Deduplicação de Nomes e Redes de Co-autoria na DBLP

Mariana O. Silva, Michele A. Brandão

Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

{mariana.santos,micheleabrandao}@dcc.ufmg.br

Abstract. *This article describes a dataset collected from the DBLP digital library, a repository with bibliographic data of Computer Science. This dataset includes approximately 15 million records collected in September 2016. From this dataset, two datasets were created. The first one has the original data collected from the DBLP with name deduplication treatment. The second one presents three co-authorship social networks built using the snowball sampling technique.*

Resumo. *Este artigo descreve um dataset coletado da biblioteca digital DBLP, um repositório com dados bibliográficos de Ciência da Computação. Este conjunto de dados inclui aproximadamente 15 milhões de registros coletados em Setembro de 2016. A partir deste dataset, foram criados dois conjuntos de dados. O primeiro possui os dados originais coletados da DBLP com tratamento de deduplicação de nomes. O segundo apresenta três redes sociais de coautoria construídas utilizando a técnica snowball sampling.*

1. Introdução

Bibliotecas Digitais (BDs) são repositórios organizados de uma ou mais coleções online, que provêm acesso à informações e conhecimento para seus usuários. Tais repositórios proporcionam diversos serviços, como pesquisa, visualização dos dados e armazenamento de metadados que descrevem seu conteúdo e suas interações. No contexto acadêmico, as bibliotecas digitais são consideradas importantes fontes de informação, fornecendo uma interface centralizada para o acesso a diversas publicações científicas.

Muitos estudos relevantes são realizados a partir dos dados coletados desses repositórios [Brandão and Moro 2017]. Por exemplo, para avaliar a qualidade e o impacto das publicações [Omodei et al. 2017; Weitzel 2006], identificar relevantes temas de pesquisa [Ohira and Prado 2002; Villarreal and Schaeffer 2016], revelar tendências e padrões de colaboração em redes sociais de coautoria [Brandão and Moro 2012; Brandão et al. 2013; Chen et al. 2017], dentre outros. Em particular, estudos focados na análise de interações entre pessoas ou organizações, bem como detectar padrões presentes nessas interações permitem prever o comportamento de uma rede e analisar diferentes aspectos da mesma.

Esses estudos podem ser usados por agências de fomento ou instituições de pesquisa e para tal pressupõem-se que os dados sejam de alta qualidade [Laender et al. 2008; Lee et al. 2007]. Porém, manter essa alta consistência geralmente não é uma tarefa simples. Um dos principais e mais complexos desafios enfrentados para melhorar a qualidade dos dados é a deduplicação de nomes [Laender et al. 2008]. Esse problema pode ocorrer de duas formas: quando um mesmo autor publica utilizando nomes similares, mas distintos (sinônimos); ou quando autores diferentes compartilham o mesmo nome, ou variações parecidas (homônimos).

Nesse contexto, a DBLP[1] (*Digital Bibliography & Library Project*) é um exemplo de biblioteca digital que possui informações bibliográficas sobre as principais publicações de Ciência da Computação. Essa biblioteca armazena dados de pesquisadores da área da computação (ou ciências vizinhas) de todo o mundo. Em setembro de 2016, essa coleção possuía cerca de 1,780,000 autores e 3,400,000 publicações. Dessa forma, a DBLP proporciona dados bibliográficos reais e úteis que podem auxiliar na análise de redes sociais acadêmicas. No entanto, a presença de sinônimos e homônimos é o problema principal. Por sua amplitude e representatividade, escolhemos a DBLP como fonte de coleta para construir o conjuntos de dados apresentado neste artigo.

Após descrever os trabalhos relacionados e aplicações do conjunto de dados (Seção 2), as principais contribuições deste artigo são: uma metodologia para construir dois conjuntos a partir dos dados da DBLP (Seção 3); uma descrição detalhada e quantitativa dos dois conjuntos além da disponibilização online dos mesmos (Seções 4 e 5); Finalmente, apresentamos as conclusões e trabalhos futuros (Seção 6).

2. Trabalhos Relacionados e Aplicações

As bibliotecas digitais são sistemas de informação extremamente complexos que envolvem conjuntos de objetos digitais e seus respectivos metadados [Gonçalves et al. 2004]. Esses dados podem ser provenientes de fontes variadas, mas relativos a uma mesma área de interesse e possuem o propósito de atender a uma determinada comunidade [Borgman 1999]. O vasto conteúdo presente nessas bibliotecas digitais podem conduzir a análises de dados interessantes, tais como tendências de pesquisa [Ferreira 2012], padrões de colaboração em redes sociais [Freitas et al. 2008], predição de links [Hasan et al. 2006], recomendação de colaborações [Brandão et al. 2013], pesquisas em qualidade da informação [Han et al. 2004], entre outras.

Dentre os diversos campos de pesquisa, a análise de redes sociais tem se tornado um assunto extremamente abordado e relevante. Trabalhos com as mais diversas finalidades têm sido realizados para analisar diferentes aspectos de uma rede social. Tais redes podem mostrar padrões de cooperação entre pesquisadores e o impacto de suas publicações [Börner et al. 2005], podem ser utilizadas para recomendação de colaboração [Brandão and Moro 2012] e avaliar grupos de pesquisa e programas de pós-graduação [Lopes et al. 2011]. Outro exemplo de aplicação de dados derivados de BDs é a análise da qualidade de agrupamento [Brandão and Moro 2017]. Ademais, utilizamos as redes sociais de co-autoria do conjunto de dados apresentado neste artigo para avaliar se métricas para força de relacionamentos podem ser usadas também para avaliar a qualidade de comunidades[2].

Geralmente, redes sociais acadêmicas apresentam uma estrutura grande e volumosa de dados, o que impossibilita uma análise manual detalhada. Para analisar essas redes, é necessário desenvolver métodos que possam tratar este grande volume de dados. Dessa forma, usuários e desenvolvedores enfrentam diversos desafios ao realizarem pesquisas mais precisas e detalhadas de tais redes. Diante desses desafios, realizamos um estudo que especifica e valida modelos para as redes sociais acadêmicas, incluindo a definição de uma infraestrutura de banco de dados que permite armazenar e manter os dados das redes[3]. Tal

[1]DBLP: `http://dblp.uni-trier.de/`

[2]Relatório técnico em `http://www.dcc.ufmg.br/~mirella/projs/apoena`

[3]Relatório técnico em `http://www.dcc.ufmg.br/~mirella/projs/apoena`

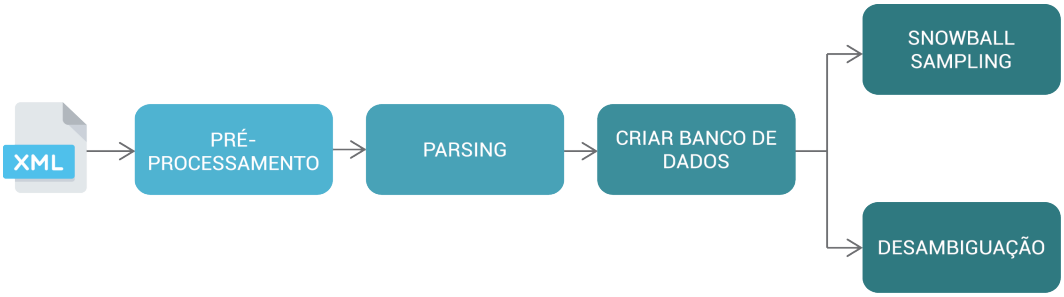


Figura 1: Etapas da metodologia.

infraestrutura também facilita a utilização de técnicas de análise de redes sociais. Esse estudo compara o desempenho de diferentes tipos de SGBDs (Sistemas de Gerenciamento de Bancos de Dados) utilizando o conjunto de dados descrito na Seção 4.

A DBLP é uma abrangente fonte de pesquisas científicas em Ciência da Computação, capaz de facilitar a análise de dados em diversas áreas. Por exemplo, Yang and Leskovec [2015] estudam um conjunto de 230 redes sociais grandes, onde uma rede científica de colaboração foi criada a partir dos dados coletados da DBLP. Os autores apresentam uma metodologia que compara e avalia quantitativamente as diferentes definições estruturais de comunidades. Verificou-se que o método proposto apresenta uma melhoria relativa de 30% em relação a métodos de agrupamento local do estado da arte. Por outro lado, Lange and Naumann [2011] utilizam o conjunto de dados da DBLP para avaliar uma abordagem proposta por eles para medir a semelhança de dois registros.

Finalmente, além de pesquisas voltadas à análise de redes sociais de co-autoria e problemas de deduplicação e desambiguação de dados, novos desafios de pesquisa também podem se beneficiar a partir dos dados coletados e tratados neste trabalho. Alguns exemplos importantes são análises de agrupamento de comunidades, formação de times na pesquisa científica e recomendação de colaborações.

3. Metodologia

A DBLP começou como uma pequena e limitada base de dados, mas tornou-se uma grande biblioteca digital contendo trabalhos de quase todos os campos de estudo em computação. Todo o conjunto de dados da DBLP está disponível online como um grande arquivo XML. O arquivo *dblp.xml* (um simples arquivo XML ASCII) contém todos os registros bibliográficos presentes na biblioteca e é acompanhado pelo arquivo *dblp.dtd*, um conjunto de regras que define quais tipos de dados e entidades fazem parte do documento XML.

A Figura 1 apresenta as etapas da metodologia, desde a obtenção do arquivo XML até a criação dos dois conjuntos de dados. Antes de realizar a leitura do *dblp.xml*, realizamos um pré-processamento para remover a codificação utilizada para caracteres especiais. Com o arquivo tratado, executamos o *parser* do mesmo. Em seguida, foram gerados dois conjuntos de dados a partir dos dados coletados: uma base de dados com nomes desambiguados e outra com três redes sociais de co-autoria.

A identificação de registros duplicados é um processo complexo e composto de várias etapas. A Figura 2 apresenta uma visão geral do processo de deduplicacão de acordo com a abordagem *Data Deduplication* [Christen 2012]. Assim, para eliminarmos

Figura 2: Processo de deduplicação dos dados.

os registros duplicados, seguimos as seguintes etapas: *(1)* pré-processamento responsável por dividir os nomes dos autores em primeiro nome, nome do meio e último nome; *(2)* indexação de todos os registros por uma chave de bloco (BK), utilizando a técnica *Soundex* [Odell and Russell 1918]; *(3)* comparação de todos os registros pertencentes a cada um dos blocos por meio da função de similaridade *Jaro Winkler* [Winkler 1990]; *(4)* classificação dos registros como duplicados, não duplicados e como possíveis duplicados de acordo com um limiar de similaridade.

Para a criação das redes, utilizamos o dataset original, sem o tratamento de nomes ambíguos. Além disso, foi utilizada a técnica de amostragem não probabilística conhecida como *snowball sampling* para filtrar e diminuir seu volume [Goodman 1961]. Essa técnica foi escolhida por ser mais direcionada do que outras técnicas de amostragem não aleatórias [Pearson 2012]. Para adquirir uma amostra, é necessária uma "semente", geralmente de indivíduos conhecidos envolvidos no comportamento sob análise [Snijders et al.]. Aqui, os nós sementes escolhidos foram os bolsistas vigentes do CNPq[4] de Ciência da Computação. Note que para resolver o problema de nomes não padronizados na DBLP, buscamos manualmente o perfil dos bolsistas. Assim, garantimos a corretude das informações de cada um deles na rede. A partir dos nós sementes, foram feitas duas coletas para aumentar a amostra e criar a três redes reais de tamanhos diferentes. Ao final, são três redes criadas a partir da DBLP: (*Rede 0*) formada apenas pelos bolsistas vigentes do CNPq, que fazem parte da DBLP; (*Rede 1*) formada pela *Rede 0* e seus vizinhos; e (*Rede 2*) formada pela *Rede 1* e seus vizinhos.

4. Descrição dos Dados Coletados

De acordo com Seção 3, os dados foram coletados do arquivo *dblp.xml* disponível na DBLP. Este arquivo é modelado a partir do formato BibTeX *.bib. Existem dois tipos de registros neste arquivo: registros de publicação e registros de pessoas. Os registros de publicação são fornecidos por um dos seis elementos: *article*, artigo em periódico ou revista; *inproceedings*, artigo publicado em conferência ou workshop; *proceedings*, volume de trabalhos de uma conferência ou workshop; *book*, autoria de monografia ou coleção editada de artigos; *incolletion*, parte ou capítulo em uma monografia; *phdthesis*, tese de doutorado; *masterthesis*, tese de mestrado; *www*, página da web.

Para formar o conjunto de dados, consideramos apenas os dados referentes às pessoas (autores e editores) e suas publicações. Para simplificar e filtrar o dataset, foram

[4]`http://cnpq.br/bolsistas-vigentes/` (Abril de 2017)

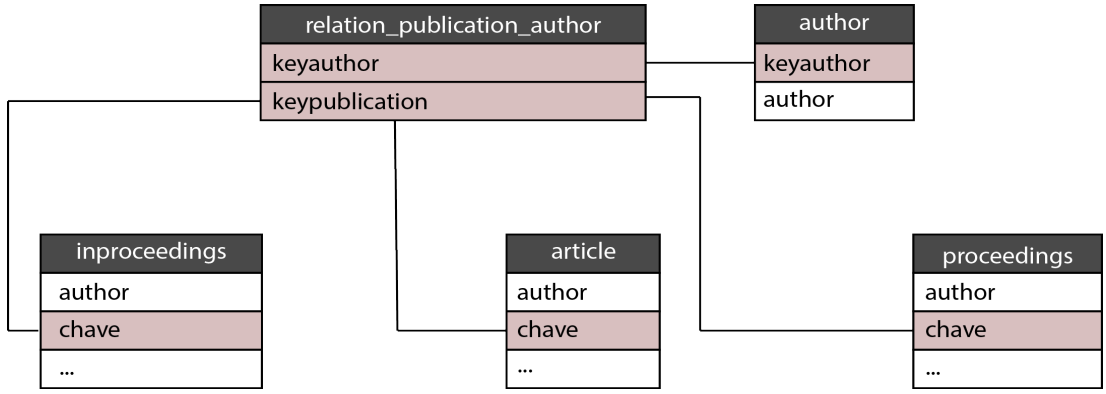


Figura 3: Esquema do banco de dados relacional. Note que exibimos apenas os principais atributos das tabelas devido a restrições de espaço.

Tabela 1: Descrição das redes sociais criadas.

Rede	# autores	# pares (# dist)	MedPubA	Modularidade	Coeficiente de clusterização médio
0	394	3898 (738)	9,89	0,686	0,377
1	68397	249352 (110357)	3,65	0,86	0,672
2	314444	1297929 (540571)	4,13	0,691	0,524

coletadas publicações originadas apenas dos elementos: *article*, *inproceedings* e *proceedings*. Além disso, criamos uma tabela que relaciona autores (ou editores) com suas publicações. O conjunto de dados coletado foi armazenado em um sistema de gerenciamento de banco de dados relacional (SGBDR), mais especificamente o MySQL. O esquema de dados relacional possui 5 tabelas conforme apresentado na Figura 3. Na tabela *author* estão armazenados o nome e o id dos pesquisadores. As tabelas *article*, *inproceedings* e *proceedings* possuem informações detalhadas de publicações. Finalmente, a tabela *relation_publication_author* representa o relacionamento entre autores e publicações, que podem estar em *article*, *proceedings* ou *inproceedings*.

Em seguida, foi realizado um processo de deduplicação de nomes e construção das três redes sociais. A Tabela 1 descreve as principais propriedades de cada rede social criada, que são o número de pesquisadores (autores de artigos), número de publicações, número médio de publicações por autor, número de pares de co-autores (e número de pares distintos de co-autores), modularidade da rede e o coeficiente de clusterização médio. A modularidade é uma medida capaz de determinar a qualidade da divisão feita na rede, enquanto o coeficiente de clusterização médio representa o valor médio do grau com que os nós de uma rede (ou grafo) tendem a se agrupar.

Para ilustrar a estrutura das redes sociais de co-autoria construídas a partir dos dados coletados da DBLP, a Figura 4 exemplifica a rede origem de coautoria com apenas pesquisadores vigentes do CNPq (*Rede 0*). Os nós do grafo representam os autores e cada aresta representa a colaboração entre dois autores. O tamanho do nó é proporcional ao seu grau, ou seja, ao número de arestas adjacentes a ele. A espessura da aresta é proporcional ao número de publicações que dois autores possuem. Já a cor dos nós identifica o componente conectado a que pertencem. Observa-se que o maior componente conectado da rede é

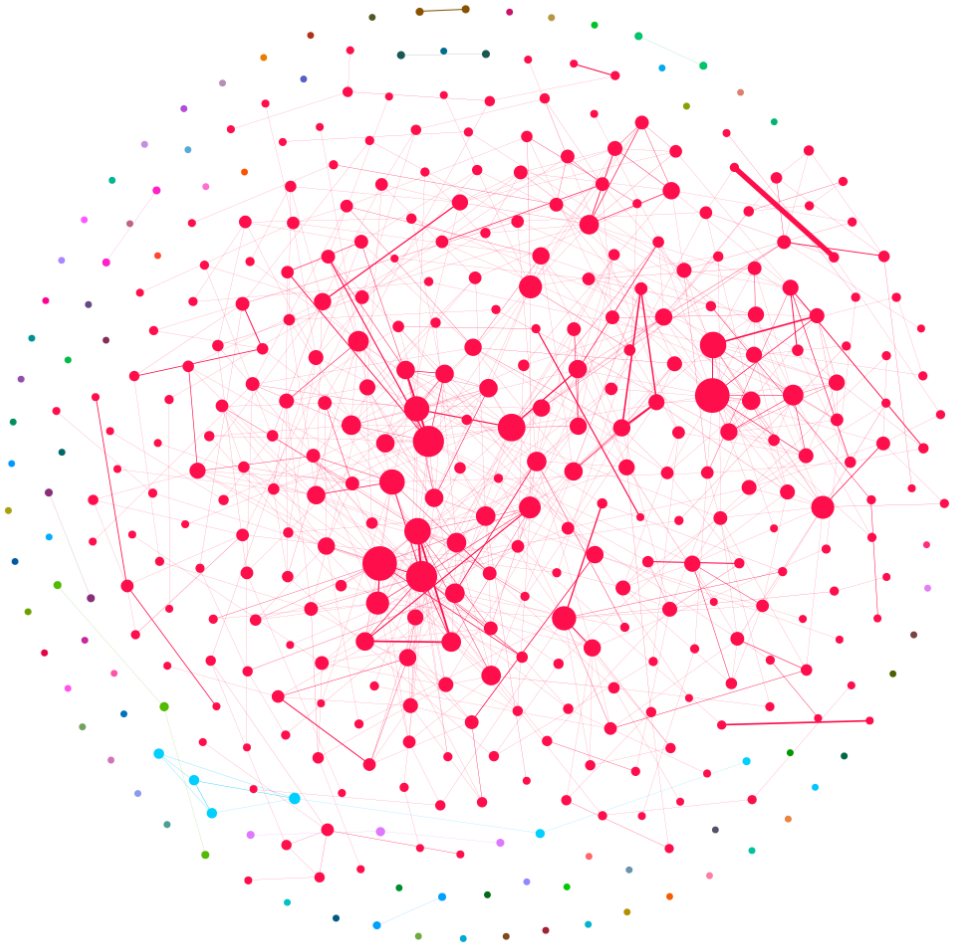


Figura 4: Representação gráfica da *Rede 0*.

grande (em vermelho), abrangendo cerca de 76% da rede, o que significa tratar-se de uma comunidade bem conectada.

O conjunto de dados completo está disponível na página do projeto Apoena[5] e é composto por três conjuntos de dados compactados:

- **Dblp.zip** - contém os dados coletados sem a deduplicação de nomes dos pesquisadores;
- **Dblp_name_desambiguation.zip** - contém os dados coletados com nomes ambíguos resolvidos;
- **Dblp_social_networks.zip** - contém os dados das três redes de coautoria.

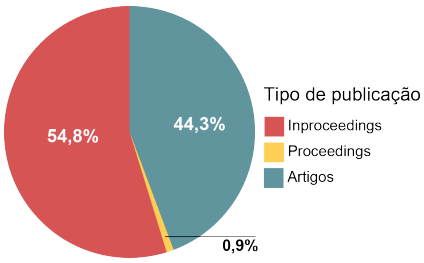
5. DBLP em Números

O conjunto de dados coletado por meio da metodologia descrita na Seção 3 consiste em aproximadamente 15 milhões de registros. A Tabela 2 apresenta uma descrição quantitativa do conjunto de dados. A Figura 5(a) mostra a distribuição dos tipos de publicações

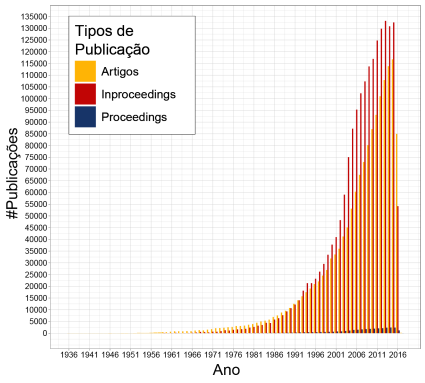
[5]`http://homepages.dcc.ufmg.br/~mirella/projs/apoena/datasets.html`

Tabela 2: Conjunto de dados coletado (16 de Setembro de 2016).

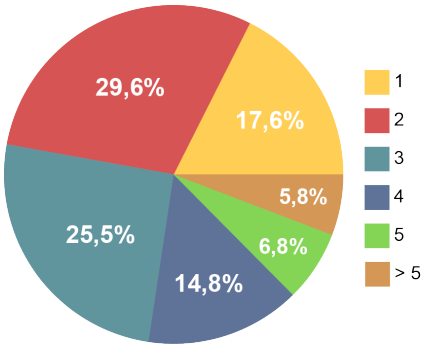
Dados	**Número de registros**
Publicações em artigos	1.505.020
Autores	1.779.971
Publicações em proceedings	31.549
Publicações em inproceedings	1.861.226
Relação entre autores e publicações	9.707.161
Total	**14.884.927**



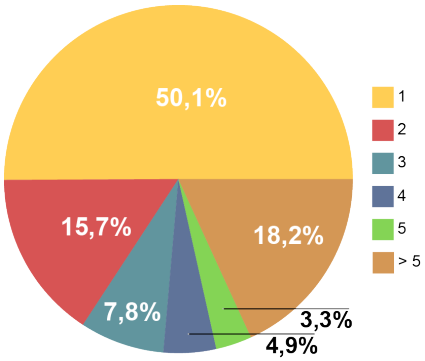
(a) Distribuição dos tipos de publicação



(b) Publicações por ano (1936−2017)



(c) Número de coautores por publicação



(d) Número de publicações por autor

Figura 5: Estatísticas sobre o conjunto de dados.

presentes no conjunto de dados. Observa-se que aproximadamente 55% das publicações são provenientes de conferências ou workshops (*inproceedings*), cerca de 44% são artigos de um periódico ou revista e apenas 0,9% representam *proceedings*. Ou seja, a maioria das publicações presentes no conjunto de dados são de artigos ou *inproceedings*.

A Figura 5(b) apresenta a evolução do número de publicações ao longo dos 81 anos da DBLP (1936 - 2017). Para este diagrama, as publicações foram agrupadas por seu tipo e ano de publicação. É possível notar que o número de publicações originadas de artigos e *inproceedings* estão correlacionados. Porém, o número de publicações em *inproceedings* cresce mais rapidamente que os outros tipos de publicação. Além disso, observa-se que a partir dos anos 90 houve um grande crescimento do número de publicações. Provavelmente,

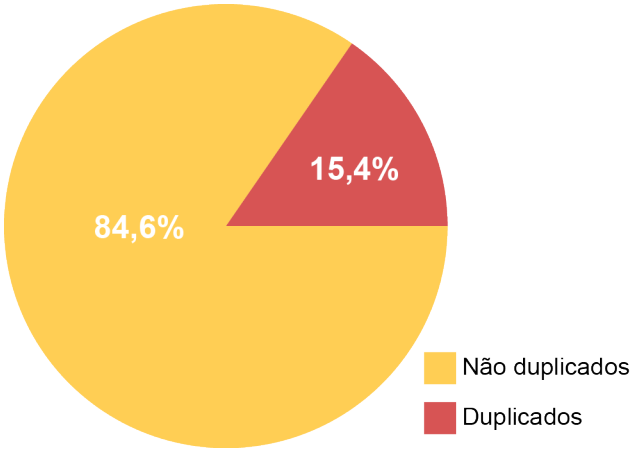


Figura 6: Distribuição dos registros duplicados no conjunto de dados.

isso pode ser explicado pelo uso da Internet que se difundiu nessa época.

A Figura 5(c) mostra a distribuição do número de co-autores por publicação. Percebe-se que cerca de 30% das publicações (seja de *articles* ou *inproceedings*) têm dois co-autores e 25,5% têm três co-autores. Além disso, apenas 5,8% das publicações apresentam mais de 5 co-autores. O maior número de pesquisadores em uma publicação é 287. Uma análise semelhante pode ser feita verificando-se a média de publicações por autor. A Figura 5(d) apresenta essa análise. Observa-se que aproximadamente metade da comunidade de autores publicam apenas uma única vez na DBLP, cerca de 16% publicam duas vezes e mais de 18% publicam mais de cinco vezes.

Segundo Lee et al. 2007, os desafios e limitações relacionados à qualidade dos dados presentes nas bibliotecas digitais têm várias origens. Por exemplo, erros na entrada de dados, ausência de padrões de execução, imperfeição de softwares de coleta, geração de metadados em larga escala, ambiguidade de nomes de autores, entre outros. Dentre esses, o problema de ambiguidade de nomes de autores vem se destacando na comunidade científica, devido à sua inerente dificuldade. Conforme Seção 3, para a criação de um dos conjuntos de dados, aplicamos um processo de deduplicação de acordo com a abordagem *Data Deduplication*. Para isso, foi utilizada uma função de deduplicação com limiar de 95% [Christen 2012], fazendo com que fosse possível alcançar resultados satisfatórios. Após o processo, foram encontrados 289.598 registros duplicados, reduzindo o número de autores de 1.779.971 para 1.593.237. Na Figura 6, podemos ver a distribuição dos registros duplicados detectados no conjunto de dados, abrangendo mais de 15% dos dados.

6. Conclusões

Neste artigo, apresentamos dois conjuntos de dados a partir dos dados da DBLP. Um conjunto possui os nomes dos autores deduplicados, enquanto que o outro é composto por três redes sociais de co-autoria. Ademais, apresentamos as principais aplicações, desafios e limitações desses conjuntos. Como trabalhos futuros, planejamos construir mais redes sociais de co-autoria considerando diferentes perfis de pesquisadores na DBLP. Além disso, pretendemos aprimorar a qualidade dos dados coletados, tanto explorando novas técnicas de deduplicação de nomes propostas na literatura, quanto incluindo metadados aos conjuntos de dados para melhorar a sua compreensão e facilitar o seu uso.

Agradecimentos. Trabalho parcialmente financiado por CAPES, CNPq e FAPEMIG.

Referências

Borgman, C. L. (1999). What are digital libraries? competing visions. *Inf. Process. Manage.*, 35(3):227–243.

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

Brandão, M. A. and Moro, M. M. (2017). Social professional networks. *Computer Communications*, 100(C):20–31.

Brandão, M. A. and Moro, M. M. (2012). Recomendação de colaboração em redes sociais acadêmicas baseada na afiliação dos pesquisadores. In *Procs. of SBBD (Short Papers) - Simpósio Brasileiro de Bancos de Dados*, pages 73–80.

Brandão, M. A. and Moro, M. M. (2017). Strength of Co-authorship Ties in Clusters: a Comparative Analysis . In *Procs. of AMW - Alberto Mendelzon International Workshop on Foundations of Data Management*, Montevideo, Uruguai.

Brandão, M. A., Moro, M. M., and Almeida, J. M. (2013). Análise de fatores impactantes na recomendação de colaborações acadêmicas utilizando projeto fatorial. In *Procs. of SBBD (Short Papers) - Simpósio Brasileiro de Bancos de Dados*, pages 5:1–5:6.

Chen, Y., Ding, C., Hu, J., Chen, R., Hui, P., and Fu, X. (2017). Building and analyzing a global co-authorship network using google scholar data. In *Procs. of WWW - International Conference on World Wide Web Companion*, pages 1219–1224, Perth, Austrália.

Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE transactions on knowledge and data engineering*, 24(9):1537–1555.

Ferreira, A. A. (2012). *Contributions for solving the author name ambiguity problem in bibliographic citations*. PhD thesis, Universidade Federal de Minas Gerais.

Freitas, C., Nedel, L. P., Galante, R., Lamb, L. C., Spritzer, A. S., Fujii, S., de Oliveira, J. P. M., Araújo, R. M., and Moro, M. M. (2008). Extração de conhecimento e análise visual de redes sociais. *SEMISH - Seminário Integrado de Software e Hardware, Belém do Pará, Brasil, SBC*, pages 106–120.

Gonçalves, M. A., Fox, E. A., Watson, L. T., and Kipp, N. A. (2004). Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. *ACM transactions on information systems (TOIS)*, 22(2):270–312.

Goodman, L. A. (1961). Snowball sampling. *The annals of mathematical statistics*, pages 148–170.

Han, H., Giles, L., Zha, H., Li, C., and Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Procs. of JCDL - Joint ACM/IEEE conference on Digital Libraries*, pages 296–305, Tucson, USA.

Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M. (2006). Link prediction using supervised learning. In *Procs. of SDM - Workshop on Link Analysis, Counterterrorism and Security*.