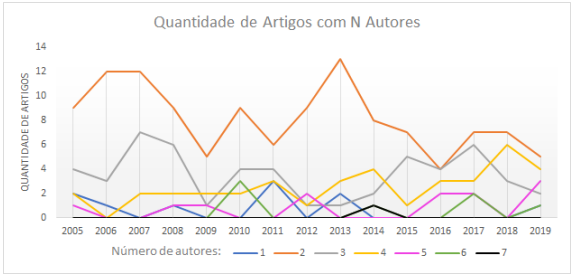

Figura 4. Número de Artigos com N Autores.

No período analisado, cada edição recebeu em média um total de 16 artigos. Nos últimos 10 anos, a categoria `Pesquisa` recebeu em média 9,3 artigos por edição, enquanto a categoria `Aplicações/Experiências` recebeu 7,4 aproximadamente. Assim, a categoria que aceita artigos completos recebeu a maior parte das submissões aceitas no evento.

A Figura 3 demonstra a média da quantidade de autores por artigo em cada edição. A partir da figura, é possível identificar uma tendência de crescimento no número de autores nos últimos anos. Esta tendência pode indicar um aumento na colaboração, dada pelo número de coautores por artigo. Entretanto, ao analisar a Figura 4 é possível identificar que ao longo dos anos a maior parte dos artigos permanece com 2 autores. O gráfico apresentado por esta figura mostra a quantidade de artigos que apresentam de 1 a 7 autores por edição. Apesar da maioria dos artigos apresentar 2 autores, percebe-se um crescimento na quantidade de artigos com 4 e 5 autores.

Os resultados apresentados pela Figura 5 mostram que a maior parte dos autores são homens. Para este levantamento considerou-se os autores distintos, contabilizando apenas uma vez os autores que publicaram mais de uma vez em uma dada edição. Ao longo das 15 edições da ERBD, dos 443 autores apenas 92 (21%) são mulheres. Este dado destaca a baixa participação feminina e enfatiza a importância das iniciativas nacionais de apoiar o ingresso deste público em cursos de áreas relacionadas à Ciência da Computação.

Os temas mais recorrentes identificados nos títulos dos artigos foram também

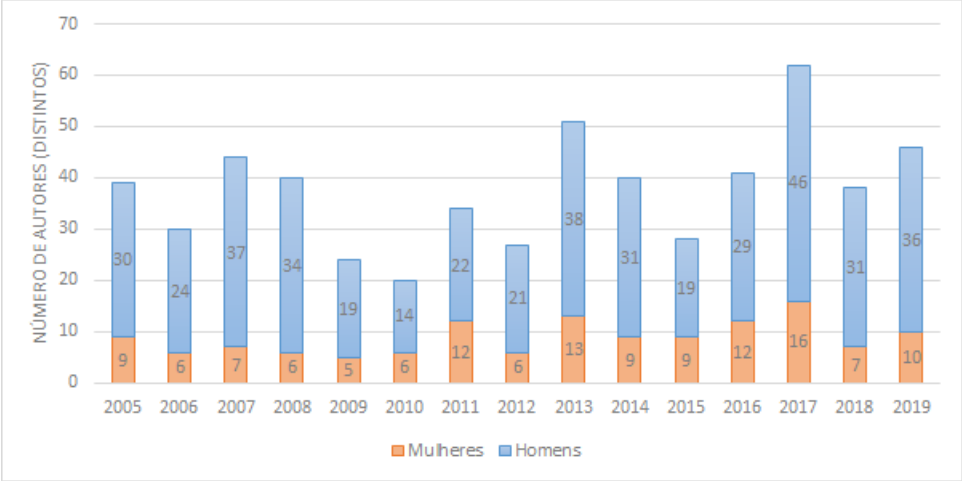

Figura 5. Número de Autores por Sexo.

Figura 6. Nuvem de Palavras - Títulos de Artigos (2015-2019).

identificados. Inicialmente optou-se por gerar nuvens de palavras para intervalos de 5 anos do evento. Entretanto, foi identificado que não houve muitas variações. Por esta razão, a Figura 6 apresenta os temas em evidência das edições de 2015 a 2019. Excetuando as *stop-words* bem como palavras gerais, destacam-se palavras como XML, extração, Web, Similaridade, Mineração e Recomendação. A seguir, são apresentadas algumas análises quanto aos autores e suas cooperações na ERBD.

4.2. Estatísticas de Autores

Ao considerar as 15 edições, os autores prolíficos da Escola Regional de Banco de Dados são evidenciados pela Tabela 1 que relaciona os autores que mais publicaram no evento, ordenados de forma decrescente pelo total de artigos publicados por cada um e o total de edições em que estes publicaram. Nesta tabela foram considerados autores com 5 ou mais artigos. Grande parte destes autores referem-se a professores da área de Banco de Dados, que possivelmente direcionaram seus alunos a publicar no evento. Este direcionamento pode ser atestado, em partes, pela Tabela 2 onde alguns destes autores são identificados

Tabela 1. Autores com 5 ou mais Artigos.

Autor	Total de Artigos	Total de Edições
Ronaldo Mello	27	11
Carina Dorneles	13	8
Sérgio Mergen	12	6
Renata Galante	11	5
Carmem Hara	10	8
Duncan Ruiz	10	7
Cristiano Cervi	10	7
Angelo Frozza	9	7
Nádia Kozievitch	9	5
Renato Fileto	9	4
Carlos Heuser	9	3
Eduardo Borges	8	5
Denio Duarte	7	6
Edimar Manica	6	6
Deise Saccol	6	4
Geomar Schreiner	6	4
Rebeca Schroeder	5	5
Gláucio R. Vivian	5	4

Tabela 2. Autores com 10 ou mais Coautores.

Autor	#coautores
Carmem Hara	28
Ronaldo Mello	28
Nádia Kozievitch	18
Duncan Ruiz	16
Carina Dorneles	14
Eduardo Borges	14
Renata Galante	14
Rebeca Schroeder	12
Carlos Heuser	11
Daniel Lichtnow	11
Sérgio Mergen	11
Angelo Frozza	10
Geomar Schreiner	10
Helena Ribeiro	10
Renato Fileto	10

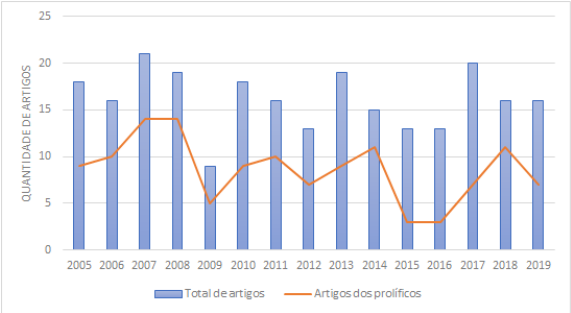

Figura 7. Total de Artigos e de Prolíficos.

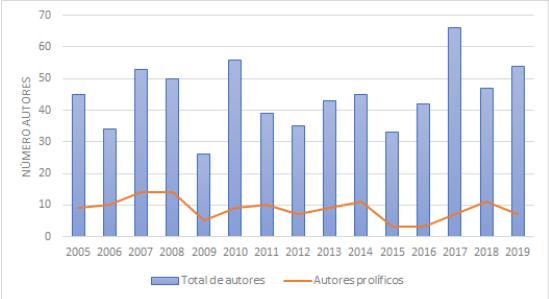

Figura 8. Total de Autores e Prolíficos.

como aqueles com o maior número de coautores. A Tabela 2 apresenta os autores que estão relacionados a 10 ou mais coautores distintos.

Com relação a expressividade na quantidade de publicações dos autores prolíficos, as Figuras 7 e 8 apresentam uma relação destes com o total de artigos, e com o total de autores distintos por ano, respectivamente. Em ambos os casos considerou-se os autores apresentados pela Tabela 1. Em média, os artigos dos prolíficos correspondem a 52% do total de artigos, assim como 22% do total de autores.

Com relação a colaboração entre autores, identificada por associações de coautoria, a Tabela 3 apresenta os pares de autores que publicaram juntos em 2 ou mais edições. Em diversos pares foi possível caracterizar a relação entre orientadores-alunos, indicando a importância da ERBD como um evento para abrigar os primeiros trabalhos de alunos, e contribuir em sua formação. Esta identificação foi possível através de uma análise exploratória nos respectivos currículos na Plataforma Lattes[4].

As colaborações ocorridas durante os primeiros 15 anos da ERBD, podem ser visualizadas de maneira geral pela rede de coautoria da Figura 9. Neste grafo, os nós correspondem a autores que se conectam uns aos outros quando há artigos publicados conjuntamente. É evidente a formação de grupos isolados ao longo dos anos. Entretanto, destaca-

[4]http://lattes.cnpq.br/

Tabela 3. Colaborações em 2 ou mais Edições.

Autor 1	**Autor 2**	**# artigos**	**# edições**
Ronaldo Mello	Angelo Frozza	6	5
Carina Dorneles	Renata Galante	6	4
Edimar Manica	Renata Galante	5	5
Cristiano Cervi	Gláucio R. Vivian	5	4
Carina Dorneles	Edimar Manica	4	4
Ronaldo Mello	Geomar Schreiner	4	3
Eduardo Borges	André Prisco	4	2
André S. Rosa	Carlos E. Pantoja	3	3
Carlos Heuser	Sérgio Mergen	3	3
Cristiano Cervi	Edimar Manica	3	3
Cristiano Cervi	Renata Galante	3	3
Angelo Frozza	Geomar Schreiner	3	2
Carlos E. Pantoja	João Victor Guinelli	3	2
Nádia Kozievitch	Rita Berardi	3	2
Sérgio Dill	Edson Padoin	3	2
Aldri Santos	Carmem Hara	2	2
André S. Rosa	João Victor Guinelli	2	2
Carmem Hara	Oliver M. Batista	2	2
Carmem Hara	Raqueline Penteado	2	2
Denio Duarte	Geomar Schreiner	2	2
Helena Ribeiro	Odacir D. Graciolli	2	2
Nádia Kozievitch	Keiko Fonseca	2	2
Rafael Garcia	Duncan Ruiz	2	2
Rodrigo Gonçalves	Ronaldo Mello	2	2
Ronaldo Mello	Cláudio Lima	2	2
Ronaldo Mello	Filipe R. Silva	2	2
Ronaldo Mello	Rebeca Schroeder	2	2

Tabela 4. Autores da ERBDxSBBD.

Autor	**#Artigos SBBD**	**#Artigos ERBD (1º autor)**
Ana Carolina Salgado	7	0
Carmem Hara	7	0
Karin Becker	6	0
Carlos Heuser	5	0
Ronaldo Mello	5	0
Carlos Eduardo Pires	3	1
Carina Dorneles	3	0
André Santanchè	2	0
Daniel S. Kaster	2	0
José Palazzo M. Oliveira	2	0
Leandro Wives	2	0
Renata Galante	2	0
Viviane Moreira Orengo	2	0
Alexandre Lazzaretti	1	2
Gustavo Kantorski	1	2
Augusto F. de Souza	1	1
Juliana B. dos Santos	1	1
Rebeca Schroeder	1	1
Roberto Walter	1	1
Rodrigo Machado	1	1
Aldo Wangenheim	1	0
Álvaro Freitas Moreira	1	0
Denio Duarte	1	0
Eduardo Borges	1	0
Nádia Kozievitch	1	0
Nina Edelweiss	1	0

se a formação de dois grandes componentes conectados, e que contêm a maior parte dos autores prolíficos da ERBD. No componente destacado por um retângulo, encontram-se os autores *Carmem Hara*, *Rebeca Schroeder*, *Ronaldo Mello*, *Angelo Frozza*, *Denio Duarte* e *Geomar Schreiner*. Já no maior componente, encontram-se *Carlos Heuser*, *Sérgio Mergen*, *Carina Dorneles*, *Renata Galante*, *Deise Saccol*, *Eduardo Borges*, *Cristiano Cervi*, *Edimar Manica* e *Gláucio R. Vivian*. Em ambos os grupos, identifica-se a relação orientador-aluno ocorrida em algum momento entre diversos destes autores. Claramente, este tipo de relação é responsável por formar os maiores componentes da rede.

Considerando o papel de formação na área de BD, procurou-se identificar autores da ERBD com publicações no fórum principal da área no Brasil, o Simpósio Brasileiro de Bancos de Dados (SBBD). Neste levantamento, foram considerados apenas artigos completos do SBBD publicados entre 2005-2019 (período de existência da ERBD). A Tabela 4 apresenta a relação completa com todos os autores, quantidade de artigos completos no SBBD e a quantidade de artigos da ERBD. Nesta última coluna foram contabilizados apenas os artigos em que o autor é o principal, indicado pela primeira posição na lista de autores dos artigos. Este levantamento permite identificar professores que contribuem em ambos os fóruns da área, bem como os possíveis alunos que um dia como autores principais de artigos da ERBD tiveram seus trabalhos publicados no fórum principal de BD no Brasil. Em trabalhos futuros serão consideradas outras categorias de artigos do SBBD para uma visão completa da intersecção das comunidades.

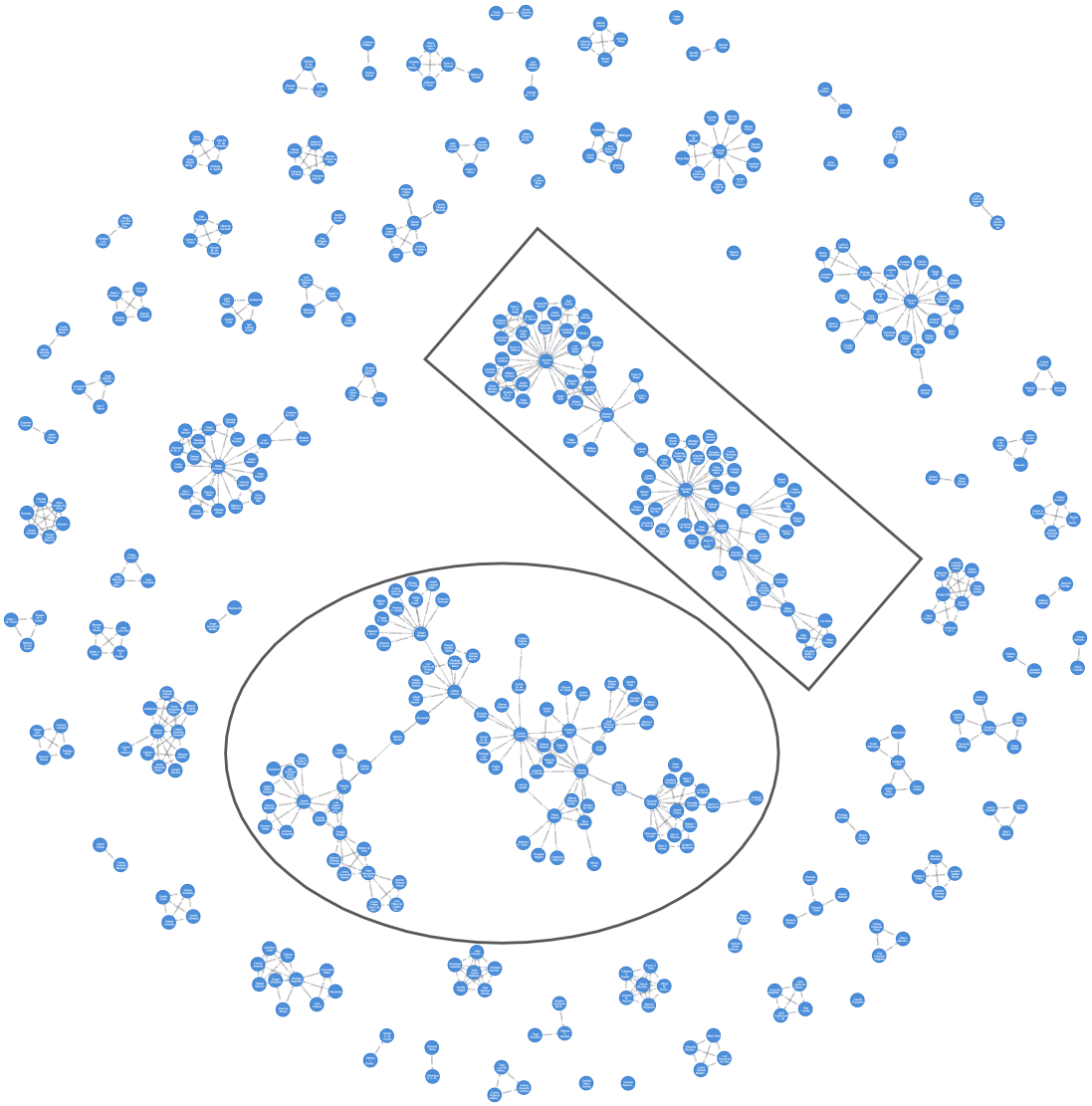

Figura 9. Rede de coautoria da ERBD.

5. Considerações Finais

Este artigo apresentou um levantamento bibliométrico produzido a partir dos artigos publicados nos primeiros 15 anos da Escola Regional de Banco de Dados. Através dos dados coletados foi possível quantificar a produção em termos de artigos, autores e suas colaborações. Além destes, a intersecção das comunidades dos dois principais eventos de banco de dados do Brasil foi identificada. Apesar de ser um trabalho inicial, acredita-se que os dados apresentados por este artigo possam contribuir para o auto-conhecimento da comunidade da ERBD, e apoiar decisões para o futuro deste evento.

Sobretudo, este artigo contribuiu com a coleta e produção de um banco de dados capaz de registrar os principais metadados das publicações da ERBD até este momento. Infelizmente, como mencionado no artigo, informações de algumas edições não puderam ser recuperadas completamente. Neste sentido, pretende-se como um próximo passo tentar a recuperação dos anais faltantes com os respectivos organizadores. Planeja-se também estender o banco de dados para que o mesmo considere as palavras-chave, resu-

mos e referências dos artigos, bem como a afiliação dos autores.

Existem diversas análises consideradas como trabalhos futuros. Encontra-se em andamento uma análise mais abrangente das redes de coautoria da ERBD, bem como sua evolução temporal. Além disto, considera-se continuar a analisar a intersecção da comunidade com o SBBD assim como com outras comunidades.

Agradecimentos. Agradecemos ao aluno Otávio Augusto de Almeida por parte da coleta e extração dos dados realizada neste trabalho.

Referências

[Amblard et al. 2011] Amblard, F., Casteigts, A., Flocchini, P., Quattrociocchi, W., and Santoro, N. (2011). On the temporal analysis of scientific network evolution. In *2011 International Conference on Computational Aspects of Social Networks (CASoN)*, pages 169–174.

[Batini et al. 1992] Batini, C., Ceri, S., and Navathe, S. B. (1992). *Conceptual Database Design: an Entity-relationship approach*. Benjamin-Cummings Publishing Co.

[Brandão et al. 2017] Brandão, M. A., de Melo, P. O. S. V., and Moro, M. M. (2017). Tie strength dynamics over temporal co-authorship social networks. In *Proceedings of the International Conference on Web Intelligence*, WI '17, page 306–313, New York, NY, USA. Association for Computing Machinery.

[Júnior et al. 2011] Júnior, P. S. P., Laender, A. H. F., and Moro, M. M. (2011). Análise da rede de coautoria do simpósio brasileiro de bancos de dados. In de Oliveira, J. P. M., editor, *XXVI Simpósio Brasileiro de Banco de Dados - Short Papers, Florianópolis, Santa Catarina, Brasil, October 3-6, 2011*, pages 131–138. SBC.

[Kazi et al. 2017] Kazi, S., Rajput, Q., and Khoja, S. (2017). Study of evolving co-authorship network: Identification of growth patterns of collaboration using sna measures. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 488–493.

[Neo4J 2010] Neo4J (2010). Neo4j graph database platform.

[PostgreSQL 1996] PostgreSQL (1996). Postgresql: The world's most advanced open source relational database.

[SBC 2020] SBC (2020). Sociedade brasileira de computação: Comissão especial de bancos de dados. Disponível em: http://comissoes.sbc.org.br/cebd/erbd.html.