October 7-10 • Ceará • Brazil

34th Brazilian Symposium on DATABASES

SBBD|2019

TÓPICOS EM GERENCIAMENTO DE DADOS E INFORMAÇÕES

October 7-10 • Ceará • Brazil

34th Brazilian Symposium on DATABASES

TÓPICOS EM GERENCIAMENTO DE DADOS E INFORMAÇÕES

Promotion
Sociedade Brasileira de Computação – SBC
Comissão Especial de Banco de Dados (CEBD) da SBC

Organization
Departamento de Computação, Universidade Federal de Ceará– UFC

Steering Committee Chair
Bernadete Farias Lóscio (UFPE, Brazil)

Local Organization Chairs
José Maria da Silva Monteiro Filho (UFC, Brazil)

Program Committee Chairs
Full Paper: Carina F. Dorneles (UFSC, Brazil)
Short Paper: Fábio Porto (LNCC, Brazil)
Demos and Applications Chair: Robson L. F. Cordeiro (ICMC-USP, Brazil)
Thesis and Dissertation Workshop Chair: Jonice Oliveira (UFRJ, Brazil)
Tutorials Chair: Altigran Soares da Silva (UFAM, Brazil)
Short course Chair: Maria Cláudia Cavalcanti (IME, Brazil)
Workshop Chair: José Antônio Macedo (UFC, Brazil)
Thesis and Dissertation Contest Chair: Caetano Traina Jr. (USP, Brazil)
Graduation Student Workshop Chair: Ticiana Linhares (UFC, Brazil)
Graduation Student Workshop Chair: Ticiana Linhares (UFC, Brazil)

Editorial

Os minicursos apresentados no XXXIV Simpósio Brasileiro de Banco de Dados (SBBD 2019) têm por objetivo apresentar temas relevantes da área de Banco de Dados e promover discussões sobre os fundamentos, tendências e desafios relacionados ao tema abordado, sendo uma excelente oportunidade de atualização para acadêmicos e profissionais que participam do evento.

Nesta edição, foram selecionadas quatro das oito propostas recebidas, para serem apresentadas durante o SBBD 2019. A seleção das propostas foi realizada por um Comitê de Avaliação formado por quatro avaliadores. Durante o processo de seleção, as propostas submetidas foram avaliadas por todos os membros do comitê. Ao final do processo, os proponentes dos minicursos selecionados prepararam os textos que constituem os capítulos deste livro.

O primeiro minicurso, "Técnicas de Privacidade de Dados de Localização", tem por objetivo apresentar os principais conceitos relacionados ao problema da violação de privacidade de dados de localização dos indivíduos, os riscos inerentes e apontar de forma detalhada as principais técnicas existentes na literatura para a preservação de privacidade em serviços de localização. No segundo minicurso, "Uma Introdução ao Combate Automático às Fake News em Redes Sociais Virtuais", é apresentada uma introdução conceitual e prática às principais abordagens computacionais de combate às Fake News, além de comentar sobre áreas e pesquisas recentes relacionadas a este tema. Já o terceiro minicurso, "Ecossistemas de Dados naWeb: da teoria aos desafios", discute os principais conceitos relacionados a este novo ambiente, abordando aspectos relevantes, tanto do ponto de vista teórico quanto de desafios e oportunidades de pesquisa nesta área. Por fim, o quarto minicurso, "Aprendizado de máquina e inferência em Grafos de Conhecimento", apresenta uma introdução aos métodos e técnicas de aprendizado de máquina utilizadas em tarefas de inferência em grafos de conhecimento, discutindo-se os desafios e oportunidades tecnológicas e científicas desse tipo de tarefa.

Agradecemos aos autores pela submissão das propostas e geração dos textos finais, bem como ao Comitê de Avaliação, pela dedicação e eficiência em todo o processo de seleção dos minicursos.

Maria Cláudia Reis Cavalcanti (IME)
Coordenadora de Minicursos do SBBD 2019

34th Brazilian Symposium on Databases

October 7-10, 2019
Fortaleza - CE - Brazil

TÓPICOS EM GERENCIAMENTO DE DADOS E INFORMAÇÕES

Promoção
Sociedade Brasileira de Computação – SBC
Comiss˜ao Especial de Banco de Dados (CEBD) da SBC

Organização
Departamento de Computação, Universidade Federal de Ceará– UFC

Coordenadores do SBBD

Coordenador do Comitê Diretivo
Bernadette Lóscio (UFPE)

Coordenadores de Organização Local
José Maria da Silva Monteiro Filho (UFC)

Coordenadora do Comitê de Programa
Carina Dorneles (UFSC, Brazil)

Coordenadoras do Comitê de Programa de Artigos Curtos
Fábio Porto (LNCC, Brazil)

Coordenador da Sessão de Demos e Aplicações
Robson L. F. Cordeiro (ICMC-USP, Brazil)

Coordenadora do Workshop de Teses e Dissertações em Banco de Dados
Jonice Oliveira (UFRJ, Brazil)

Coordenadora de Minicursos
Maria Cláudia Cavalcanti (IME, Brazil)

Coordenador de Tutoriais
Altigran Soares da Silva (UFAM, Brazil)

Coordenador do Concurso de Teses e Dissertações
Caetano Traina Jr. (USP, Brazil)

Coordenador de Workshops
José Antônio Macedo (UFC, Brazil)

Comitê de Avaliação de Minicursos
Maria Cláudia Cavalcanti (IME) (coordenadora)
Kelli de Faria Cordeiro (MB/IME) (vice-coordenadora)
Eugênio da Silva (UEZO)
Maria Camila Nardini Barioni (UFU)
Vaninha Vieira (UFBA)

Table of Contents

Capítulo 1

Técnicas de Privacidade de Dados de Localização

Javam C. Machado, Eduardo R. Duarte Neto, and Manuel E. Bento Filho [1]

Abstract

The development of mobile devices has led to the growing popularity of location services, allowing unknown or untrusted servers to collect a large amount of information from the user's location data, which has raised serious concerns about the privacy of the user's sensitive information in the use of these services. Thus, maintaining user privacy while ensuring the quality of service is a complex problem that has received much attention in recent years. The objective of this chapter is to describe the problem of violating individuals' location privacy, presenting their main concepts, the inherent threats and pointing out in detail the main techniques in the literature for preserving privacy in location services. Finally, we will highlight research opportunities in the area and present relevant conclusions on the subject.

Resumo

O desenvolvimento dos dispositivos móveis tem proporcionado o crescimento da popularidade dos serviços de localização, permitindo que servidores desconhecidos ou não confiáveis coletem uma grande quantidade de informação dos usuários a partir de seus dados de localizações, o que tem gerado sérios questionamentos sobre a privacidade das informações sensíveis dos usuários no uso destes serviços. Sendo assim, manter a privacidade dos usuário, e simultaneamente, garantir a qualidade do serviço é um problema complexo que tem recebido bastante atenção nos últimos anos. Este capítulo tem por objetivo descrever o problema da violação de privacidade de dados de localização dos indivíduos, apresentando seus principais conceitos, os riscos inerentes e apontando de forma detalhada as principais técnicas existentes na literatura para a preservação de privacidade em serviços de localização. Por fim, iremos destacar oportunidades de pesquisas na área e apresentar conclusões relevantes sobre o tema.

[1]LSBD/DC – Universidade Federal do Ceará

1.1. Introdução

Com o passar dos anos a quantidade de dados coletados por aplicativos a fim de prover serviços tem crescido bastante. Estes dados são muito valiosos para os diversos tipos de organizações, sejam elas de saúde, varejo, dentre outras. Por exemplo, muitas empresas da área de varejo traçam estratégias de vendas de acordo com o perfil de seus consumidores, através da análise dos dados de seus consumidores, assim, potencializando o lucro da empresa. Já na área de saúde é possível identificar quais regiões estão mais sujeitas ou não a uma doença. Esse tipo de análise só é possível graças à análise de dados privados de indivíduos. Entretanto, isto leva a sérios riscos de exposição de dados sensíveis dos indivíduos. Logo, encontrar uma forma de permitir esta análise sem que haja riscos a exposição dos mesmos tem sido objeto de estudo na área de privacidade de dados.

O desenvolvimento dos dispositivos móveis tem contribuído bastante para a popularidade dos serviços baseado em localização (LBS), que utilizam de informações de localização para atender seus usuários. Através dos sensores destes dispositivos, as coordenadas de latitude e longitude são obtidas e utilizadas por estes serviços. Segundo Schiller [27], os serviços de localização são definidos como serviços que integram a localização ou posição de um dispositivo móvel a outras informações, de modo a fornecer valor agregado a um usuário. Estes numerosos serviços, tais como navegação, redes sociais, serviços de recomendação, jogos de realidade aumentada, entre outros, tem sido desenvolvidos e integrados às atividades diárias das pessoas, provendo informações úteis sobre seus arredores e sendo capazes de responder perguntas do dia a dia como: qual a melhor rota a ser percorrida para um determinado endereço? Quais os pontos turísticos mais próximos da minha localização atual? Em quanto tempo o táxi que eu solicitei irá demorar para chegar em meu apartamento?

O uso das informações geradas pelos serviços de localização pode beneficiar várias aplicações. De fato, muitas empresas e agências governamentais tem obtido conhecimento sobre os dados associados às atividades praticadas nas localizações, seja para melhorar o serviço prestado, para o lançamento de um novo produto, ou até mesmo para gerar uma nova política pela empresa. Entretanto, acessar dados de localizações de usuários desses serviços, mesmo que com permissão, levanta severas preocupações de privacidade para a maioria dos usuários. Dessa forma, a utilização de serviços baseados em localização pode levar a sérios riscos de violação de privacidade devido a provedores de serviços não confiáveis [20], que podem expor os dados de localização de seus usuários ou até mesmo vender suas informações de localizações a terceiros [35]. De posse dessas informações, os dados obtidos por terceiros são utilizados para descoberta de dados sensíveis dos usuários, *i.e.*, dados de saúde, crenças religiosas, ideologias políticas, questões raciais, preferências sexuais, dentre várias outras.

A Figura 1.1 ilustra um típico exemplo de violação de privacidade no uso de serviços de localização. Em questão, o usuário Bob ao longo do tempo realiza várias requisições a um serviço baseado em localização. No tempo t_1 Bob estava próximo ao pronto-socorro de um hospital, já em um tempo t_2 Bob realiza uma nova consulta próxima a um laboratório de patologia, em outros dois momentos sua localização também esta próximo à localizações associadas a área de saúde. Considerando que é de conhecimento do provedor do serviço as requisições feitas pelos usuários, o próprio provedor

facilmente consegue inferir, com alta probabilidade, que Bob possui algum tipo de doença em razão das localizações enviadas por ele ao provedor, revelando uma informação sensível do usuário. Desta forma, considerando que o provedor de serviço pode não ser confiável, o risco de uma violação de privacidade, portanto, é bastante alto, deixando o usuário exposto.

Figura 1.1. Exemplo de requisições realizadas próximas a hospitais e clínicas, permitindo inferências de dados sensíveis do usuário.

A aplicação de modelos de privacidade sobre requisições de usuários é imprescindível para evitar que as localizações dos indivíduos não sejam identificadas pelos provedores no uso destes serviços. Todavia, em geral os modelos de privacidade acabam provocando mudanças nos dados, afetando diretamente a sua utilidade, com impacto direto na qualidade do serviço. Portanto, gerenciar essa solução de compromisso (*trade-off*, Figura 1.2) entre privacidade dos indivíduos e utilidade dos seus dados se torna um outro grande desafio. Desta forma, vários modelos de privacidade de dados têm sido propostos por pesquisadores com o objetivo de resolver esta questão.

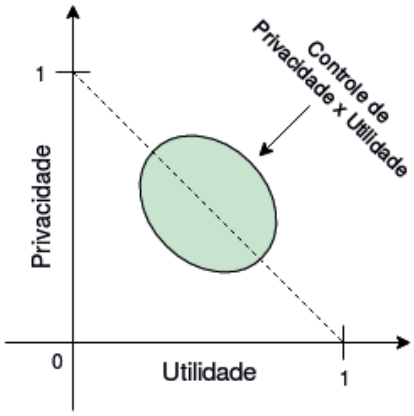

Figura 1.2. Trade-off entre privacidade e utilidade

Este capítulo tem por objetivo introduzir os fundamentos e técnicas para preservação da privacidade de dados dos indivíduos, procurando apresentar os riscos mais comuns

e as técnicas mais populares na solução do problema. Em seguida, apresentaremos um aprofundamento sobre o tema privacidade de dados de localização, apontando os conceitos básicos sobre dados de localização, os tipos de ataques a que estão sujeitos, e os principais modelos de preservação de privacidade de dados de localização na atualidade. A Seção 1.2 apresenta os princípios básicos sobre o tema, que tipos de dados estão sujeitos a violação, e como a preservação de privacidade pode ser alcançada. A Seção 1.3 apresenta os principais tipos de ataque utilizados para violação à privacidade dos indivíduos. Os modelos sintáticos mais populares para preservação de privacidade são descritos na Seção 1.4. A Seção 1.5 apresenta o modelo de privacidade diferencial. Na Seção 1.6 apresentamos o tema privacidade de localização, descrevendo o modelo de serviços de localização, bem como as características dos dados de localização. A Seção 1.7 descreve os principais tipos de ataque que dados de localização estão sujeitos. Os modelos de privacidade em dados de localização são descritos na Seção 1.8. E por fim, a Seção 1.9 apresenta as considerações finais do capítulo.

1.2. Privacidade de Dados

Privacidade é o direito que um indivíduo tem de manter seus assuntos pessoais e relacionamentos secretos [8]. É um consenso entre os pesquisadores que a privacidade é um assunto complexo, com muitas questões envolvidas. Sendo bastante comum confundir os conceitos de privacidade e segurança. Apesar de privacidade e segurança serem temas relacionados, suas definições tratam de pontos bem distintos. Em se tratando de dados, a segurança visa regular o acesso durante todo o ciclo de vida do dado, enquanto a privacidade define como será realizado esse acesso, em geral com base em leis e políticas de privacidade. Neste ponto, também surge o conceito de controle de acesso como forma de fornecer segurança a um conjunto de dados. O controle de acesso se refere a regras específicas de quem está autorizado a acessar (ou não) determinados recursos, isto é, quando um conjunto de usuários está apto a acessar um conjunto de dados. A privacidade aqui, está associada a regras de controle de acesso efetivas, que permitem a revelação da informação apenas por usuários autorizados. Contudo, a privacidade dos indivíduos não está garantida apenas com o controle de acesso eficiente, visto que os usuários com acesso àquelas informações podem ser maliciosos, e assim capazes de divulgar informações sensíveis acerca daqueles indivíduos.

Com o desenvolvimento dos dispositivos móveis, a quantidade de dados coletados para o uso dos diversos serviços existentes tem crescido bastante, tornando a privacidade de dados muitas vezes uma moeda de troca, onde o usuário abre mão da sua privacidade em favor da prestação destes serviços. Nesse contexto, é importante identificar quais os tipos de dados não devem ser divulgados, a fim de garantir a aplicação de técnicas que permitam a proteção destes dados. Todavia os dados essenciais aos serviços devem ser fornecidos ao provedor para que este possa prestar o serviço na qualidade necessária ao usuário.

1.2.1. Privacidade em Microdados

Em geral, os dados são representados por uma tabela, onde cada linha corresponde a um registro do conjunto de dados e as colunas a atributos destes registros. A estes dados assim representados dá-se o nome de microdados. Neles cada registro corresponde a um

indivíduo e os atributos se referem a características ou propriedades do indivíduos. Por sua vez, no contexto de privacidade, os atributos podem ser classificados em [14]:

1. **Identificadores explícitos**: são aqueles atributos que identificam de maneira única os indivíduos, como "CPF", "nome", etc., e devem ser removidos antes da publicação dos dados;

2. **Semi-identificadores**: são aqueles que não são identificadores explícitos, mas podem identificar o usuário, quando relacionados. "Data de nascimento"e "CEP"são exemplos de atributos semi-identificadores;

3. **Atributos sensíveis**: possuem informações sensíveis a cerca dos indivíduos, como "doença", salário", etc.;

4. **Atributos não sensíveis**: são aqueles que não se enquadram em nenhuma das categorias citadas anteriormente.

Os atributos sensíveis são aqueles de maior interesse no nosso contexto porque apresentam potenciais danos ao seus donos em caso de divulgação. Por esse motivo, tais atributos necessitam ser protegidos. A Tabela 1.1 ilustra um exemplo de registros de indivíduos contendo atributos identificadores explícitos e semi-identificadores, que precisam ser protegidos.

Identificadores Explícitos		**Semi-identificadores**			
ID	**Nome**	**Idade**	**Gênero**	**Endereço**	**Telefone**
1	Isabela	22	Feminino	Av. I	99998 1324
2	João	25	Masculino	Av. K	99998 1454
3	Iago	25	Masculino	Av. K	99998 3245
4	Maria	31	Feminino	Rua J	99998 3465

Tabela 1.1. Exemplos de identificadores explícitos e semi-identificadores em dados tabulados de indivíduos.

1.2.2. Proteção e Privacidade de Dados

A fim de estabelecer a confiança dos indivíduos e o consentimento para a utilização dos seus dados, faz-se necessário garantir a proteção dos dados pessoais coletados. A abordagem mais promissora para solucionar o problema da preservação de privacidade consiste em anonimizar os dados antes de sua liberação para uso [15], visando impedir a exposição de dados sensíveis dos indivíduos. No processo de anonimização, um conjunto de dados D é transformado em um conjunto de dados D', por meio de modificações sobre os dados. Esta transformação se dá por meio de técnicas de *generalização*, *supressão* e *pertubação*.

A generalização modifica os atributos semi-identificadores dos registros por valores mais gerais, aumentando a incerteza de um adversário associar um indivíduo a seus dados, ou a atributos sensíveis, no conjunto de dados. Na abordagem mais comum de

generalização, o valor de um atributo semi-identificador que se deseja proteger nos diferentes registros é substituído por um valor generalizado. A Figura 1.3 ilustra o processo de generalização sobre o atributo Telefone dos registros da Tabela 1.1, onde, nas folhas da árvore, têm-se os valores originais para o atributo Telefone. No segundo nível agrupam-se os telefones cujos 6 primeiros dígitos correspondem, enquanto que no nível seguinte em direção ao topo da hierarquia são agrupados os telefones cujos 5 primeiros dígitos correspondem.

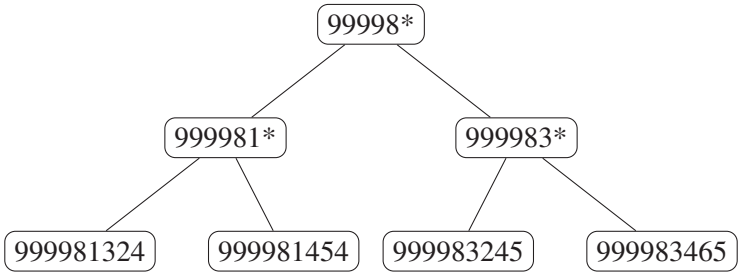

Figura 1.3. Exemplos de generalização do atributo semi-identificador Telefone.

Voltando à Tabela 1.1, após passar por um processo de anonimização por generalização dos valores do atributo Telefone até o topo da hierarquia, a publicação dos dados definidos como semi-identificadores poderia ser realizada de acordo com os valores observados na Tabela 1.2. Veja que este processo dificulta a re-identificação de um dos indivíduos caso seja de conhecimento externo o número do telefone deste indivíduo.

Identificadores Explícitos		**Semi-identificadores**			
ID	**Nome**	**Idade**	**Gênero**	**Endereço**	**Telefone**
1	Isabela	22	Feminino	Av. I	99998*
2	João	25	Masculino	Av. K	99998*
3	Iago	25	Masculino	Av. K	99998*
4	Maria	31	Feminino	Rua J	99998*

Tabela 1.2. Exemplo de semi-identificador anonimizado por generalização.

A supressão de dados remove valores ou substitui um ou mais valores de um conjunto de dados por algum valor especial, possibilitando a não descoberta de semi-identificadores por adversários. Alguns dos principais tipos de supressão:

- **Supressão de registro**: a supressão de registro remove um registro inteiro do conjunto de dados, consequentemente nenhum valor de atributo é disponibilizado para uso [5, 19].
- **Supressão de valor**: a supressão de valor remove ou substitui todas as ocorrências de um valor de um atributo semi-identificador por um valor especial, como "*". Por exemplo, em uma tabela de funcionários de uma empresa, os valores de atributo salário abaixo de R$ 30.000,00, podem ser removidos ou substituídos por "*", enquanto os demais valores não sofrem distorções [33, 34].

- **Supressão de células**: nessa técnica, apenas algumas instâncias de valores de um atributo são removidas ou substituídas por um valor especial, caracterizando uma *supressão local* [23]. Por exemplo, pode-se remover apenas metade dos valores de atributo salário abaixo de R$ 30.000,00, em uma tabela de empregados. Assim, instâncias de salário podem conter valores abaixo ou acima de R$ 30.000,00, além de valores suprimidos. Entretanto, essa estratégia pode levar a inconsistências em eventuais análises de dados.

Por último, mas, não menos importante, a perturbação substitui os valores dos atributos semi-identificadores originais por valores fictícios, de modo que informações estatísticas calculadas a partir dos dados originais não se diferenciem significativamente de informações estatísticas calculadas sobre os dados perturbados. As técnicas mais comuns de perturbação de dados são:

- **Adição de ruído**: essa técnica é aplicada comumente sobre atributos numéricos. Ela substitui o valor original "v" de um atributo por "$v + r$", em que "r" é um valor, chamado de ruído, escolhido aleatoriamente a partir de uma distribuição. O valor "v" do atributo também pode ser substituído por "$v \times r$". Os valores dos atributos são portanto, perturbados com um determinado nível de ruído, que pode ser adicionado ou multiplicado pelo valor original de cada atributo [31];

- **Permutação de dados**: nesta abordagem os valores de um mesmo atributo de dois registros diferentes são permutados. Isso mantém algumas características estatísticas dos dados, como frequência dos atributos e contagem [10]. Essa técnica não altera o domínio dos atributos, mas as possíveis permutações de valores diferentes podem levar a valores nos registros sem sentido, e com isso, informações equivocadas;

- **Geração de dados sintéticos**: nesta técnica, um modelo estático é inicialmente gerado a partir do conjunto de dados e, após isso, são gerados dados sintéticos que seguem o modelo gerado [1]. Esses dados sintéticos são os que devem ser disponibilizados para o uso final. A vantagem desta técnica é que todas as propriedades estatísticas dos dados são mantidas. Entretanto, pode-se gerar também alguns valores sem sentido e que não são condizentes com o mundo real.

Uma vez anonimizado os dados, através de técnicas de generalização, supressão ou perturbação, é possível permitir o compartilhamento de informações com outras entidades, as quais poderão utilizá-las para diversas finalidades, sem que haja violação de privacidade. Todavia, a modificação dos dados originais no processo de anonimização causa perda de utilidade dos mesmos. Portanto, é necessário encontrar um equilíbrio entre a proteção desejada e a utilidade dos dados, a fim de se permitir operações de agregação ou mesmo análise dos dados.

1.3. Ataques à Privacidade

O objetivo dos modelos de privacidade é preservar ao máximo possível a privacidade dos donos dos registros em um conjunto de dados. Já o objetivo do atacante, se opondo a

isso, é justamente utilizar de todos os recursos a sua disposição para retirar o máximo de informação dos registros. Dessa forma, o atacante muitas vezes necessita de conhecimentos prévios que o auxiliem a fazer inferências sobre o conjunto de dados. Por exemplo, o adversário que trabalha no mesmo local da vítima, pode possuir conhecimento sobre a mesma, tais como, endereço residencial e cargo na empresa, permitindo a inferência de informações sensíveis, como localização, opção sexual, etc. Em se tratando de publicação de dados, o atacante pode ter acesso a outros conjuntos de dados previamente publicados, e assim, cruzar referências para descobrir novas informações sensíveis da vítima. Portanto, podemos concluir que o conhecimento do adversário é imensurável e imprevisível e deve ser levado em consideração nas soluções de preservação de privacidade, apesar de suas características.

Um adversário é capaz de violar a privacidade dos usuários por meio de diversos ataques que citaremos a seguir:

- **Ataque de Ligação ao Registro:** este ataque tem por objetivo re-identificar o registro de um usuário, cujas informações pertencem ao conjunto de dados publicado;
- **Ataque de Ligação ao Atributo:** o objetivo do adversário é ser capaz de inferir atributos sensíveis do usuário mesmo sem re-identificar seu registro, com base nos valores sensíveis relacionados ao grupo que o usuário pertence.
- **Ataque de Ligação à Tabela:** este tipo de ataque assume que o adversário sabe que o registro do usuário foi publicado. Neste ataque o intuito é inferir se a vítima está presente ou ausente nos dados publicados.
- **Ataque Probabilístico:** este ataque tem o foco de destacar como o adversário mudaria seu pensamento probabilístico sobre um usuário depois de ter acesso ao conjunto de dados disponível.

1.4. Modelos de Privacidade Sintático

Os modelos de privacidade sintáticos procuram garantir a privacidade dos indivíduos ao exigir que, após o processo de anonimização, o conjunto de dados vai atender certas condições específicas da técnica aplicada. Para isto, estes modelos, usualmente aplicam uma transformação nos registros por meio de técnicas de supressão e/ou generalização até que esta condição seja alcançada. Iremos apresentar alguns dos modelos de privacidade sintático mais utilizados em preservação de dados.

1.4.1. k-anonimato

O k-anonimato é o modelo de privacidade mais conhecido no campo da anonimização de dados [30]. Esse modelo assegura que, para cada combinação de valores de semi-identificadores, existem pelo menos k registros no conjunto de dados, formando uma classe de equivalência. O k-anonimato atua sobre o princípio da indistinguibilidade, isto é, cada registro em um conjunto de dados k-anônimo é indistinguível de pelo menos outros $k-1$ registros em relação ao conjunto de semi-identificadores. Assim, garante-se que cada registro não pode ser ligado a um indivíduo por um adversário com probabilidade maior que $\frac{1}{k}$.

O valor de k define o nível de privacidade e, consequentemente, afeta diretamente a perda de utilidade. Assim, um valor de k grande implica em uma maior proteção dos dados, entretanto, diminui a utilidade dos mesmos, por ser necessário adicionar grande volume de ruído a fim de se alcançar classes de equivalência com pelo menos k registros. É importante ressaltar que não existem abordagens analíticas para determinar um valor ótimo para o parâmetro k [9], sendo este um problema NP-difícil [23]. Dessa forma, cabe aos *dataholders* esta complexa tarefa quando da aplicação do processo de anonimização por k-anonimato sobre um conjunto de dados.

A Tabela 1.3 ilustra a aplicação do modelo k-anonimato para $k = 2$. São atributos identificadores explícitos Placa, Motorista e CPF enquanto que são atributos sensíveis Tipo de Multa e Valor da Multa. Os atributos restantes semi-identificadores: Data de Nascimento e Data da Infração.

	Placa	**Motorista**	**CPF**	**Data de Nascimento**	**Data da Infração**	**Tipo de Multa**	**Valor da Multa (R$)**
1	UVW-1840	Gigi	223.512.956	14/03/1980	03/01/2013	1	170
2	AXO-2064	André Luis	523.512.511	04/03/1980	03/01/2013	2	250
3	AUG-1046	Juçara Silva	123.998.687	24/05/1980	03/01/2013	1	170
4	FBI-1001	Bruno Lima	230.320.523	20/04/1982	04/01/2013	1	170
5	ACO-6241	Abu Ali	221.320.876	20/05/1982	04/01/2013	2	250
6	ABA-5012	Pedro Ramires	210.329.890	13/05/1982	05/01/2013	2	250
7	HBV-2002	Eduardo Neto	538.687.045	15/05/1982	05/01/2013	1	170

Tabela 1.3. Dados sobre infrações de trânsito.

Em um processo de anonimização dos dados da Tabela 1.3 são aplicadas a supressão nos identificadores explícitos e a generalização nos atributos sensíveis, gerando a Tabela anonimizada 1.4. Nesta tabela podemos perceber quatro classes de equivalência para os semi-identificadores: Classe A = "$03/1980, 01/2013$" nas linhas 1 e 2; Classe B = "$05/1980, 01/2013$" registro 3; Classe C = "$04/1982, 01/2013$" com o registro 4 e Classe D = "$05/1982, 01/2013$" nas linhas 5, 6 e 7. Observe, que mesmo após aplicar um processo inicial de anonimização, o k-anonimato ainda não foi alcançado, já que as classes B e C, não possuem uma quantidade mínima requerida de 2 registros, sendo, portanto, necessário, algum novo processo de transformação. Uma estratégia, seria então remover os registros 3 e 4, como podemos observar na Tabela 1.5. Desta forma, agora observamos apenas 2 classes de equivalência, contendo a quantidade mínima de dois registros por classe, garantindo o k-anonimato.

1.4.2. l-diversidade

Assim como o k-anonimato, o l-diversidade age sobre o princípio da indistinguibilidade. Entretanto, o k-anonimato apesar de apresentar uma alta eficácia na prevenção contra ataques de ligação ao registro, não se mostra adequado contra ataques de ligação ao atributo, *i.e.*, ataques em que um adversário procura inferir informações sensíveis sobre registros mesmo sem identificá-los. Tomamos como exemplo a Tabela 1.5 que garante o k-anonimato, para $k = 2$, contendo pelo menos dois registros em cada uma das classes de equivalência. Observe, que se o atacante tiver conhecimento que o usuário Pedro Ramires nasceu em 05/1982, e recebeu uma infração em janeiro de 2013, ele poderá inferir com

	Placa	Motorista	CPF	Data de Nascimento	Data da Infração	Tipo de Multa	Valor da Multa (R$)
1	*	*	*	03/1980	01/2013	1	170
2	*	*	*	03/1980	01/2013	2	250
3	*	*	*	05/1980	01/2013	1	170
4	*	*	*	04/1982	01/2013	1	170
5	*	*	*	05/1982	01/2013	2	250
6	*	*	*	05/1982	01/2013	2	250
7	*	*	*	05/1982	01/2013	1	170

Tabela 1.4. Dados sobre informações de trânsito anonimizados.

	Placa	Motorista	CPF	Data de Nascimento	Data da Infração	Tipo de Multa	Valor da Multa (R$)
1	*	*	*	03/1980	01/2013	1	170
2	*	*	*	03/1980	01/2013	2	250
3	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*
5	*	*	*	05/1982	01/2013	2	250
6	*	*	*	05/1982	01/2013	2	250
7	*	*	*	05/1982	01/2013	1	170

Tabela 1.5. Tabela no modelo *2*-anonimato.

uma probabilidade de $\frac{2}{3}$ que a multa recebida por Pedro foi de 250 reais. Superior à $\frac{1}{2}$, desejada pelo modelo k-anonimato.

O l-diversidade busca prover proteção contra ataques de ligação ao atributo, garantindo que para cada classe de equivalência, exista pelo menos l valores distintos para cada atributo sensível. Assim, o que se pretende é que um atacante, mesmo com conhecimento prévio sobre a classe de equivalência de um registro, não seja capaz de inferir o atributo sensível do mesmo com probabilidade maior que $\frac{1}{l}$.

Tomando por exemplo a Tabela 1.6, onde a probabilidade de se identificar que o indivíduo tem asma, valor do atributo sensível "Doença", caso o atacante tenha conhecimento de que o CEP do indivíduo é 540040, é de 100%, superior a $\frac{1}{4}$ exigido pelo modelo 4-anonimato. Convertendo a Tabela 1.6 para o modelo 3-diversidade, não é preciso fazer nenhuma alteração nos registros da classe A (linhas 1 a 4), pois esta já possui no mínimo 3 valores distintos para o atributo sensível. Entretanto, a classe B (linhas 5 a 8) possui todos os valores de atributos sensíveis iguais. Uma solução simples seria suprimir os registros das linha 5 a 8. Outra solução seria modificar os valores do atributo sensível destas linhas por valores diferentes que garantam a diversidade, conforme Tabela a 1.7 que atende, portanto, o modelo 4-anonimato e 3-diversidade.

Outros modelos foram propostos como uma extensão do l-diversidade, como o t-proximidade [21] e o p-sensibilidade [32], com a finalidade de prover uma maior garantia de preservação de privacidade tanto contra ataques de ligação ao registro, como ao atributo.

	Idade	CEP	Cidade	Doença
1	<70	560001	*	Sinusite
2	<70	560001	*	Gripe
3	<70	560001	*	Zika
4	<70	560001	*	Hérnia
5	<35	540040	*	Asma
6	<35	540040	*	Asma
7	<35	540040	*	Asma
8	<35	540040	*	Asma

Tabela 1.6. *4*-anonimato

	Idade	CEP	Cidade	Doença
1	<70	560001	*	Sinusite
2	<70	560001	*	Gripe
3	<70	560001	*	Zika
4	<70	560001	*	Hérnia
5	<35	540040	*	Sinusite
6	<35	540040	*	Zika
7	<35	540040	*	Asma
8	<35	540040	*	Asma

Tabela 1.7. *4*-anonimato e *3*-diversidade.

1.4.3. δ-presença

O δ-presença, uma extensão ao k-anonimato, é um modelo que busca proteger a privacidade de dados dos indivíduos contra ataques de ligação à tabela [24]. O modelo define o limite $\delta = \delta_{max}, \delta_{min}$ para a probabilidade de um adversário inferir a a presença de um indivíduo na tabela. Desta forma, indiretamente, o modelo também garante a privacidade contra ataques de ligação ao registro e ao atributo, uma vez que a probabilidade de um ataque de ligação ao registro ou ao atributo sensível ser bem sucedido está limitado por δ_{max}.

Para ilustrar um ataque de ligação à tabela, imagine um atacante que tem conhecimento sobre a Tabela A (1.8), no formato 4-anônimo, com duas classes de equivalência: E_1 (Vendedor, Feminino, [30,35]), 5 indivíduos; E_2 (Professor, Masculino,[35-40]), 4 indivíduos. Caso a Tabela B (1.9) seja publicada, onde todos os indivíduos de B estão em A, é possível identificar que a probabilidade de a vendedora Maria estar na Tabela B é de $\frac{4}{5}$, uma vez que há 5 registros na mesma classe de equivalência de Maria (E_1 em A), e que em B há apenas 4 registros na classe de equivalência E_1.

Nome	Profissão	Gênero	Idade
Lucas	Professor	Masculino	[35-40)
Isaías	Professor	Masculino	[35-40)
João	Professor	Masculino	[35-40)
Mateus	Professor	Masculino	[35-40)
Maria	Vendedor	Feminino	[30-35)
Fátima	Vendedor	Feminino	[30-35)
Marta	Vendedor	Feminino	[30-35)
Irene	Vendedor	Feminino	[30-35)
Natália	Vendedor	Feminino	[30-35)

Tabela 1.8. Tabela no formato *4*-anonimato.

Profissão	Gênero	Idade	Multa
Professor	Masculino	[35-40)	250
Professor	Masculino	[35-40)	300
Professor	Masculino	[35-40)	250
Vendedor	Feminino	[30-35)	250
Vendedor	Feminino	[30-35)	300
Vendedor	Feminino	[30-35)	450
Vendedor	Feminino	[30-35)	250

Tabela 1.9. Tabela de pacientes no formato *3*-anonimato.

A Tabela 1.10 apresenta a aplicação do δ-presença para um $\delta_{max} = \frac{1}{2}$, através da supressão de registros. As linhas 2, 5 e 6 foram removidas. Desta forma, a probabilidade de se identificar a presença de qualquer indivíduo da Tabela 1.8 na Tabela 1.9 é inferior ou igual a $\frac{1}{2}$.

Profissão	**Gênero**	**Idade**	**Multa**
Professor	Masculino	[35-40)	250
*	*	*	*
Professor	Masculino	[35-40)	250
Vendedor	Feminino	[30-35)	250
*	*	*	*
*	*	*	*
Vendedor	Feminino	[30-35)	250

Tabela 1.10. Tabela de pacientes no formato 3-anonimato e δ-presença.

1.5. Modelo de Privacidade Diferencial

Diferentemente dos modelos apresentados até agora, que buscam garantir a preservação de privacidade dos indivíduos na publicação de dados em formato tabulado , o modelo de Privacidade Deferencial procura garantir a preservação de privacidade na publicação de resultados de consultas. Seu objetivo é evitar que o conhecimento adversário do atacante aumente a probabilidade de se expor os indivíduos do conjunto de dados, ou seja, evitando ataques probabilísticos. Para isto, as respostas destas consultas são perturbadas, com a adição de ruído, como forma de garantir a privacidade dos indivíduos.

Proposta por Dwork [11], a Privacidade Diferencial (PD) consiste em um modelo matemático que oferece sólidas garantias de privacidade. A PD é um modelo semântico, cujo objetivo é garantir a utilidade dos dados ao mesmo tempo que fornece proteção contra ataques de conhecimento prévio. Em um contexto geral, seu objetivo é proteger os dados dos usuários na publicação de informação agregada sobre o conjunto de dados. Para isto, este método requer que a adição ou remoção de um único indivíduo tenha um efeito insignificante sobre a resposta de uma requisição. De forma mais precisa, a PD requer que, para quaisquer dois conjuntos de dados vizinhos (conjuntos de dados que se diferenciam em apenas um registro, Figura 1.4), a probabilidade de uma consulta sobre estes conjuntos retornar o mesmo valor v deve estar limitada por $exp(\varepsilon)$. Tipicamente, alcança-se ε-Privacidade Diferencial ao adicionar um ruído aleatório controlado à resposta das consultas, utilizando para isto de um mecanismo, Figura 1.5.

ID	Peso (Kg)	Altura (m)
1	87,2	1,70
2	81,2	1,62
3	74,2	1,75
4	60,0	1,61
5	78,5	1,58

(a)

ID	Peso (Kg)	Altura (m)
1	87,2	1,70
2	81,2	1,62
4	60,0	1,61
5	78,5	1,58

(b)

Figura 1.4. Exemplo de conjuntos de dados vizinhos.

Um mecanismo M garante ε-Privacidade Diferencial se para quaisquer conjuntos de dados vizinhos D_1 e D_2,

$$Pr[M(D_1)] \leq exp(\varepsilon) \times Pr[M(D_2)],$$

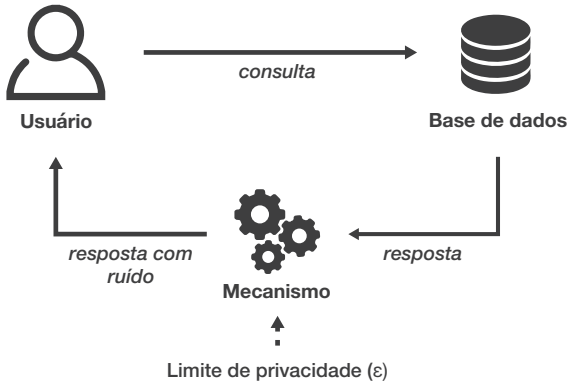

Figura 1.5. Ambiente interativo no modelo de Privacidade Diferencial.

onde Pr é a probabilidade da resposta adicionada de ruído aplicado por M. Isto é, a diferença entre as probabilidades de uma consulta retornar o mesmo valor v em dois conjuntos de dados vizinhos é limitada por ε.

1.5.1. Mecanismo

Como já dito anteriormente, o mecanismo é o responsável pela adição de ruído controlado à resposta da consulta a fim de garantir ε-Privacidade Diferencial. A quantidade de ruído necessária depende do tipo de consulta f aplicada sobre o conjunto de dados D. Isto é importante para introduzirmos o conceito de sensibilidade de uma função de consulta, que vai medir, justamente, quanta diferença na resposta da consulta um usuário faz ao ser removido ou adicionado a D. Desta forma, podemos definir a sensibilidade da função f como sendo:

$$\Delta f = \max_{D_1, D_2 \in D} \| f(D_1) - f(D_2) \|_1,$$

para todo D_1, D_2 diferindo de no máximo um elemento, ou seja, D_1 e D_2 são vizinhos [12]. A Figura 1.6 ilustra um conjunto de dados simples D. Para uma consulta f sobre D que retorna a soma de imóveis, a resposta é 14. A Figura 1.7 apresenta conjuntos de dados vizinhos e sua respectivas respostas para a mesma consulta f. Podemos então calcular a sensibilidade de Δf sobre D como 7, que é a maior diferença entre as respostas da consulta f sobre os conjuntos de dados vizinhos.

ID	Nome	N° de Imóveis
1	José	4
2	Antônio	2
3	Raimundo	7
4	Francisco	1

Figura 1.6. Exemplo de conjunto de dados original contendo o número de imóveis de cada indivíduo (Fonte: [8]).

O mecanismo de Laplace é normalmente utilizado para alcançar a Privacidade Diferencial em consultas sobre dados numéricos que retornam valores agregados. A adição de ruído segue uma função de densidade de probabilidade de uma variável aleatória com

ID	Nome	N° de Imóveis
2	Antônio	2
3	Raimundo	7
4	Francisco	1

$f(D_1) = 2 + 7 + 1 = 10$

ID	Nome	N° de Imóveis
1	José	4
3	Raimundo	7
4	Francisco	1

$f(D_2) = 4 + 7 + 1 = 12$

ID	Nome	N° de Imóveis
1	José	4
2	Antônio	2
4	Francisco	1

$f(D_3) = 4 + 2 + 1 = 7$

ID	Nome	N° de Imóveis
1	José	4
2	Antônio	2
3	Raimundo	7

$f(D_4) = 4 + 2 + 7 = 13$

Figura 1.7. Conjuntos de dados vizinhos gerados a partir da base original e suas respectivas respostas da consulta f (Fonte: [8]).

distribuição de Laplace com média μ e escala b de forma que

$$Laplace_{\mu,b}(x) = \frac{1}{2b}\exp(-\frac{|x|}{b}).$$

A Figura 1.8 mostra a distribuição de probabilidades de respostas desejada em um modelo ε-Privacidade Diferencial quando aplicado sobre dois conjuntos de dados vizinhos D_1 e D_2 utilizando o mecanismo de Laplace.

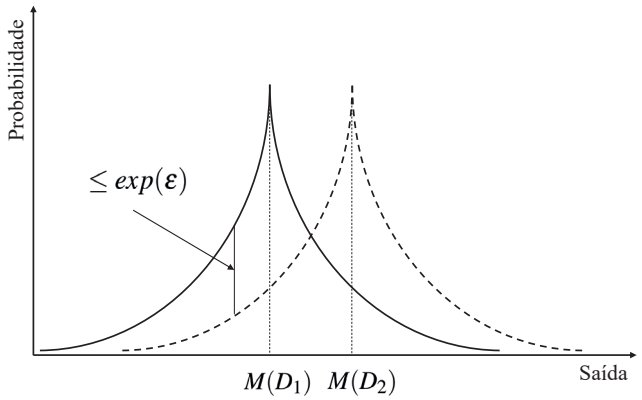

Figura 1.8. Probabilidades de saída de um algoritmo aleatório M sobre os conjuntos de dados vizinhos D_1 e D_2.

Apresentemos agora a definição formal do mecanismo de Laplace (Definição 1).

Definição 1 *Dada uma função de consulta $f : D \to \Re$, o mecanismo de Laplace M:*

$$M_f(D) = f(D) + Laplace(0, \Delta f/\varepsilon)$$

Ruído	*f(D) + ruído*	*Pr(f(D) + ruído)%*
-4,58	9,42	3,70
-0,15	13,85	6,98
12,15	26,15	1,25
-6,43	7,57	2,85
2,89	16,89	4,72

Tabela 1.11. Cinco possíveis valores de ruído, resposta e probabilidade de ocorrência após a aplicação da Privacidade Diferencial.

fornece ε-Privacidade Diferencial, onde $Laplace(0, \Delta f/\varepsilon)$ retorna uma variável aleatória da distribuição de Laplace com média zero e escala $\Delta f/\varepsilon$.

Considere a Tabela 1.11. Ela apresenta cinco possíveis ruídos aplicados pelo mecanismo de Laplace para a mesma consulta f sobre o conjunto de dados da Tabela 1.6, o ruído de $-4,58$ possui uma probabilidade de ocorrência de $3,70\%$, resultando em uma resposta anonimizada de $9,42$ imóveis, frente à resposta 14 que seria retornada caso o ruído não tivesse sido adicionado pelo mecanismo. Observe que repetidas execuções da consulta f retornam valores distintos devido a aleatoriedade do ruído adicionado pelo mecanismo de Laplace à resposta de cada execução.

1.6. Privacidade de Localização

O desenvolvimento dos dispositivos móveis tem contribuído para um crescimento na popularidade dos serviços de localização. Serviços que, como o próprio nome diz, dependem da localização dos usuários para sua prestação. São alguns exemplos dos mais comuns serviços de localização:

- **Navegação**: permite o usuário obter direções para um ponto de interesse geograficamente localizado. Os dados de localização do usuário são coletados para prover instruções de direção em tempo real. São algumas aplicações: Google Maps e Waze.
- **Aplicações de tempo (clima)**: estes serviços provêm condições do tempo, bem como previsões. A localização do usuário é usada para obter informações relevantes sobre o clima do local atual.
- **Jogos**: utilizam a localização do usuário no contexto do ambiente virtual do jogo. Os mais recentes usam tecnologia de realidade aumentada, onde a movimentação do usuário em tempo real se reflete no jogo. Exemplo desse tipo de jogo é Pokemon GO.
- **Serviços de Recomendação**: estes serviços utilizam da localização do usuário para enviar recomendações de locais de interesse próximos. São exemplos: Foursquare e Yelp.

Apesar da popularidade destes serviços, a natureza dos dados de localização tem levado a sérios questionamentos quanto a preservação da privacidade dos usuários no

uso destes serviços. Por carregarem consigo muita informação, estes dados são capazes de potencializar violações de privacidade, requisitando assim a utilização de técnicas de anonimização a fim de garantir a preservação de privacidade dos indivíduos.

Há um consenso entre os pesquisadores que a privacidade é um assunto complexo, com muitas questões envolvidas, sendo um direito dos indivíduos a ser preservado [8]. Entretanto, o que temos muitas vezes visto na prática, é a privacidade dos usuários sendo utilizada como moeda de troca nos serviços de localização, nos quais o usuário fornece informações pessoais para fazer uso dos serviços. Nesta seção iremos apresentar os principais fundamentos em preservação de dados de localização, destacando a natureza deste tipo de dado, alguns dos principais tipos de ataques a dados de localização, e as técnicas usadas para preservar a privacidade dos indivíduos que fazem uso desse tipo de serviço.

1.6.1. Dados de Localização

O vazamento da informação de localização dos usuários pode permitir uma série de ataques de indivíduos maliciosos, que vão desde vigilância física e perseguição, até roubo de identidade. Outro risco é o de inferências de informações sensíveis. Estes ataques são possíveis devido a natureza dos dados de localização, que carregam consigo muita informação. Por exemplo, se a informação de uma localização de um indivíduo indicar um hospital. Neste caso o dado já sugere uma série de informações relacionadas ao local, por exemplo, doenças, horário de funcionamento, profissão, visita a conhecidos, dentre outras.

As informações de localização são obtidas por meio de sistemas de posicionamento global (GPS), que estão contidos na maioria dos aparelhos celulares da atualidade, o que tem impulsionado o crescimento de serviços baseados em localização (LBS). Estes serviços fornecem valor agregado aos seus usuários através da integração da localização ou posição de seus dispositivos móveis a outras informações. A popularidade destes serviços tem aumentado vertiginosamente a quantidade de informações de localização coletadas, o que por si só tem ampliado os riscos de quebra de privacidade.

A Figura 1.9 ilustra um típico serviço baseado em localização. São alguns componentes básicos de um LBS:

- **GPS**: permite determinar a localização dos objetos envolvidos, *i.e.*, usuários, ou outra entidade qualquer. O GPS é o mais popular sistema de posicionamento. Ele é um mecanismo de posicionamento por satélite que fornece a um aparelho receptor a sua posição.

- **Usuários**: são participantes que irão usufruir do serviço baseado em localização prestado. Através de dispositivos como *smartphones*, *notebooks*, *wearables*, os usuários se conectam ao meio de comunicação e enviam requisições ao provedor do serviço

- **Rede de comunicação**: é o meio através do qual acontece o tráfego de informações entre os participantes. Normalmente o meio utilizado é a rede de banda larga móvel, como a 4G.

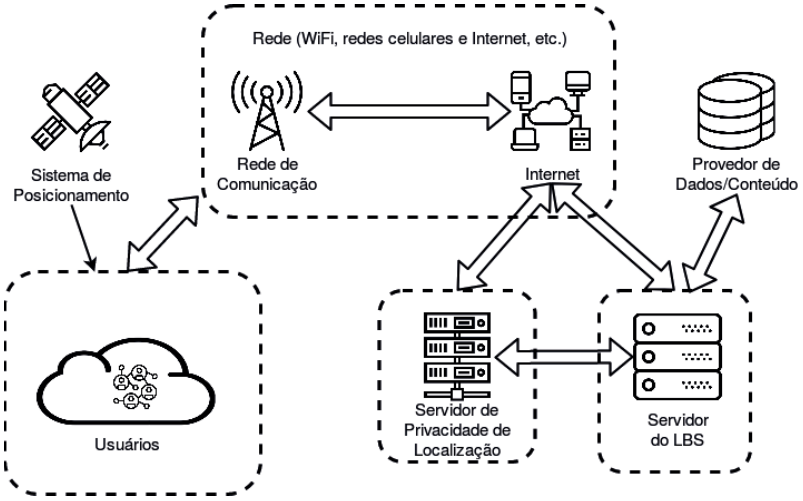

Figura 1.9. Modelo de sistema de serviços baseados em localização

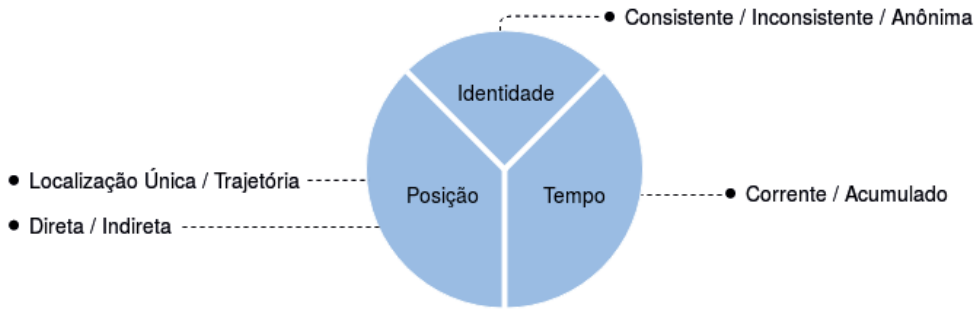

Figura 1.10. Três atributos da informação de localização

- **Servidor do LBS**: é o responsável por receber as requisições dos usuários e prestar o serviço baseado em localização de acordo com sua natureza, seja para encontrar uma localização, seja para auxiliar na navegação do usuário, ou um outro tipo de serviço qualquer que utiliza a informação de localização enviada na requisição.

- **Provedor de Conteúdo/Dados**: provedor de Conteúdo/Dados fornecem dados e conteúdo ao servidor LBS. Alguns provedores de LBS possuem seus próprios dados e conteúdo, enquanto outros usam um terceiro para fornecer esse serviço.

- **Servidor de Privacidade**: o servidor de privacidade de localização executa os algoritmos de preservação de privacidade, como anonimização e criptografia e pode ser de propriedade e operado pelo provedor LBS ou por terceiros.

Essas informações de localização são constituídas de três partes, que podem ser observadas na Figura 1.10, identidade, tempo e posição.

A *identidade* pode ser um endereço de email, nome ou qualquer outra informação que torne um indivíduo distinguível dos outros. Ela pode ser: *(i)* consistente, que é aquela requisitada obrigatoriamente para o acesso a um LBS, como um nome de usuário; *(ii)*

inconsistente, quando pode haver o uso de pseudônimos; *(iii)* anônima, onde há a ausência de uma identificação.

O *tempo* é referente ao momento a qual as localizações estão associadas. Em geral, os serviços de localização associam as localizações a marcos de tempo. Esta informação temporal pode ser classificada como acumulada ou corrente. As aplicações em tempo acumulado não publicam as informações de localização em tempo real, mais em um tempo posterior ao atual. Um exemplo destas aplicações é o sistema de rastreio do *Fitbit*, que coleta as informações percorrida pelos usuários, mas só publica a trajetória computada depois de encerrado a atividade [22]. As aplicações em tempo real publicam as informações de localizações associados ao marco de tempo atual, de forma imediata.

A *informação espacial (posição)* é o principal meio para determinar a localização do usuário. As localizações podem ser descritas como um conjunto de coordenadas (latitude e longitude), ou por alguma outra forma de informação que pode ser vinculada a um local. As localizações podem ser únicas, quando não são correlacionadas umas com as outras, ou podem formar uma trajetória no caso em que são fortemente correlacionadas, o que acaba por gerar maiores riscos de exposição.

As localizações também podem ser classificadas em diretas ou indiretas. Os LBS tradicionais usam localizações diretas definidas por coordenadas de GPS, ou seja, utilizam o padrão adotado na realidade, em que a localização é composta de latitude e longitude. As localizações indiretas são aquelas estabelecidas com base na proximidade física, substituindo-se a localização exata pelo ponto de interesse mais próximo (POI) - entidade que representa uma localização, dotada de informação complementar sobre a mesma, como o nome do estabelecimento, horário de funcionamento, endereço, entres outras.

A privacidade de localização pode ser definida como a proteção desses três atributos que formam a informação de localização de uma pessoa. Blumberg e Eckersley [7] definem privacidade de localização como a capacidade de um indivíduo de se deslocar no espaço público com a expectativa de que, em circunstâncias normais, sua localização não será sistematicamente e secretamente registrada para uso posterior. Entretanto, é importante destacar que a garantia de privacidade de localização de indivíduos não é absoluta, uma vez que, no momento que você sai de casa, sua privacidade já está sendo exposta. Por exemplo, ao sair de casa, seu vizinho pode saber a hora que você chega em casa, que horas sai para trabalhar, se você tem algum tipo de animal de estimação, além de possíveis outras informações. Desta forma, pode-se identificar dois requisitos principais da privacidade de localização dos indivíduos: a expectativa de privacidade dos indivíduo de "circunstâncias normais", ou seja, o que o indivíduo espera em termos de exposição da sua localização, e a maneira como as informações são coletadas e usadas. A expectativa de privacidade de uma pessoa pode mudar com o tempo, assim como a forma como as informações de localização são coletadas e usadas também mudam. Logo, para avaliar a privacidade de localização do indivíduo, seus principais requisitos devem ser definidos do ponto de vista dos usuários.

Com isso, temos dois fatores que servem de base para avaliar a privacidade do indivíduo quanto a suas localizações. Estes fatores são caracterizados pelos seguintes pronomes interrogativos:

- *Como*: Como as informações são reveladas? É revelado secretamente ou publicamente? É criptografado ou não? E como a informação será usada?

- *Que*: Que tipo de informação é revelada? Por exemplo, um conjunto de coordenadas, um momento do tempo, a identidade do usuário anexada, dentre outros.

1.7. Tipos de Ataques

Um atacante é qualquer entidade que possa ter acesso aos dados de localização de um ou de vários indivíduos. Sendo assim, um adversário pode ser desde o próprio provedor do serviço de localização, ou até mesmo um cientista de dados que tenha acesso a uma publicação dos dados [26]. Na maioria das vezes, o atacante é considerado honesto, mas curioso [16]. Ou seja, o provedor do serviço ou um cientista de dados, potenciais adversários, se comportam conforme se espera deles na prestação do serviço, entretanto, eles são capazes de explorar de todas as formas possíveis quaisquer dados que tenham acesso. Para exemplificar, vamos supor um serviço de localização onde o usuário busca informações sobre uma determinada loja de animais. O provedor do serviço irá prestar este serviço, informando ao usuário todas as informações disponíveis sobre a loja requisitada. Estas informações são coletadas pelo provedor do serviço e podem ser usadas para obter outras informações que lhe permitam tirar algum proveito, como por exemplo, traçar o perfil do usuário e enviar sugestões de propagandas e serviços de seu interesse, atividades que vão além daquela fornecida pelo serviço.

Por ser carregado de informação, o conhecimento adversário torna os usuários ainda mais vulneráveis a ataques. Entretanto, o poder desse conhecimento vai depender se estes dados sofreram ou não algum tipo de transformação, tais como supressão, generalização ou pertubação já apresentados na Seção 1.2. Em se tratando de dados de localização, o conhecimento de contexto pode gerar potenciais riscos de violação de privacidade. São exemplos de conhecimento de contexto: o número de usuários em uma área em um determinada hora do dia; a relação entre diferentes usuários; as restrições de localização de uma determinada área, como rede de ruas, área de preservação; a distribuição e a probabilidade estatística associada às localizações.

Desta forma, em função da expectativa de privacidade dos indivíduos e na forma como as informações de localização são coletadas e usadas, um atacante e seu ataque podem ser caracterizados por "como" se obtem a informação, "como" o ataque é lançado, "que" informação ou conhecimento se detem, e "que/quem" é o alvo.

Em particular, assume-se que o atacante possui qualquer base de dados que contém conhecimentos adicionais sobre a semântica das informações de localização dos usuários. Além disso, o provedor do LBS pode identificar que o usuário está utilizando alguma técnica de preservação de privacidade de localização a fim de garantir a utilização do serviço sem expor sua localização real.

1.7.1. Ataque de Identidade

Os ataques de identidade, também chamados de ataques de desanonimização, procuram cruzar conhecimentos adversários de diversas fontes a fim de determinar a identidade do alvo. São alguns exemplos deste tipo de ataque:

- **Ataque de identificação pessoal**: através do conhecimento prévio pessoal de um indivíduo, busca-se identificar o indivíduo dentro do conjunto de dados, a fim de se obter toda a informação a ele associada no conjunto de dados. Considere o exemplo: o atacante tem o conhecimento sobre o endereço residencial de um indivíduo. Através dele, mesmo em um conjunto de dados anonimizado, se o atributo endereço não tiver sido protegido, o atacante poderá identificar o dono do registro em função do endereço residencial exposto.

- **Ataque de presença agregada**: identificar a identidade com base na relação entre dois indivíduos ou através de uma propriedade agregadora, por exemplo pessoas agrupadas próximas a um evento, uma estação de Pokemon Go, ou uma loja com ofertas, dentre outros eventos.

1.7.2. Ataque de Localização

Os ataques de localização consistem em identificar as informações espaciais e temporais referentes a um indivíduo. São alguns exemplos de ataques de localização: *(i)* ataque a localizações sensíveis procura identificar localizações importantes, como residência ou local de trabalho; *(ii)* ataque de revelação de presença ou ausência determina se um usuário está presente ou ausente em determinadas localizações em um determinado horário do dia; *(iii)* ataque de rastreamento identifica uma sequência de eventos para rastrear um usuário.

1.7.3. Métodos de Ataque

Os métodos de ataque dizem respeito à forma como o ataque é realizado. São alguns destes métodos:

- **Ataques de vinculação de contexto**: é a forma mais comum em ataques de localização. O conhecimento de contexto é combinado com a informação de localização obtida para se chegar à localização precisa da vítima em um ataque de localização. Por exemplo, um indivíduo ao realizar um *check-in* em um hospital, preenche seus dados informando seu endereço residencial. Se um atacante tiver conhecimento do endereço residencial do indivíduo, ele poderá usá-lo para identificar este indivíduo na lista de *check-in* do hospital.

- **Ataques probabilísticos**: este tipo de ataque é baseado na coleta de informações estatísticas sobre o ambiente [28]. Pode ser aplicado tanto para ataques de identidade, como de localização. Sendo assim, a localização do usuário pode ser inferida em razão da probabilidade do usuário estar em uma determinada localização em um horário preciso.

- **Ataque de conluio de usuários maliciosos**: é realizado por usuários que usam o mesmo provedor de serviços baseado em localização, que colidem para realizar vários ataques. Por exemplo, usuários em conluio utilizam sua posição para obter, do serviço, a distância da vítima, e baseado nisso calculam a exata localização da vítima.

1.8. Modelos de Privacidade em Serviços de Localização

Como já citado nas seções anteriores, a exposição dos dados de localizações a agentes maliciosos pode levar a sérios riscos de violação de privacidade. Portanto, diversas técnicas de privacidade foram propostas a fim de mitigar o problema. Nesta seção, iremos agrupar algumas das principais técnicas de preservação de privacidade de dados de localização em dois modelos: modelos de anonimização e modelos de ofuscação.

1.8.1. Anonimização

As técnicas de anonimização em privacidade de localização buscam impedir ataques de ligação, ou seja, conforme discutido na Seção 1.3, buscam proteger a ligação entre a identidade do usuário e a informação de localização do mesmo [22], dificultando a re-identificação dos indivíduos. Em outras palavras, o objetivo é garantir que os dados de localização de um usuário, dentro de um conjunto de dados, não poderão ser ligados à identidade de seu dono. São exemplos de técnicas de anonimização: o k-Anonimato de localizações e as Zonas de Mixagem.

1.8.1.1. k-anonimato de Localizações

Conforme discutido na Seção 1.4.1, o k-anonimato foi proposto por Sweeney et al. [30] com o objetivo de prevenir ataques de ligação ao registro, alcançando a preservação de privacidade através de generalização ou supressão de dados. Em sua forma original, esse modelo assegura que, para cada combinação de k atributos semi-identificadores, existem pelo menos k registros distintos no conjunto de dados publicado, formando uma classe de equivalência.

Em privacidade de localização, um sujeito é tido k-anônimo se sua localização é indistinguível da localização de outros $k-1$ usuários. Portanto, a probabilidade de um usuário malicioso violar a privacidade de um indivíduo através de um ataque não poderá ser maior do que $\frac{1}{k}$.

Vale a pena lembrar que o parâmetro k do modelo é responsável por balancear a utilidade e a privacidade dos dados. Assim, quanto maior o valor de k, maior será a privacidade dos dados e, consequentemente, menor sua utilidade, o que pode ser um problema em serviços de localização, pois a precisão da localização afeta diretamente a qualidade do serviço. Desta forma, encontrar um equilíbrio entre privacidade e utilidade se faz ainda mais importante no contexto de LBS. Entretanto, encontrar um valor ótimo para o parâmetro k é um problema NP-difícil, como citado na Seção 1.4.1. Desta forma, os responsáveis pela anonimização devem especificar o grau de privacidade desejada em função desse parâmetro.

O conceito básico da aplicação do k-anonimato em dados de localização requer que o LBS seja operado por uma terceira entidade confiável, o anonimizador, responsável por anonimizar as localizações das requisições. Este anonimizador tem conhecimento da localização de todos os usuários que usam o serviço. Dessa forma, sempre que um usuário necessita realizar uma requisição, enviando sua localização, o anonimizador calcula um conjunto de k usuários e reporta uma área de ofuscação contendo k posições, incluindo a

localização da requisição do usuário.

A Figura 1.11 ilustra uma abordagem da utilização dessa técnica, para um $k = 3$, onde o usuário em laranja deseja enviar uma requisição ao LBS. Um terceiro confiável responsável por anonimizar a sua localização, agrupa o usuário e sua localização, com outras $k-1$ localizações, enviando uma requisição ao LBS contendo k localizações no total. Este, por sua vez, irá responder a requisição em função de cada uma das localizações enviadas. Como o terceiro confiável tem conhecimento das localizações dos usuários, ele irá filtrar a resposta referente a localização real presente na requisição, enviando-a ao usuário. Assim, para um atacante, a localização do usuário pode ser qualquer uma das k localizações que fazem parte da requisição, garantindo que a probabilidade de se identificar a localização do usuário não seja superior a $\frac{1}{k}$.

Figura 1.11. Processo de anonimização utilizando k-anonimato.

1.8.1.2. Zona de Mixagem

A zona de mixagem é outra abordagem que procura proteger a privacidade do usuário contra ataques de ligação, ao evitar que seja possível vincular a identidade dos usuários à sua localização. Entretanto, diferente do k-anonimato, a zona de mixagem pode ser aplicada sem qualquer informação de identidade do usuário. O conceito de zona de mixagem foi proposto por Beresford et al. [6]. Ele propõe um *framework*, onde os usuários utilizam pseudo-ids que são modificados constantemente garantindo que estes não sejam identificados no uso de serviços de localização. Sendo assim, a real identidade do usuário é protegida através do uso de pseudônimos.

As zonas de mixagens são definidas como áreas circulares de raio r, onde todo usuário dentro da zona possui um único pseudo-id, não registrado por nenhuma das aplicações cobertas por ela. Em outras palavras, esta técnica procura garantir a indistinguibilidade dos usuários no uso de qualquer das aplicações dentro da zona, através do uso de pseudo-ids e a não presença de qualquer informação de localização presente na requisição. Assim, como o k-anonimato, a técnica de zona de mixagem em sua forma clássica exige a figura de um terceiro confiável, responsável pela anonimização. Na zona de mixagem,

este terceiro confiável tem o papel de gerenciar os pseudo-ids dos usuários, garantindo que sempre que um usuário entre na zona de mixagem, ele possua um pseudo-id único que não tenha sido registrado por nenhuma aplicação. Desta forma, como não há a presença de localizações presentes na requisição, todas as aplicações que proveem serviços de localização dentro da zona veem os usuários em uma mesma localização definida pela zona de mixagem de alguma forma, como por exemplo um centroide que represente a zona.

A Figura 1.12 ilustra a aplicação da técnica de zona de mixagem. No primeiro quadro, o usuário ao entrar na zona de mixagem 1 obtém um pseudo-id único, não registrado por nenhuma aplicação, e só então passa a dispor de qualquer aplicação coberta pela zona de mixagem. Ao se deslocar para a zona de mixagem 2, no quadro 2, o usuário obtém um novo pseudo-id, também único e não registrado, e passa a utilizar dos serviços cobertos dentro dessa nova zona.

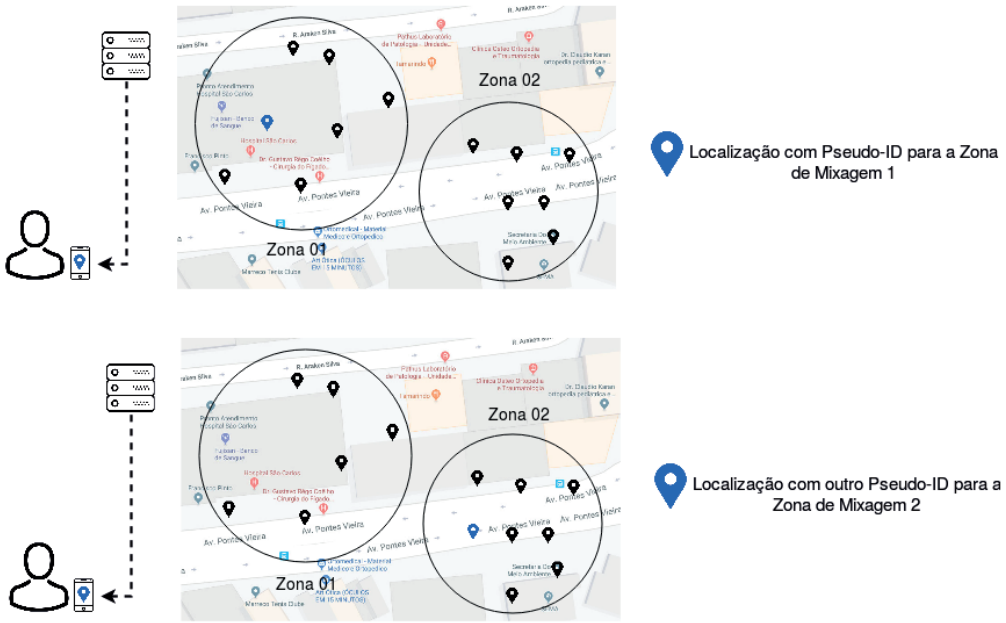

Figura 1.12. Processo de anonimização utilizando zona de mixagem.

1.8.2. Ofuscação

Diferente dos modelos de anonimização, os modelos de ofuscação atuam sobre os dados de localização em si, reduzindo sua precisão a fim de preservar a privacidade de localização dos usuários [22]. As principais técnicas de ofuscação são: técnica de Localizações falsas, Ofuscação de localização e Privacidade Diferencial (PD).

1.8.2.1. Localizações Falsas

Uma típica requisição feita a provedores de serviços de localização é composta de dois componentes básicos: *(i)* **informação de localização** é toda a informação contida na requisição referente a localização enviada na consulta. De forma geral e simplificada, é a

localização composta por suas coordenadas geográficas de latitude e longitude; *(ii)* **conteúdo de requisição** é a solicitação do serviço a ser prestado, por exemplo, a solicitação do hospital mais próximo dada a localização contida na informação de localização.

A técnica de localização falsas proposta por Kido et al. [18] procura garantir a privacidade dos dados de localização do usuário através de um processo de sanitização da *informação de localização* presente na requisição do usuário aos provedores de serviços de localização. Assim, numa requisição, a localização do usuário é acompanhada de múltiplas localizações falsas, cujo objetivo é mascarar a localização verdadeira do usuário.

Diferentemente dos modelos tradicionais de k-anonimato e zona de mixagem, a técnica de seleção de localizações falsas não necessita da presença de servidores confiáveis, diminuindo o risco de exposição. O processo de sanitização é de responsabilidade dos dispositivos móveis dos próprios usuários, que são responsáveis por selecionar, através de um mecanismo aleatório, as localizações falsas que irão compor a informação de localização da requisição de seus usuários. O objetivo é garantir que a probabilidade de se identificar a localização real dentre aquelas presentes na requisição não seja maior que $\frac{1}{k}$, onde k é o grau de privacidade desejado para uma requisição com uma quantidade de k localizações presentes.

Figura 1.13. Técnica de Localizações Falsas.

A Figura 1.13 ilustra a aplicação da técnica de Localizações falsas, onde o usuário, por intermédio de seu dispositivo, envia uma requisição anonimizada ao provedor de serviço. O processo de anonimização seleciona $k-1$ localizações, em preto, que serão enviadas na informação de localização da requisição, juntamente com a localização real do usuário, em laranja. O servidor receberá a requisição contendo k localizações, e responderá a solicitação contida no conteúdo de requisição, tendo como referência cada uma das localizações presentes na informação de localização. O dispositivo então filtra a resposta referente à localização real.

O mecanismo de seleção aleatória das localizações falsas é fundamental para garantir a privacidade de localização do usuário, uma vez que esta técnica, assim como as técnicas que abordam o modelo de anonimização estão sujeitas a ataques de conhecimento, que, conforme discutido na seção 1.7, utilizam de conhecimentos prévios para

inferir dados sensíveis dos usuários. Desta forma, várias abordagens foram propostas, a fim de solucionar este problema e garantir uma seleção de localizações falsas que minimize o risco de exposição dos dados de localização do usuário, utilizando para isto, do próprio conhecimento prévio disponível ao provedor de serviço [29, 25].

1.8.2.2. Ofuscação de Localizações

A técnica de ofuscação de localização procura garantir a preservação de privacidade do usuário através da redução deliberada da precisão da localização do mesmo. Em sua abordagem tradicional apresentada por Ardagna *et al.* [3, 4], o usuário ao realizar uma requisição ao LBS, ao invés de enviar suas coordenadas de localização real (x_u,y_u), envia uma área circular $Area(r,x_c,y_c)$, centralizada nas coordenadas geográficas (x_c,y_c) e raio r, onde a probabilidade de a localização do usuário estar presente dentro dessa área é igual a 1. A Figura 1.14 ilustra uma requisição anonimizada pela técnica de ofuscação, onde o usuário, com localização em laranja, ao enviar sua requisição, envia como informação de localização uma área circular de raio r, centrada nas coordenadas do ponto central em preto na figura, semelhante a todos os outros usuários presentes na área de ofuscação.

Figura 1.14. Técnica de Ofuscação de Localização usando área circular.

Outra abordagem desta técnica é proposta por Gutscher [17], onde a precisão da localização real é reduzida através de simples operações geométricas (i.e., rotação, translação) sobre suas coordenadas geográficas antes de serem enviadas na requisição ao provedor LBS. A Figura 1.15 ilustra a aplicação desta técnica através de uma simples operação de translação onde a localização enviada na requisição é substituída pela localização em laranja.

1.8.2.3. Privacidade Diferencial para Localizações

Como já explanado na Seção 1.5, o modelo de Privacidade Diferencial foi proposto com o objetivo de garantir a utilidade dos dados, ao mesmo tempo que fornece proteção contra ataques municiados por conhecimento adversário.

No contexto de dados de localização diversas abordagens que utilizam privacidade diferencial foram propostas. Os modelos propostos em [2, 13] procuram estender a no-

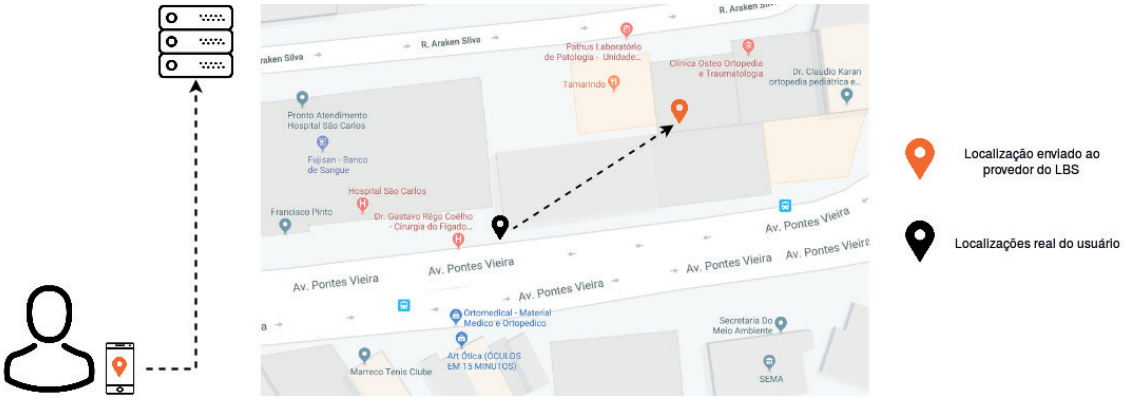

Figura 1.15. Técnica de Ofuscação de Localização usando operações geométricas.

ção utilizada na abordagem original de Privacidade Diferencial sobre conjuntos de dados vizinhos para o contexto de localização. Assim, duas localizações são ditas vizinhas se a distância física entre elas é menor ou igual a um raio r. Desta forma, é definida uma área de raio $r > 0$, onde supostamente a localização do usuário dentro desta área esta protegida. O nível de privacidade, portanto, depende diretamente de r e é alcançado através da adição de um ruído controlado à localização do usuário. Ou seja, um r grande implica em uma maior proteção à privacidade do usuário, entretanto, a adição do ruído tende a ser maior, diminuindo a utilidade dos dados e consequentemente, a qualidade do serviço. Já um r pequeno garante uma maior qualidade do serviço já que a adição do ruído é menor, entretanto, a proximidade das localizações pode afetar a garantia de privacidade do usuário. A Figura 1.16 apresenta em azul um conjunto de localizações vizinhas da localização l em preto, e em vermelho localizações que estão a uma distância da localização l maior que r, portanto, não vizinhas de l.

Figura 1.16. Localizações vizinhas.

Podemos, agora, definir (r,ε)-privacidade de localização para localizações vizinhas da mesma forma que definido no modelo padrão de privacidade diferencial para conjuntos de dados vizinhos. Assim, um mecanismo responsável pela adição de ruído satisfaz (r,ε)-privacidade de localização se quaisquer duas localizações dentro do raio r são indistinguíveis quando observadas as saídas do mecanismo K para estas localizações.

Definição 2 *(r,ε)-privacidade de localização: Para um raio $r > 0$ e $\varepsilon > 0$, um mecanismo*

$K : X \rightarrow E^2$ satisfaz (r,ε)-privacidade de localização), se e somente se, para todo i,j $\in X$ com d(i,j)$\leq$ r,

$$P(K(i) \in S) \leq exp^{\varepsilon} P(K(j) \in S) \quad \forall S \subseteq E^2$$

De acordo com a definição, a probabilidade de a localização retornada pelo mecanismo K aplicado sobre a localização i é semelhante a probabilidade de a localização retornada pelo mecanismo K quando aplicado por uma localização vizinha j, limitada pela exponencial de ε.

A Figura 1.17 ilustra a aplicação do processo de anonimização garantindo (r, ε)-*privacidade de localização*. O Mecanismo irá adicionar um ruído controlado às coordenadas da localização real do usuário. Este processo irá garantir uma nova localização anonimizada à uma distância de no máximo r da localização real.

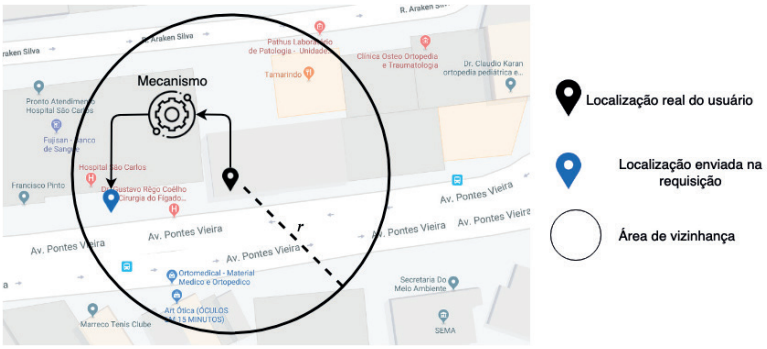

Figura 1.17. Anonimização por $(r-\varepsilon)$-privacidade de localização.

1.9. Conclusão

Este capítulo conclui que a preservação da privacidade de dados acerca de indivíduos é um problema desafiador. Técnicas de anonimização têm sido utilizadas para a disponibilização de dados sensíveis, procurando encontrar o melhor balanceamento entre privacidade e utilidade que atenda às diversas partes envolvidas no processo de disponibilização de dados. Diferentes tipos de ataques à privacidade têm sido empregados por usuários maliciosos com a intenção de violar informações sensíveis de bases de dados abertas. Para tal fim, os atacantes utilizam conhecimento que muitas vezes é imensurável, devido aos diversos cenários em que informações podem ser obtidas. No contexto de dados de localização, este risco se potencializa, em virtude das informações agregadas ao dado geográfico buscado quando de uma solicitação a um serviço de localização, que servem de munição para os agentes maliciosos. Este capítulo apresentou as principais técnicas no estado da arte em preservação de privacidade de dados de localização. Os modelos de anonimização buscam proteger contra ataques de ligação ao registro, ou seja, prevenir a vinculação entre a identidade do usuário e sua localização, evitando a re-identificação de indivíduos, geralmente utilizando técnicas de supressão e generalização. Os modelos de ofuscação, por sua vez, buscam proteger a localização em si, garantindo que esta não seja revelada, mesmo no uso de serviços de localização. A Privacidade Diferencial se destaca por fornecer soluções de preservação de privacidade, onde um ruído aleatório controlado

é adicionado a localização do usuário, garantindo que a localização real do usuário estará protegida independentemente do conhecimento do atacante.

Finalmente, entendemos que o problema da garantia de privacidade de dados de localização dos usuários de LBS continua cientificamente relevante. A busca por um ponto ideal na curva de solução de compromisso entre privacidade do indivíduo e a utilidade do dado fornecido para esse tipo de serviço deve pautar os próximos passos da pesquisa. Este aspecto é particularmente importante no contexto de LBS pois a qualidade do serviço é dependente da precisão do dado de localização, portanto o envio de dado perturbado para o provedor de serviço tende a impactar negativamente na qualidade. Tanto o paradigma de anonimização sintática, quanto o modelo de Privacidade Diferencial apresentam aspectos de revisão que devem ser vistos como oportunidades de pesquisas e desenvolvimento. Avanços em ambos os paradigmas são necessários para garantir que o futuro ofereça cada vez mais proteção à privacidade de indivíduos e ao mesmo tempo haja dados úteis e disponíveis para pesquisadores, testadores e analistas de dados.

Agradecimentos

Esta trabalho foi parcialmente financiada pela Lenovo, como parte do seu investimento em pesquisa e desenvolvimento de acordo com a Lei de Informática, pela CAPES (1836136), CNPq (122201/2018-3) e pelo LSBD/UFC.

Referências

[1] Aggarwal, C. C. and Philip, S. Y. (2008). A framework for condensation-based anonymization of string data. *Data Mining and Knowledge Discovery*, 16(3):251–275.

[2] Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., and Palamidessi, C. (2013). Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, pages 901–914.

[3] Ardagna, C. A., Cremonini, M., Damiani, E., De Capitani di Vimercati, S., and Samarati, P. (2007). Location privacy protection through obfuscation-based techniques. In Barker, S. and Ahn, G.-J., editors, *Data and Applications Security XXI*, pages 47–60.

[4] Ardagna, C. A., Cremonini, M., De Capitani di Vimercati, S., and Samarati, P. (2011). An obfuscation-based approach for protecting location privacy. *IEEE Transactions on Dependable and Secure Computing*, 8(1):13–27.

[5] Bayardo, R. J. and Agrawal, R. (2005). Data privacy through optimal k-anonymization. In *21st International conference on data engineering (ICDE'05)*, pages 217–228.

[6] Beresford, A. R. and Stajano, F. (2003). Location privacy in pervasive computing. *IEEE Pervasive computing*, (1):46–55.

[7] Blumberg, A. J. and Eckersley, P. (2009). On locational privacy, and how to avoid losing it forever. *Electronic frontier foundation*, 10(11).

[8] Brito, F. T. and Machado, J. C. (2017). Preservação de privacidade de dados: Fundamentos, técnicas e aplicações. In Delicato, F. C., Pires, P. F., and Silveira, I. F., editors, *Jornadas de A tualização em Informática 2017*. Sociedade Brasileira de Computação - SBC.

[9] Dewri, R., Ray, I., Ray, I., and Whitley, D. (2008). On the optimal selection of k in the k-anonymity problem. In *24th ICDE International Conference on Data Engineering*, pages 1364–1366, Cancun,Mexico.

[10] Domingo-Ferrer, J. and Torra, V. (2001). A quantitative comparison of disclosure control methods for microdata. *Confidentiality, disclosure and data access: theory and practical applications for statistical agencies*, pages 111–134.

[11] Dwork, C. (2006). Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming*, pages 1–12, Venice, Italy.

[12] Dwork, C (2008). *Differential privacy: a survey of results. InInternational conference on theory and applications of models of computation (pp. 1–19).*

[13] Elsalamouny, E. and Gambs, S. (2016). Differential Privacy Models for Location-Based Services. *Transactions on Data Privacy*, 9(1):15 – 48.

[14] Fung, B. C., Wang, K., Fu, A. W.-C., and Yu, P. S. (2010a). *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Chapman & Hall/CRC, 1st edition. ISBN 978-1-4200-9148-9.

[15] Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S. (2010b). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4):1–53.

[16] Goldreich, O. (2003). Cryptography and cryptographic protocols. *Distributed Computing*, 16(2-3):177–199.

[17] Gutscher, A. (2006). Coordinate transformation - a solution for the privacy problem of location based services? In *Proceedings 20th IEEE International Parallel Distributed Processing Symposium*, pages 7 pp.–.

[18] Kido, H., Yanagisawa, Y., and Satoh, T. (2005). An anonymous communication technique using dummies for location-based services. In *Proceedings of the Int. Conf. on Pervasive Services, ICPS'05*, pages 88–97. IEEE.

[19] LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. (2005). Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 49–60. ACM.

[20] Li, H., Sun, L., Zhu, H., Lu, X., and Cheng, X. (2014). Achieving privacy preservation in wifi fingerprint-based localization. In *INFOCOM, 2014 Proceedings IEEE*, pages 2337–2345. IEEE.

[21] Li, N., Li, T., and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE.

[22] Liu, B., Zhou, W., Zhu, T., Gao, L., and Xiang, Y. (2018). Location privacy and its applications: A systematic study. *IEEE Access*, 6:17606–17624.

[23] Meyerson, A. and Williams, R. (2004). On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228. ACM.

[24] Nergiz, M. E., Atzori, M., and Clifton, C. (2007). Hiding the presence of individuals from shared databases. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 665–676, New York, NY, USA. ACM.

[25] Neto, E. R. D., Mendonça, A. L. C., Brito, F. T., and Machado, J. C. (2018). Privlbs: uma abordagem para preservação de privacidade de dados em serviços baseados em localização. In *Brazilian Symposium on Databases SBBD*, Rio de janeiro, Brazil.

[26] Primault, V., Boutet, A., Mokhtar, S. B., and Brunie, L. (2018). The long road to computational location privacy: A survey. *IEEE Communications Surveys & Tutorials*.

[27] Schiller, J. and Voisard, A. (2004). *Location-based services*. Elsevier.

[28] Shokri, R., Theodorakopoulos, G., Le Boudec, J.-Y., and Hubaux, J.-P. (2011). Quantifying location privacy. In *2011 IEEE symposium on security and privacy*, pages 247–262. IEEE.

[29] Sun, G., Chang, V., Ramachandran, M., Sun, Z., Li, G., Yu, H., and Liao, D. (2017). Efficient location privacy algorithm for internet of things (iot) services and applications. *Journal of Network and Computer Applications*, 89:3–13.

[30] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

[31] Tan, V. Y. F. and Ng, S.-K. (2007). Generic probability density function reconstruction for randomization in privacy-preserving data mining. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 76–90. Springer.

[32] Truta, T. M., Campan, A., and Meyer, P. (2007). Generating microdata with p-sensitive k-anonymity property. In *Workshop on Secure Data Management*, pages 124–141. Springer.

[33] Wang, K., Fung, B. C., and Yu, P. S. (2005). Template-based privacy preservation in classification problems. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE.

[34] Wong, R. C.-W. and Fu, A. W.-C. (2010). Privacy-preserving data publishing: An overview. *Synthesis Lectures on Data Management*, 2(1):1–138.

[35] Zhu, X., Chi, H., Niu, B., Zhang, W., Li, Z., and Li, H. (2013). Mobicache: When k-anonymity meets cache. In *Global Communications Conference (GLOBECOM), 2013 IEEE*, pages 820–825, Atlanta, GA, USA. IEEE.

Capítulo 1

Uma Introdução ao Combate Automático às *Fake News* em Redes Sociais Virtuais

Paulo Márcio Souza Freire, Ronaldo Ribeiro Goldschmidt

Abstract

Combating Fake News (i.e., false news intentionally spread) is not a recent problem. However, its complexity has increased mainly due to the growth of volume and speed of news dissemination provided by the virtual social networks. In this scenario, computational approaches are becoming essential devices to combat this type of news. Thus, this Chapter presents a conceptual and practical introduction to the main computational approaches to combat Fake News, besides some comments on related areas and recent research on this theme.

Resumo

O problema de combater Fake News (ie., notícias falsas veículadas de forma intencional) não é recente. Contudo, sua complexidade vem aumentando em função do crescimento do volume e da velocidade de divulgação de notícias proporcionado pelas redes sociais virtuais. Diante deste cenário, abordagens computacionais que possam auxiliar no combate automático deste tipo de notícia estão se tornando cada vez mais necessárias. Assim sendo, o presente Capítulo apresenta uma introdução conceitual e prática às principais abordagens computacionais de combate às Fake News, além de comentar sobre áreas e pesquisas recentes relacionadas a este tema.

1.1. Considerações Iniciais

Historicamente, a publicação de notícias estava restrita à mídia tradicional, como rádio, TV, jornais e revistas impressos. Com o surgimento das redes sociais virtuais de fácil acesso e baixo custo (também conhecidas como, simplesmente, redes sociais), as pessoas vêm, a cada dia, aumentando o consumo de notícias *on-line*, em vez daquelas fornecidas pelos canais tradicionais [Vosoughi et al. 2017].

Apesar de seus benefícios, as redes sociais permitem que qualquer pessoa, independentemente de sua credibilidade, divulgue (publique/propague) notícias com intenso poder de espalhamento [Shu et al. 2017a][Wang et al. 2018a]. Portanto, as redes sociais amplificaram um problema antigo: a disseminação de notícias falsas [Conroy et al. 2015] [Zhang et al. 2018]. Este problema abrange uma questão ainda mais difícil: *Fake News* que é a divulgação de uma notícia falsa de forma intencional [Shu et al. 2017a]. Este espalhamento de notícias propositalmente falsas costuma ser prejudicial, pois uma inverdade deliberada tende a ser melhor elaborada e, portanto, mais eficaz em seu objetivo principal, que é influenciar na mudança de opinião. A proliferação de *Fake News*, geralmente, afeta não apenas a integridade jornalística, mas também perturba as áreas social, política, econômica, cultural e da segurança [Wang 2017][Mustafaraj and Metaxas 2017].

Como materialização do poder de influência deste tipo de notícia, pode-se destacar que, somente nos Estados Unidos da América (EUA), mais de sessenta e dois porcento dos adultos recorrem às redes sociais para receberem notícias. Como consequência deste elevado percentual, alguns casos relevantes, ocorridos em 2016, podem ser destacados [Farajtabar et al. 2017]:

• Nos três meses finais das eleições presidenciais, as notícias falsas publicadas no *Facebook*, que favoreceram qualquer um dos dois candidatos, foram compartilhadas 37 milhões de vezes;

• Uma análise do *Buzzfeed News* [1] mostra que, a partir das 20 principais notícias falsas sobre as eleições, criadas por sites fraudulentos, foram geradas quase 1,5 milhões de atividades de engajamento de usuários no *Facebook*;

• Um homem, carregando um rifle AR-15, aterrorizou os frequentadores de uma pizzaria na capital *Washington*, porque ele havia lido uma notícia falsa *on-line*, afirmando que o referido estabelecimento usava crianças jovens como escravas sexuais.

Inclusive, casos relacionados às *Fake News* não se limitam aos EUA. Em 2018, na Índia, após notícias falsas terem, supostamente, levado a linchamentos, o *WhatsApp* anunciou um limitador para a quantidade de encaminhamentos de mensagem [2]. Portanto, há um apelo urgente para desenvolver estratégias efetivas para mitigar o impacto deste tipo de notícia falsa.

Nos últimos anos, tanto a academia quanto a indústria estudam como combater *Fake News* nas redes sociais [Flintham et al. 2018] [Wang et al. 2018a] [Zhou et al. 2019] [Campan et al. 2017] [Kshetri and Voas 2017]. Este combate apresenta-se como não trivial, tanto pelo volume de publicações quanto pela velocidade das suas respectivas propagações. Assim, o emprego de apordagens computacionais, devido à sua maior velocidade de atuação, vem se destacando no combate às *Fake News* nas redes sociais [Ruchansky et al. 2017].

Baseado nesta necessidade computacional, o presente Capítulo provê uma introdução ao referido combate através da seguinte estrutura: A Seção 1.2 apresenta diferentes definições para o termo *Fake News*, assim como aborda o comportamento disseminativo deste tipo de notícia nas redes sociais. Um levantamento sobre os trabalhos relacionados

[1] https://www.buzzfeed.com/

[2] BBC News Brasil - https://www.bbc.com

é realizado na Seção 1.3. A Seção 1.4, por sua vez, realiza um estudo de caso onde aplica detecção automática baseada na reputação do usuário via *Crowd Signals*. Por fim, na Seção 1.5, são abordados os problemas em aberto.

1.2. Fundamentos

Como a utilização do termo *Fake News* é relativamente recente, a sua caracterização se faz necessária. Para tal, são agrupadas as diferentes definições para *Fake News*, assim como são categorizadas as razões que levam ao seu comportamento disseminativo nas redes sociais.

1.2.1. Definição de *Fake News*

Apesar da originalidade da expressão, as *Fake News* não surgiram com o uso das redes sociais. Haja vista que, mesmo com as mídias tradicionais, já existiam pessoas que, por diferentes razões, divulgavam notícias falsas de forma proposital [Golbeck et al. 2018]. Independente do surgimento, devido à contemporaneidade do termo, *Fake News* apresenta diversas definições que podem ser organizadas em dois grupos.

O primeiro grupo considera que o aspecto proposital é fundamental, pois define as *Fake News* como publicações intencionalmente e verificadamente falsas [Shu et al. 2017a] [Mustafaraj and Metaxas 2017] [Reis et al. 2019] [Zhou et al. 2019] [Campan et al. 2017] [Flintham et al. 2018] [Wang et al. 2018a] [Zhou and Zafarani 2018] [Conroy et al. 2015]. Para enfatizar a diferença entre uma notícia falsa e uma intencionalmente falsa, pode-se utilizar dois termos denominados *misinformation* e *disinformation* [Golbeck et al. 2018] [Campan et al. 2017]. Enquanto *misinformation* corresponde às notícias falsas publicadas pela falta da informação verdadeira, a *disinformation* diz respeito às notícias falsas divulgadas com algum propósito. Com base nestas correspondências, é possível caracterizar *Fake News* como sendo uma *disinformation* [Kshetri and Voas 2017]. Cabe ressaltar que, apesar de pertencente ao primeiro grupo, o trabalho [Zhou and Zafarani 2018] é ainda mais específico em sua definição, pois só considera *Fake News* quando a notícia intencionalmente falsa é divulgada por uma agência de notícias. Ademais, ainda de acordo com este primeiro grupo, existem outras áreas que, apesar de não abordarem a questão do combate às *Fake News*, apresentam relação com as notícias intencionalmente falsas. Algumas destas áreas se encontram descritas abaixo:

• Classificação de Rumores (*Rumor Classification*) - Rumor é uma informação em circulação cuja veracidade não foi verificada no momento da publicação. Um rumor pode ser classificado como verdadeiro, falso ou ainda não verificado [Shu et al. 2017a] [Liu and Xu 2016] [Vosoughi et al. 2017] [Ma et al. 2015]. Portanto, uma *Notícia* não verificada antes da publicação é um *Rumor*, que pode ser caracterizado como *Fake News* a partir do momento que seja identificado como falso e intencional. A tarefa mais relacionada com o combate às *Fake News* é a classificação da veracidade dos rumores;

• Descoberta da Verdade (*Truth Discovery*) - é a descoberta da verdade de fatos conflitantes entre diferentes fontes [Shu et al. 2017a] [Li et al. 2015]. Assim, uma mesma *Notícia* pode conter afirmações diferentes (distintas opiniões), onde as intencionalmente falsas podem ser caracterizadas como *Fake News*. Assim, o combate às *Fake News* pode se beneficiar da Descoberta da Verdade para determinar a veracidade das afirmações;

• Detecção de Iscas de Cliques (*Clickbait Detection*) – procura identificar, nas páginas *Web*, as chamadas iscas de cliques que, praticamente, forçam o usuário a selecionar a opção apresentada. Neste caso, o corpo do texto (*bodytext*) dos artigos é, frequentemente, pobre em relação ao seu cabeçalho (*headlines*). Esta discrepância pode ser encontrada não só em *Clickbait*, como também em *Fake News*. Sendo assim, o *Clickbait* pode ser usado como um indicador de *Fake News* [Shu et al. 2017a];

• Detecção de Bots (*Bot Detection*) – procura identificar o envio automático de informações nas redes sociais por meio de robôs [Braz and Goldschmidt 2017]. Estes envios podem potencializar tanto a publicação quanto a respectiva propagação da *Fake News* [Wang et al. 2018a] [Nasim et al. 2018] [Ferrara et al. 2016];

• Checagem de fatos (*Fact Checking*) - são *Websites* ou *Frameworks* responsáveis pela verificação, normalmente realizada com a ajuda de especialistas, da veracidade de fatos divulgados em redes sociais [Ciampaglia et al. 2015] [Ruchansky et al. 2017] [Sethi 2017] [Vo and Lee 2018]. Inclusive, existem abordagens voltadas para a seleção automática de notícias a serem enviadas para a referida checagem [Kim et al. 2018] [Tschiatschek et al. 2018]. A verificação da verdade dos fatos pode ser utilizada na tarefa de detecção de *FakeNews* [Cazalens et al. 2018], assim como na criação de *datasets*;

• Sistemas de Reputação (*Reputation System*) - são sistemas que buscam determinar o nível de confiança em redes sociais baseados na obtenção de graus de reputação [Vavilis et al. 2014] [Hendrikx et al. 2015] [Seo J. 2013] [Deng et al. 2014] [Sherchan et al. 2013]. A determinação de graus de reputação dos usuários pode ser utilizada na tarefa de identificação das *Fake News*.

O segundo grupo, entretanto, tem uma definição mais genérica. Para este segmento, as *Fake News* são todas as notícias falsas, independente da sua natureza intencional [Sharma et al. 2019] [Castelo et al. 2019] [Ajao et al. 2019]. Inclusive, consideram-se como *Fake News* outros tipos de notícia, como, por exemplo, Rumor.

Este Capítulo adota a definição do primeiro grupo, consequentemente considera *Fake News* como sendo, somente, uma notícia intencionalmente falsa. A principal razão para esta escolha é que uma notícia propositalmente divulgada tende a ser mais bem elaborada, podendo, assim, causar mais malefícios aos usuários das redes sociais.

1.2.2. Comportamento disseminativo das *Fake News*

A disseminação e, consequente, divulgação de uma notícia se inicia pela sua publicação e provável propagação na rede social (Efeito de Câmara de Eco) [Shu et al. 2017a]. Desta forma, é importante destacar o momento no qual uma notícia pode ser caracterizada como *Fake News*. Basicamente, uma notícia intencionalmente falsa pode surgir de duas formas. A primeira é quando a *Fake News* é iniciada na rede social por meio da sua publicação e, posteriormente, potencializada pela sua possível propagação. A segunda é quando uma notícia não *fake* é publicada, porém se tornar *fake*, a partir do seu espalhamento, de acordo com as contribuições intencionalmente falsas feitas durante a sua propagação.

Independente do momento de criação, a recente proliferação de notícias falsas e mal-intencionadas nas redes sociais tem sido uma fonte de preocupação generalizada. Esta apreensão se deve pelo seu poder de espalhamento e, consequente, influência na

sociedade [Flintham et al. 2018]. As razões que potencializam a divulgação das *Fake News* nas redes sociais podem ser divididas em quatro categorias. A primeira tem relação com poder de influência ocasionado pelos fatores inerentes ao ser humano, dentre eles podemos destacar que as pessoas [Shu et al. 2017a]:

• Preferem receber informações que confirmem as suas opiniões sem, necessariamente, verificarem a veracidade da notícia;

• Tendem a aceitar as informações não pela análise da verdade, mas pela relação de ganhos e perdas que a notícia pode trazer para elas;

• Tendem a avaliar as informações não pela busca da veracidade, pois acabam acompanhando a aceitação dos outros.

A segunda categoria é a carência de legislação punitiva, sendo uma das alegações para tal fato é que as referidas leis poderiam cercear a liberdade de expressão. A terceira categoria está vinculada ao potencial ganho financeiro com a divulgação de determinadas notícias [Kshetri and Voas 2017] nas redes sociais. Já a quarta categoria advém da facilidade de criação de contas nas redes sociais [Conroy et al. 2015]. Um aspecto importante inerente à esta facilidade é a criação de contas digitais maliciosas por meio de divulgadores de natureza humana e/ou computacional [Shu et al. 2017a]. Estes divulgadores subdividem-se em:

• *Bot* - robôs responsáveis por divulgar *Fake News*;

• Humano - pessoas (*trolls*) intencionadas em disseminar *Fake News*;

• *Cyborg* - mecanismos híbridos (Humano/*Bot*) que divulgam *Fake News*.

Ainda se tratando da facilidade de divulgação de notícias intencionalmente falsas nas redes sociais, uma das formas mais simples de criar uma *Fake News* é se infiltrar em uma comunidade de pessoas engajadas em discutir um determinado assunto. Portanto, segundo [Mustafaraj and Metaxas 2017], devem ser realizados os seguintes passos: Criar um domínio falso (*website*), criar contas anônimas, identificar comunidades e usuários interessados em um determinado assunto, contaminar estes usuários com a notícia falsa e, finalmente, incentivar a discussão para que a *Fake News* seja espalhada.

1.3. Trabalhos Relacionados

Para apresentar os trabalhos vinculados ao combate automático às *Fake News* nas redes sociais é proposto e, em seguida, aplicado um modelo comparativo que viabilize uma distinção entre abordagens computacionais. Desta forma, as abordagens, juntamente com o seu respectivo *dataset*, são enquadrados no citado modelo.

1.3.1. Proposta de Modelo Comparativo

O combate às *Fake News* em redes sociais, por meio de abordagens computacionais, possui uma variedade de aspectos que podem ser considerados. Com o objetivo de facilitar a comparação e a consequente classificação das referidas abordagens, tais aspectos são categorizados na Figura 1.1. As próximas subseções detalham cada um destes aspectos.

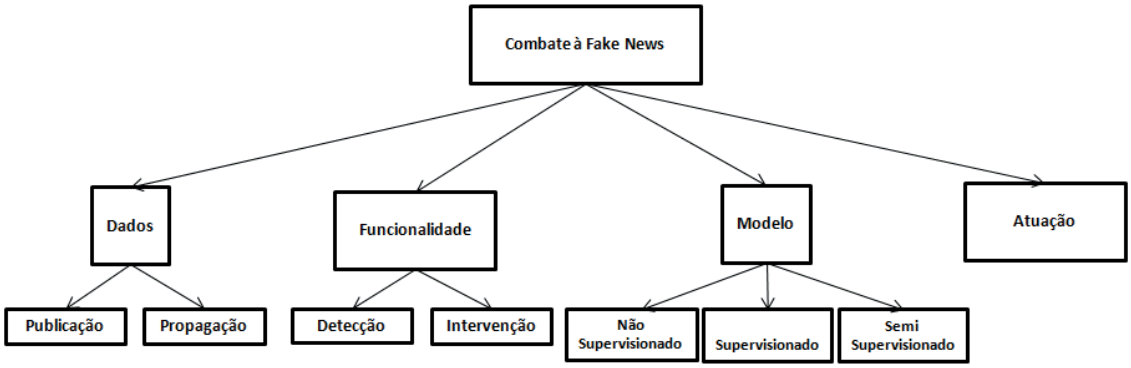

Figura 1.1: Aspectos considerados em Abordagens de Combate Automático às *Fake News*

1.3.1.1. Dados

Aspecto relacionado aos dados que podem ser utilizados pelas abordagens computacionais de combate às *Fake News*. Este aspecto subdivide-se em dados obtidos a partir da *Publicação* da notícia, como também aqueles associados com a sua *Propagação*.

Os dados de *Publicação* representam as informações inerentes ao surgimento da notícia na rede social. Estes dados podem ser classificados em *Notícia, Usuário, Assunto e Temporalidade*. No que diz respeito à Notícia, a abordagem pode ser capaz de analisar dados oriundos da publicação a partir de diferentes tipos de *Mídia* (*Texto, Áudio e Imagem*). Independente da *Mídia*, a análise do *Conteúdo* pode ser realizada de forma *Léxica, Sintática, Semântica e Legibilidade*. Com relação ao *Usuário* publicador, a abordagem pode identificar diferentes *Tipos*, tais como: humano, *bot* ou *cyborg*. Pode-se analisar também dados referentes ao *Perfil* do usuário na rede social, tais como: identificação e idade. Outro aspecto relevante está relacionado à *Reputação* do publicador, que pode estar vinculada à sua capacidade em identificar ou publicar *Fake News*. A abordagem pode também utilizar o *Assunto* abordado no momento da publicação. Assim, é possível tratar Especificidades, tais como: relacionamento entre assuntos, assuntos controversos ou análise de tópicos. Outro aspecto leva em consideração a *Relevância* do assunto publicado, haja vista que assuntos em voga motivam a criação de *Fake News*. A variação das características de uma notícia com o passar do tempo, torna a *Temporalidade* mais um relevante recurso para a identificação de *Fake News*.

Os dados de *Propagação* representam as informações obtidas após a publicação, consequentemente, aquelas inerentes às contribuições devido ao espalhamento da notícia na rede social (ex: curtida/*like*, comentário/*reply* ou compartilhar/*retweet*). Portanto, estes dados podem ser classificados em *Contribuição, Usuário, Assunto, Temporalidade e Rede*. No que diz respeito à *Contribuição*, *Usuário*, *Assunto* e *Temporalidade* a abordagem pode ser capaz de analisar os dados oriundos da *Propagação*, a partir dos mesmos aspectos anteriormente citados na *Publicação*. Ademais, as informações relacionadas à *Rede* criada, a partir da propagação da notícia, possibilitam não só a identificação de uma *Fake News* como uma possível atuação contra a mesma.

1.3.1.2. Funcionalidade

Além dos dados coletados, as abordagens automáticas de combate às *Fake News* podem, basicamente, possuir duas funcionalidades: *Detecção e Intervenção*.

A *Detecção* automática da *Fake News* pode ser, basicamente, um problema de classificação binária onde dada uma rede social $\mathscr{G}$, uma notícia a e um conjunto de postagens (publicações/propagações) $\mathscr{P}$, relacionadas à a, são espalhadas através da $\mathscr{G}$ por um conjunto de usuários U em um intervalo de tempo t. Assim o referido classificador binário $\mathscr{F}$ deve, aprendendo a partir dos dados, prever se a é uma *fake news* ou não, como formalmente indicado na equação 1. Uma outra forma é a utilização de técnicas mais subjetivas que definam a probabilidade, peso ou pertinência de uma notícia a ser *fake*.

$$\mathscr{F}(\mathscr{G},a,\mathscr{P},U,t) = \begin{cases} 1, & \text{se } a \text{ é uma } \textit{fake news}; \\ 0, & \text{caso contrário.} \end{cases} \tag{1}$$

Independente da forma, para que uma notícia a possa ser detectada como *Fake News* é necessária a realização de duas subfuncionalidades: *Autenticidade e Intencionalidade* [Janze and Risius 2017] [Vosoughi et al. 2017]. A *Autenticidade* analisa se a notícia é verdadeira ou falsa, enquanto que a *Intencionalidade* busca determinar a intenção dos divulgadores em ludibriar os receptores. Esta *Intencionalidade* pode ser mensurada como pontuação, peso ou score e obtida, por exemplo, por intermédio da análise de sentimentos que a notícia disponibiliza, pela associação entre usuários, assim como pelas características de perfil, tipo e reputação (credibilidade/confiança) dos divulgadores.

Já a *Intervenção* automática procura atacar as *Fake News*, nas redes sociais, de forma proativa ou reativa [Shu et al. 2017a][Farajtabar et al. 2017]. A intervenção reativa busca combater os efeitos da notícia a partir do momento da sua detecção como notícia propositalmente falsa. Por outro lado, a intervenção proativa tenta atuar antes mesmo da referida detecção, agindo então como uma forma de prevenção. Além disso, a tarefa de intervenção pode ser dividida em dois segmentos: *o Bloqueio e a Mitigação*. O *Bloqueio* atua de forma reativa. Na sua forma mais branda, o bloqueio interrompe a propagação da notícia e/ou a atuação do(s) usuário(s) responsáveis. Uma outra forma mais incisiva seria remover a(s) notícia(s) e/ou o(s) usuário(s) divulgador(es). Já a *Mitigação* pode agir de forma reativa ou proativa buscando enfraquecer as consequências causadas pela *Fake News*. Na reatividade, a mitigação pode, por exemplo, imunizar os usuários provendo notícias verdadeiras [Farajtabar et al. 2017]. Uma forma de proatividade na mitigação é prover alertas, mesmo que a notícia ainda não tenha sido detectada com propositalmente falsa. Estes alertas podem estar relacionados com o nível de reputação da fonte (usuário) ou sobre o assunto estar relacionado com outras *Fake News* já identificadas.

Independente da funcionalidade da abordagem, a coleta dos dados inerentes à divulgação da notícia se faz necessária para subsidiar a detecção e a intervenção das notícias intencionalmente falsas. Assim, tanto a coleta de dados na rede social quanto as tarefas de detecção e intervenção são fases iterativas, conforme ilustra a Figura 1.2. Cabe ressaltar que quanto mais cedo acontecer a detecção e a intervenção da *Fake News*, os impactos negativos desta notícia tendem a ser menores.

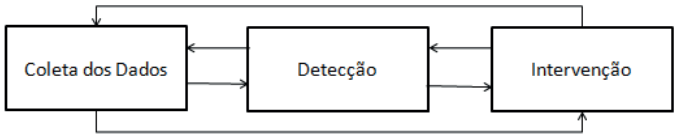

Figura 1.2: Fluxo do processo de combate às Fake News

1.3.1.3. Modelo

Quando a solução é por aprendizado de máquina, pode-se utilizar modelos computacionais para, a partir dos dados coletados, detectar as *Fake News*. Estes modelos são categorizados em *Não Supervisionado, Semi-Supervisionado e Supervisionado*.

No modelo *Não Supervisionado* são categorizadas as técnicas que normalmente levam mais tempo para realizar a identificação, porém, como não necessitam de rótulos, podem utilizar *datasets* mais simples [Shu et al. 2017a].

Os modelos *Supervisionados* são lentos na fase do treinamento, entretanto, tendem a ser mais rápidos do que os não supervisionados no momento da sua utilização na identificação das *Fake News*. Devido à necessidade de treinamento, os modelos supervisionados precisam de *datasets* mais completos [Shu et al. 2017a].

O modelo *Semi-Supervisionado* procura realizar a tarefa de identificação da *Fake News*, em redes sociais, de uma forma híbrida que busque utilizar, tanto as técnicas supervisionadas quanto não supervisionadas. Esta abordagem pode utilizar *datasets* mais simples do que aqueles manipulados pelos modelos supervisionados, porém mais complexos do que os utilizados pelos não supervisionados [Shu et al. 2017a].

1.3.1.4. Atuação

As abordagens computacionais que visam o combate às *Fake News*, independemente dos dados coletados, funcionalidade e modelo utilizados, podem ter diferentes formas de atuação.

Uma das possibilidades de atuação está associada à localização física do combate dentro da rede social. Uma abordagem *Centralizada* encontra-se, fisicamente, em um único ponto da rede social. Portanto, todas as tarefas relacionadas com a detecção/intervenção da *Fake News* são executadas em um mesmo local.

Por outro lado, uma abordagem *Descentralizada* encontra-se, fisicamente, espalhada na rede social. Assim, esta forma de atuação possibilita, inclusive, uma execução paralela e/ou ditribuida [Wu and Liu 2018] no combate às *Fake News*.

1.3.2. Revisão dos Trabalhos Relacionados

Nesta Subseção são apresentados alguns trabalhos relacionados ao combate automático às *Fake News* em redes sociais. Para tal, foram realizadas buscas onde as principais fontes de

consulta foram os artigos [Zhou and Zafarani 2019] [Reis et al. 2019] [Shu et al. 2017a] [Conroy et al. 2015] [Zhou et al. 2019] [Zhou and Zafarani 2018] [Sharma et al. 2019].

Para um melhor entendimento, os citados trabalhos são identificados e enquadrados no Modelo Comparativo tratado na Subseção 1.3.1, conforme mostram as Tabelas 1.1, 1.2 e 1.3. Cabe ressaltar que, nestas três Tabelas, as células não preenchidas indicam a não utilização do respectivo aspecto no trabalho correspondente.

Além disso, os referidos trabalhos são brevemente descritos, podendo seus detalhes serem consultados através das respectivas referências:

T1) *A Topic-Agnostic Approach for Identifying Fake News Pages* [Castelo et al. 2019]: O trabalho propõe um *topic-agnostic* (TAG) classificador que usa dados linquísticos e *Web-Markup* (padrões de layout das páginas) para detectar Fake News. Assim, ao invés de usar o *bag of words*, o trabalho explora as topic-agnostic, incluindo características morfológicas, psicológicas e de legibilidade que são comuns em *Fake News*. O trabalho propõe que páginas com *Fake News* normalmente têm inclinação sensacionalista, assim como a ocorrência de termos, tais como: *"Just in" e "Read this"*. Foram utilizados 3 classificadores *Support Vector Machine (SVM), K-Nearest Neighbors (KNN) e Random Forest (RF)*. Comparou o TAG com os resultados obtidos em [Pérez-Rosas et al. 2018] (T2), separando-os ano a ano (2013 até 2018);

T2) *Automatic Detection of Fake News* [Pérez-Rosas et al. 2018]: Este trabalho cria uma ferramenta de detecção de *Fake News* por classificação com *Support Vector Machines* (SVM), combinando informações léxicas, sintáticas, semânticas e de legibilidade. O presente trabalho compara os resultados com a detecção humana;

T3) *Automatic Detection of Fake News on Social Media Platforms* [Janze and Risius 2017]: Este artigo implementa a detecção com os classificadores binários *Logistic Regression, Support Vector Machines (SVM), Decision Tree, Random Forest e Extreme Gradient Boosting*. O referido trabalho compara os resultados entre os classificadores, onde os melhores resultados foram alcançados com SVM;

T4) *Automatically Identifying Fake News in Popular Twitter Threads* [Buntain and Golbeck 2017]: O trabalho apresenta um método para detecção de *Fake News* no *Twiter* que acumula, ao longo do tempo, as características de rede, usuário e conteúdo para gerar uma regressão linear. Assim, a abordagem realiza a sua análise, levando em consideração os aspectos temporais relacionados à notícia. O artigo avalia os resultados nos *datasets PHEME (Twitter* para rumor), *CredBank (Twitter)* e *BuzzFeed News Fact-Checking Dataset (Facebook)* que precisaram ser alinhados com as mesmas características e rótulos. Os resultados apontam que o *dataset CredBank* foi o mais indicado para a detecção automática de *Fake News* praticada;

T5) *Beyond News Contents: The Role of Social Context for Fake News Detection* [Shu et al. 2019b]: Este artigo explora as correlações da postura da notícia, o bias e engajamento do usuário. Assim, é apresentado um Tri-Relacionamento (TriFN) onde tanto informações partidárias quanto níveis de confiança do usuário podem ser utilizados para detecção de *Fake News*. Além disso, os usuários tendem a formar relacionamentos com pessoas afins que podem aumentar o espalhamento das *Fake News*. Esta abordagem compara os seus resultados com outros trabalhos, como [Rubin et al. 2015] (T23);

T6) *CIMTDetect: A Community Infused Matrix-Tensor Coupled Factorization Based Method for Fake News Detection* [Gupta et al. 2018]: Através da modelagem de Câmara de Ecos, o trabalho representa uma notícia como um *3-mode tensor* <News, User, Community> e propõe um método baseado em *tensor factorization*. Além disso, apresenta uma extensão deste método com a junção de modelos que utilizam o conteúdo da notícia através de um *framework coupled matrix-tensor factorization*. Este artigo usou o algoritmo de detecção da comunidade *Girvan-Newman* para identificar, na rede social, comunidades representativas de câmaras de eco. Os seus resultados são comparados com métodos que utilizam o classificador SVM, porém com diferentes formas de análise de conteúdo (ex. N-Gram). Os dois métodos propostos *CITDetect (community-infused tensor information) e CIMTDetect (community-infused tensor information + conteúdo da notícia)* utilizam o classificador SVM;

T7) *Combining Neural, Statistical and External Features for Fake News Stance Identification* [Bhatt et al. 2018]: Neste estudo a ferramenta, desenvolvida para o primeiro desafio (FNC-1) [3], não tem o objetivo final de detectar se a notícia é *Fake News*. Nesta abordagem, as notícias são classificadas de acordo com a relação existente entre a manchete e o corpo do texto. Portanto os possíveis rótulos são *Agree* - o texto do corpo concorda com a manchete, *Disagree* - o texto do corpo discorda da manchete, *Discuss* - o texto do corpo discute a mesma afirmação que o título, mas não toma uma posição ou *Unrelated* - o texto do corpo discute uma alegação que difere do título. A ferramenta combina as abordagens neural e estatística com recursos externos. Para isto, a solução implementa um modelo profundo recorrente (*Neural Embedding*), um modelo ponderado de características estatísticas (*n-gram bag-of-words*) e recursos externos criados à mão com a ajuda de uma heurística de engenharia de recursos. Por fim, usando uma camada de rede neural profunda, todas as referidas abordagens são combinadas. Os resultados foram comparados com as demais ferramentas participantes do referido desafio;

T8) *CSI: A Hybrid Deep Model for Fake News Detection* [Ruchansky et al. 2017]: O trabalho procura melhorar a acurácia na detecção de *Fake News* por meio de um modelo híbrido de rede neural profunda chamado CSI. Este modelo utiliza três características: o texto da notícia, a resposta do usuário que recebeu a notícia e o usuário fonte da notícia. O CSI trabalha com o comportamento temporal dos usuários e da notícia. Este modelo se divide em três partes: *Capture, Score e Integrate*. O primeiro módulo é baseado no texto e na resposta, por meio de uma rede neural recorrente (LSTM) para capturar um padrão temporal de atividades do usuário sobre a notícia e a representação *Doc2Vec*. O segundo usa uma rede neural para aprender as características da fonte, baseado nas interações dos usuários, gerando um score por meio de um grafo. Os dois módulos são integrados com o terceiro para caracterizar ou não a notícia como *Fake News*. O trabalho propõe a sua utilização em diferentes domínios, inclusive, em bancos de dados. Os resultados foram comparados com técnicas criadas para detecção de rumores;

T9) *DistrustRank: Spotting False News Domains* [Woloszyn and Nejdl 2018]: Esta solução propõe uma estratégia de aprendizagem semi-supervisionada para separar automaticamente notícias falsas a partir de fontes não confiáveis de notícias. O trabalho utiliza como fonte *experts* de portais de checagem de fatos para classificar manualmente as no-

[3] http://www.fakenewschallenge.org/

tícias. A partir disto, é criado um grafo de pesos com os *ranks* de confiança sobre os *sites* e as arestas representam a similaridade dos mesmos. A pesquisa computa a centralidade, utilizando o *PageRank* em busca de uma similaridade entre os *sites* não confiáveis. O resultado da análise é a classificação em *Trust* ou *Distrust* para a fonte da notícia. O trabalho verificou que a semelhança entre os sites de notícias falsas é estatisticamente superior aos sites de notícias verdadeiras. Esta abordagem cita e compara os seus resultados com outros trabalhos a partir do mesmo *dataset*;

T10) *EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection* [Wang et al. 2018b]: O artigo aponta que a maioria das abordagens existentes aprendem a detectar *Fake News* a partir de características específicas do evento, consequentemente, não podem ser transferidas para outros eventos ainda não aplicados. Assim, este trabalho desenvolveu um *framework*, de ponta a ponta, denominado *EANN*, que pode derivar características invariantes de um evento para outro. Desta forma, propõe uma detecção de *Fake News* para eventos recém-chegados. Isso consiste de três componentes principais: o extrator de características multimodais para texto e imagem (rede neural Convolucional), o detector de *Fake News (fully connected layer com softmax)* e o discriminador de eventos (rede neural) que é o responsável por remover as características específicas do evento e manter as características compartilháveis entre os eventos para poder rotulá-los. Assim, o *framework* mede as características não similares entre diferentes eventos e remove-os para capturar as características invariantes entre eventos. Para avaliar seus resultados, realizou testes com técnicas de identificação de texto e imagem, porém utilizadas em trabalhos não ligados à detecção de *Fake News*;

T11) *Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks* [Liu and BrookWu 2018]: O artigo propõe um modelo para detecção precoce de Fake News através da classificação dos caminhos de propagação da notícia. O referido trabalho modela o caminho de propagação de cada notícia como uma série temporal multivariada, na qual cada tupla é um vetor numérico que representa as características do usuário empenhado em espalhar a notícia. Para tal, é construído um classificador de série temporal que incorpora redes recorrente e convolucional. Estas redes capturam as variações globais e locais das características do usuário, ao longo do caminho de propagação, para detectar *Fake News*. Esta abordagem cita e compara os seus resultados com outros trabalhos a partir do mesmo *dataset*;

T12) *Evaluating Machine Learning Algorithms for Fake News Detection* [Gilda 2017]: Este artigo explora técnicas de linguagem natural para a detecção de *Fake News*. O trabalho aplicou *term frequency-inverse document frequency (TF-IDF) de bigrams e probabilistic context free grammar (PCFG)* para um conjunto de 11.000 artigos em um *dataset* obtido pela *Signal Media* [4] e uma lista de fontes da *OpenSources.com* [5]. Este *dataset* foi testado com os algoritmos de classificação *Support Vector Machines, Stochastic Gradient Descent, Gradient Boosting, Bounded Decision Trees e Random Forests*. Os modelos com melhor desempenho foram os *Stochastic Gradient Descent*, treinados apenas no conjunto de recursos do TF-IDF;

[4] https://research.signal-ai.com/newsir16/signal-dataset.html
[5] http://www.opensources.co

T13) *FActCheck: Keeping Activation of Fake News at Check* [Srivastava et al. 2018]: Esta abordagem de Intervenção sobre *Fake News* propõe uma melhoria na abordagem *competing cascades*, onde os *AFC (algoritmos polynomial time greedy) e RAFC (fast graph-pruning)* procuram escolher quais usuários têm maior poder de mitigação. Assim, os usuários com maior capacidade de influência na rede social realizam a mitigação através da divulgação de notícias alternativas *(Real News)*;

T14) *Fake News Detection in Social Networks via Crowd Signals* [Tschiatschek et al. 2018]: A ferramenta desenvolvida trabalha na detecção e consequente intervenção de *Fake News*. Esta solução possui um algoritmo, chamado de *Detective* que usa inferência Bayesiana para detectar *Fake News* a partir de *Crowd Signals*. Este *Crowd* é formado pela opinião dos usuários sobre a notícia, juntamente com a sua capacidade em opinar corretamente. O objetivo é detectar, de forma antecipada, a *Fake News* e bloqueá-la. Os resultados foram comparados a partir de variações na própria abordagem, sendo as mesmas denominadas pelo artigo como Opt, Oracle, Fixed-CM e No-Learn;

T15) *Fake News Mitigation Via Point Process Based Intervention* [Farajtabar et al. 2017]: Neste artigo, o enfoque está na intervenção de *Fake News*. A proposta é intervir, mitigando a notícia falsa, fornecendo recompensas na forma de notícias verdadeiras para quem recebeu a *Fake News*. A nível de influência da *Fake News* e a respectiva mitigação são quantificadas por contadores. O modelo utilizado foi baseado em *least-squares temporal difference learning* (LSTD). Um dos experimentos foi real, com a criação de cinco contas no *Twitter*;

T16) *FakeNewsTracker: A Tool for Fake News Collection, Detection, and Visualization* [Shu et al. 2019a]: Apresenta o FakeNewsTracker, um sistema para detecção de notícias falsas. O FakeNewsTracker pode coletar, automaticamente, dados para notícias e contexto social. Este tratalho propõe um *framework end to end* para realizar a coleta de dados, a detecção das *Fake News* e a visualização dos resultados. Esta pesquisa usa *autoencoders* para aprender o conteúdo de notícias e *RNN* para capturar o padrão temporal dos usuários de acordo com o seu engajamento com a notícia. O trabalho compara os seus resultados, internamente, a partir de variações do próprio *FakeNewsTracker*, onde são considerados somente o conteúdo da notícia ou o contexto social. Além disso, os resultados são comparados também com *Support Vector Machine, Logistic Regression and Naive Bayes*;

T17) *Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection* [Wang 2017]: Além de propor um *dataset*, cria uma técnica de detecção de *Fake News* híbrida, usando redes neurais convolucionais (CNNs) para analisar, não somente textos, mas também os dados do usuário. O artigo obteve os melhores resultados ao ser comparado com os de outros três detectores implementados com *Logistic Regression Classifier* (LR), *Support Vector Machine Classifier* (SVM) e *bi-directional long short-term memory* (Bi-LSTMs);

T18) *Neural User Response Generator: Fake News Detection with Collective User Intelligence* [Qian et al. 2018]: O trabalho enfatiza a rápida propagação das *Fake News* nas redes sociais e, portanto, destaca a importância da sua detecção nos estágios iniciais, onde considera que apenas o texto da notícia está disponível. Tal afirmação se baseia no fato de que informações adicionais, como respostas dos usuários e padrões de

propagação, podem ser obtidas somente após a notícia se espalhar. Contudo, como as respostas propagadas podem ajudar na tarefa de detecção, os autores propõe um *Two-Level Convolutional Neural Network with User Response Generator (TCNN-URG)* onde o TCNN captura a semântica do texto da notícia e o URG cria um modelo generativo de resposta dos usuários propagadores. O URG, a partir de respostas históricas, é treinado para aprender como os usuários respondem às notícias publicadas, gerando respostas de usuários para ajudar a TCNN na detecção da *Fake News*. Esta abordagem cita e compara os seus resultados com outros trabalhos a partir do mesmo *dataset*;

T19) *Ranking-based Method for News Stance Detection* [Zhang et al. 2018]: Mais uma pesquisa relacionada ao primeiro desafio (FNC-1). A solução do artigo é criada a partir de uma rede neural *Multi-Layer Perceptron*. Os resultados foram comparados com as demais ferramentas participantes do referido desafio;

T20) *Real-time Detection of Content Polluters in Partially Observable Twitter Networks* [Nasim et al. 2018]: Esta pesquisa procura encontrar um tipo específico de *bots*, chamados de poluidores de conteúdo, para poder distinguir notícias verdadeiras de *Fake News*. Segundo o artigo, o estado da arte de detecção de *bots*, normalmente, necessita de um histórico completo da rede. Assim, o trabalho propõe uma abordagem baseada em informações parciais onde, ao invés de mapear um grafo com seguidores e seguidos, utiliza um grafo com a (dupla de Usuário) x (Evento). Esta dupla é obtida a partir do momento em que o par tenha *tweetado* no mesmo dia do evento. Desta forma, os dados são clusterizados para que os usuários possam ser classificados como *bots* pela análise dos respectivos perfis e a frequência dos *tweets*. Os resultados do trabalho foram comparados com os obtidos por uma ferramenta citada pelo artigo, denominada de *Truthy*;

T21) *Sentiment Aware Fake News Detection on Online Social Networks* [Ajao et al. 2019]: O trabalho se aplica tanto a *Fake News* como Rumor. Assim, o artigo propõe a hipótese de que existe uma relação entre mensagens falsas ou rumores com os sentimentos dos textos. Foram utilizados dois modelos para extrair os escores de emoção (positividade, negatividade ou neutralidade) do texto: *Latent Semantic Analysis (LSA) e Latent Dirichlet Allocation (LDA)*. O objetivo foi desenvolver um classificador que utilize os escores de sentimento. Assim, utilizando classificadores distintos, compara os resultados a partir da abordagem proposta com sentimentos;

T22) *This Just In: Fake News Packs A Lot In Title, Uses Simpler, Repetitive Content in Text Body, More Similar To Satire Than Real News* [Horne and Adali 2017]: Detecção por meio da análise do texto. Este trabalho usa um classificador SVM e compara os seus resultados entre detecção de *Fake News, Real News* e Sátira. Este estudo determinou que as *Fake News* são mais próximas das Sátiras do que as notícias Reais;

T23) *Towards News Verification: Deception Detection Methods for News Discourse* [Rubin et al. 2015]: O trabalho propõe a ferramenta RST-SVM que analisa a notícia para extrair o estilo por meio da combinação do *Rhetorical Structure Theory (RST) e Vector Space Modeling (VSM)* para Clusterização. A detecção da notícia como enganosa ou real foi feita por meio de um classificador SVM. Os resultados obtidos foram comparados com a detecção humana;

T24) *Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate* [Wu and Liu 2018]: Este trabalho busca a detecção de *Fake News*, pela modelagem da propagação da notícia através da mineração de grafos em Florestas. Segundo o artigo, classificar notícias pelo seu conteúdo é muito difícil, devido à atual similaridade entre as divulgações *fake* e não *fake*. Em contra partida, as *Fake News* tendem a ter as mesmas fontes e sequências. O trabalho propõe a ferramenta paralelizável chamada TraceMiner que utiliza *Recurrent Neural Networks* (LSTM-RNNs), para classificar o caminho de propagação das mensagens no *Twitter*. O artigo comparou os seus resultados por meio de técnicas de análise de conteúdo criadas com SVM e XGBoost;

T25) *Weakly Supervised Learning for Fake News Detection on Twitter* [Helmstetter and Paulheim 2018]: Neste estudo, como existe uma dificuldade em conseguir um grande volume de dados para análise *(datasets)*, os *tweets* são rotulados, automaticamente, durante a coleta, de acordo com a confiança na sua fonte. Assim é criado um *dataset*, denominado de *Large-scale Training Dataset*, onde cada tweet de uma fonte confiável é rotulado como uma notícia real, assim como, cada tweet de uma fonte não confiável é rotulado como uma *Fake News*. Esperasse que neste *dataset* a classe de notícias reais contenha apenas uma quantidade negligenciável de ruído, pois fontes confiáveis raramente divulgam *Fake News*. Também é criado um segundo *dataset*, denominado de *Small-scale Evaluation Dataset*, possuindo *tweets* rotulados manualmente, como *fake* e não *fake*, a partir do site PolitiFact [6]. O objetivo principal do trabalho é treinar um classificador, a partir do primeiro *dataset*, para aplicá-lo no segundo *dataset*. Portanto, este classificador, apesar de ter sido treinado em um *dataset* desenvolvido a partir da confiança, é utilizado para detectar *tweets fakes* e não *fakes* no segundo *dataset*. Portanto, o artigo considera que o classificador foi treinado e avaliado com alvos distintos *(weakly supervised learning, mais especificamente, learning with inaccurate supervision)*. Para a referida detecção foram levadas em consideração as características do usuário (ex: engajamento e qtd seguidores), do tweet (ex: dia da semana, hora e texto), do tópico (assunto) e do sentimento. Como algoritmos de aprendizado foram usados o *Naive Bayes*, Árvores de Decisão, *Support Vector Machines* (SVM) e Redes Neurais. Além disso, foram usados dois *ensemble methods* conhecidos como *Random Forest e XGBoost*. Os resultados foram comparados, utilizando diferentes combinações para os classificadores;

T26) *XFake: Explainable Fake News Detector with Visualizations* [Yang et al. 2019]: O detector *XFake* é composto por 3 *frameworks: MIMIC, ATTN e PERT*. O *MIMIC* é construído para análise de atributos (ex. contexto da notícia e publicador) por meio de uma *deep neural network*. O *ATTN* é para análise semântica através de *pre-trained word embedding*, rede neural convolucional e *self-attention mechanism*. O *PERT* é para análise linguística utilizando um classificador *XGBoost*. A ferramenta, além de realizar as predições, também possui um módulo de interface para prover os usuários de explicações sobre as predições. O *XFake* é implementado em Python e deployed em *FLASK* com *front-end em HTML*. Para comparar seus resultados, o trabalho utilizou mão-de-obra humana realizada pela Amazon Mechanical Turk [7].

[6] https://www.politifact.com/

[7] https://www.mturk.com/

Tabela 1.1: Comparação entre abordagens - Dados de Publicação

Id	Dados							
	Publicação							
	Notícia		Usuário			Assunto		Temporalidade
	Mídia (Texto, Áudio e Imagem)	Conteúdo (Léxica, Sintática, Semântica e Legibilidade)	Tipo (Humano, Bot e Cyborg)	Perfil	Reputação	Especificidades	Relevância	
T1	Texto	Léxica e Semântica						
T2	Texto	Léxica, Sintática, Semântica e Legibilidade						
T3	Texto e Imagem	Léxica						
T4	Texto	Léxica e Semântica		X				X
T5	Texto	Léxica e Semântica		X	X			
T6	Texto	Léxica		X				
T7	Texto	Semântica						
T8	Texto	Léxica e Semântica		X	X			X
T9					X	Assuntos controversos		
T10	Texto e Imagem	Léxica						
T11				X				X
T12	Texto	Léxica e Semântica						
T13								
T14					X			X
T15								
T16	Texto	Léxica e Semântica		X				X
T17	Texto	Léxica e Semântica		X		Relaciona Assuntos		
T18	Texto	Semântica						
T19	Texto	Semântica						
T20			Bot	X				X
T21	Texto	Léxica e Semântica						
T22	Texto	Léxica e Sintática						
T23	Texto	Semântica						
T24					X			
T25	Texto	Léxica e Semântica		X		Análise dos Tópicos		X
T26	Texto	Léxica e Semântica		X		Análise dos Tópicos		

Tabela 1.2: Comparação entre abordagens - Dados de Propagação

Id	Dados								
	Propagação								
	Contribuição		Usuário			Assunto		Temporalidade	Rede
	Mídia (Texto, Áudio e Imagem)	Conteúdo (Léxica, Sintática, Semântica e Legibilidade)	Tipo (Humano, Bot e Cyborg)	Perfil	Reputação	Especificidades	Relevância		
T1									
T2									
T3	Texto	Léxica							X
T4	Texto	Léxica e Semântica		X				X	X
T5				X	X				X
T6				X					X
T7									
T8	Texto	Léxica e Semântica		X	X			X	X
T9									
T10									
T11				X				X	X
T12									
T13									X
T14					X			X	X
T15									X
T16				X				X	X
T17									
T18	Texto	Semântica							
T19									
T20			Bot	X				X	X
T21	Texto	Léxica e Semântica							
T22									
T23									
T24					X				X
T25	Texto	Léxica e Semântica							
T26									

Tabela 1.3: Comparação entre abordagens - Modelo, Funcionalidade e Atuação

Id	Modelo			Funcionalidade				Atuação (Centralizada ou Descentralizada)
				Detecção		Intervenção		
	Não Supervisionado	Semi Supervisionado	Supervisionado	Autenticidade	Intencionalidade	Bloqueio (Reativa)	Mitigação (Proativa e Reativa)	
T1			X	X				Centralizada
T2			X	X				Centralizada
T3			X	X				Centralizada
T4			X	X	Análise das características dos usuários			Centralizada
T5		X		X	Pontuação de credibilidade para os usuários			Centralizada
T6			X	X				Centralizada
T7			X	X				Centralizada
T8			X	X	Score para os usuários			Centralizada
T9		X		X	Atribui pesos de confiança aos websites			Centralizada
T10			X	X				Centralizada
T11			X	X				Centralizada
T12			X	X				Centralizada
T13							Reativa	Centralizada
T14			X	X		X		Centralizada
T15							Reativa	Centralizada
T16		X		X				Centralizada
T17			X	X	Associação com o usuário			Centralizada
T18			X	X				Centralizada
T19			X	X				Centralizada
T20		X		X	identificação de bots			Centralizada
T21			X	X	Análise de Sentimentos			Centralizada
T22			X	X				Centralizada
T23		X		X				Centralizada
T24			X	X	Relação entre os usuários			Centralizada (pode ser paralelizada)
T25			X	X	Análise de Sentimentos			Centralizada
T26			X	X				Centralizada

1.3.3. Datasets

Apesar da relevância do problema de combate às *Fake News* nas redes sociais, os *datasets* que contêm dados reais ainda estão raramente disponíveis para download. Como consequência, a maioria das pesquisas relacionadas ao combate às *Fake News* adaptou *datasets* originalmente criados para investigar outros problemas em redes sociais, como divulgação de *Rumor*. Esses datasets adaptados, geralmente, não contêm informações importantes para a detecção de *Fake News*, como rótulos *fake* / não *fake*. Além disso, a maioria desses *datasets*, adaptados ou originalmente criados para detecção de *Fake News*, não descrevem a propagação das notícias nas redes sociais, como uma mesma notícia divulgada por vários usuários e várias notícias divulgadas por um mesmo usuário. Assim, não há um consenso sobre os *datasets* de referência para este problema [Shu et al. 2017a]. Outro fator complicador para a criação de *datasets* é a carência de informação, proveniente das redes sociais, para combate às *Fake News*. Tal carência acontece pois, muitas vezes estas informações são apagadas, impossibilitando a sua análise [Mustafaraj and Metaxas 2017].

Independente das informações fornecidas pelo *dataset*, cabe salientar as diferentes formas pelas quais os referidos dados são disponibilizados:

- No Dataset: a informação está armazenada na própria base de dados;
- Link para o dado: a informação não está armazenada na base de dados, mas o

dataset disponibiliza um link direto para o dado específico;

• Link para a notícia: Nesta caso, o *dataset* simplesmente disponibiliza o link para a notícia. Assim, se faz necessário o acesso à notícia original para a retirada das informações desejadas.

Com o objetivo de apresentar alguns *datasets*, a Tabela 1.4 relaciona os trabalhos apresentados na Subseção 1.3.2 com os seus respectivos *datasets*. Em seguida, a Tabela 1.5 enquadra estes repositórios de acordo com os dados fornecidos por cada um deles. Este enquadramento é realizado no Modelo Comparativo, restrito ao aspecto *Dados*, tratado na Subseção 1.3.1.1. Cabe ressaltar que, na Tabela 1.5, as células não preenchidas indicam o não fornecimento do respectivo dado no *dataset* correspondente.

Além disso, os referidos *datasets* são brevemente descritos, podendo seus detalhes serem consultados através das respectivas referências:

D1) *BS Detector* [Shu et al. 2017a]: Este *dataset* é coletado de uma extensão de *browser* chamada *BS Detector* que foi desenvolvido para checagem da veracidade de notícias. Os rótulos existentes são *"Fake news", "Satire", "Extreme bias", "Conspiracy theory", "Rumor mill", "State news", "Junk science", "Hate group"e "Clickbait"*;

D2) *BuzzFace* [Santia and Williams 2018]: Este repositório foi criado pela equipe do BuzzFeed. Ele contém 2.282 artigos rotulados como *"Mostly true", "Mixture of true and false", "Mostly false"e "No factual content"*;

D3) *BuzzFeedNews (2016-10-facebookfact-check modificado)*
[Janze and Risius 2017]: Conjunto de dados criado a partir do *BuzzFeedNews (2016-10-facebookfact-check)* (D4), contudo os artigos são rotulados com *"Fake"e "Non-Fake"*;

D4) *BuzzFeedNews (2016-10-facebookfact-check)* [Shu et al. 2017a]: Este *dataset* compreende as notícias, do *Facebook*, oriundas de nove agências para a eleição presidencial americana de 2016. Os eventos e artigos ligados foram checados por jornalistas do *BuzzFeed*. Ele contém 1.627 artigos rotulados como *"Mostly true", "Mixture of true and false", "Mostly false"e "No factual content"*;

D5) *Celebrity* [Pérez-Rosas et al. 2018]: Este *dataset* fornece os dados da notícia para análise de texto. As notícias verdadeiras e falsas foram retiradas da *Web*, sendo relacionadas com assuntos de celebridades;

D6) *CredBank* [Shu et al. 2017a]: Conjunto de dados criado a partir do cruzamento de várias fontes, com aproximadamente 60 milhões de *tweets*, que cobrem 96 dias, iniciados em outubro de 2015. Todos os *tweets* são relacionados com mais de 1.000 eventos de notícias. Cada evento foi avaliado por 30 anotadores da *Amazon Mechanical Turk*. Os rótulos existentes são *"[-2]Certainly inaccurate", "[-1]Probably inaccurate", "[0] Uncertain (doubtful)", "[+1] Probably accurate"e "[+2] Certainly accurate"*;

D7) *DataSet Emergent* [Zhang et al. 2018][Bhatt et al. 2018]: Neste repositório, as notícias são rotuladas como *"Agree"* (o texto do corpo concorda com a manchete), *"Disagree"* (o texto do corpo discorda da manchete), *"Discuss"* (o texto do corpo discute a mesma afirmação que o título, mas não toma uma posição) e *"Unrelated"* (o texto do corpo discute uma alegação que difere do título). Esta base faz parte do primeiro desafio (FNC-1) e foi criado a partir do *dataset* para detecção de rumor chamado *Emergent*;

D8) *DistrustRank Datasets* [Woloszyn and Nejdl 2018]: Foram desenvolvidos dois *datasets*. O primeiro, gerado com sites confiáveis, por meio do *SimilarWeb* [8], tem 502 domínios e 396.422 *URLs* de notícias. O segundo, obtido com sites não confiáveis, através do *Wikipedia's list of prominent Fake News* [9], possui 47 dominios e 37.320 *URLs* de notícias;

D9) *Facebook para Detective* [Tschiatschek et al. 2018]: Repositório que considera os círculos sociais do *Facebook*, consistindo de 4.039 usuários (nós) e 88.234 arestas;

D10) *Fake News vs Satire* [Golbeck et al. 2018]: *DataSet* para diferenciar *Fake News* e Sátiras onde as notícias são codificadas manualmente. A base, oriunda de diversas fontes, é composta por 283 relatos rotulados como *Fake News* e 203 como *Satirical*. Estes relatos são compostos pelo título, texto e um link para cada artigo;

D11) *FakeNewsAMT* [Pérez-Rosas et al. 2018]: As notícias falsas e legítimas são fornecidas em duas pastas separadas. Cada pasta contém 40 notícias de seis domínios diferentes: tecnologia, educação, negócios, esportes, política e entretenimento;

D12) *FakeNewsData1* [Horne and Adali 2017]: São dois *datasets* onde o primeiro contém notícias rotuladas como *Fake e Real* retiradas a partir do *BuzzFeed*. Já o segundo contém notícias políticas rotuladas como *Real, Fake* e Sátira obtidas, randomicamente, durante as eleições americanas de 2016;

D13) *FakeNewsNet1* [Shu et al. 2017a] [Shu et al. 2019b] [Sharma et al. 2019] [Gupta et al. 2018] [Shu et al. 2019a]: Esta base de dados, coletada do Twitter, fornece 211 notícias *Fake* e 211 notícias *Real*, rotuladas a partir do *BuzzFeed* e *PolitiFact*;

D14) *FakeNewsNet2* [Shu et al. 2018][Sharma et al. 2019]: Esta base de dados, coletada do Twitter, fornece 6.480 notícias *Fake* e 17.441 notícias *Real*, rotuladas a partir do *GossipCop* [10] e *PolitiFact*;

D15) *Kaggle* [11]: Este conjunto de dados contém texto e metadados de 244 sites, totalizando 12.999 postagens. Os dados foram extraídos usando a *API webhose.io*. Cada site foi rotulado de acordo com o *BS Detector*, sendo que as fontes de dados sem rótulo foram categorizadas como "Bs";

D16) *KV* [Dong et al. 2014]: Nesta base as notícias têm sujeito, predicado e objeto. Cada notícia tem um rótulo que indica a probabilidade da mesma ser verdadeira. A ferramenta, por meio de uma fusão de conhecimentos, cria um grafo relacionando o sujeito com o objeto para medir a quantidade de interações e, assim, gerar automaticamente o *dataset*;

D17) *Large-scale Training Dataset e Small-scale Evaluation Dataset* [Helmstetter and Paulheim 2018]: No *Large-scale Training Dataset* cada tweet de uma fonte confiável é rotulado como notícia real e cada tweet de um uma fonte não confiável é rotulado como uma *Fake News*. As 46 fontes confiávies e 65 não confiáveis foram obtidas através de pesquisas em sites e os *tweets* foram coletados a partir destas fontes. No total,

[8]https://www.similarweb.com/top-websites/category/News-and-media
[9]https://en.wikipedia.org/wiki/List-of-fake-News-websites
[10]https://https://www.gossipcop.com/
[11]https://www.kaggle.com/datasets

foram coletados 401.414 exemplos, nos quais 110.787 (27,6 porcento) foram rotulados como *Fake News*, enquanto 290.627 (24,4 porcento) foram rotulados como *Real News*. O *Small-scale Evaluation Dataset* contém 116 *tweets* rotulados manualmente e obtidos no *PolitiFact*;

D18) *LIAR* [Wang 2017]: Esta base de dados é coletada do *PolitiFact*. Ele inclui 12.836 notícias rotuladas manualmente como *"Pants-fire", "False", "Barely-true", "Half-true", "Mostly true"e "True"*. Cabe salientar que os dados referentes ao usuário se resumem ao nome do autor da postagem;

D19) *PoliticalNews* [Castelo et al. 2019]: Para criar o dataset foram usados os sites *Politifact, BuzzFeed, OpenSources.co e Alexa's top 500 news* [12]. O resultado foi um *dataset* com 14.240 páginas de notícias sendo 7.136 páginas vindas de 79 sites não confiáveis e 7.104 vindos de 58 sites confiáveis;

D20) *PolitiFact para XFake* [Yang et al. 2019]: Repositório criado a partir do site *PolictFact* com 5.104 notícias contendo os atributos *Subject, Context, Speaker, Targeting e Statement*. As notícias foram rotuladas como *True e False*;

D21) *RST-SVM Dataset* [Rubin et al. 2015]: Esta base de dados foi criada a partir de codificadores, usando notícias do *Bluff the Listener* [13]. Este repositório consiste de 144 notícias selecionadas, aleatoriamente, de 2010 até 2014;

D22) *Signal Media para Evaluating Machine Learning Algorithms for Fake News Detection* [Gilda 2017]: *Dataset* rotulado com *"Fake"* ou *"Não fake"* criado a partir de uma base de notícias da *Signal Media* e uma lista do repositório de confiança de fontes *OpenSources.co*. O citado *dataset* contém 11.051 artigos, sendo 3.217 categorizados com falsos;

D23) *Soc-LiveJournal* [Srivastava et al. 2018]: Este repositório não rotulado contém uma rede de relacionamentos formada por 4.847.571 nós e 68.475.391 arestas;

D24) *Twitter e Sina Weibo para CSI* [Ruchansky et al. 2017]: *Dataset* criado com 2.811 artigos rotulados como *"Fake"* e 2.845 como *"True"*. A citada base de dados foi obtida a partir do repositório, para detecção de rumores, gerado no artigo [Ma et al. 2016];

D25) *Twitter e Sina Weibo para EANN* [Ruchansky et al. 2017]: A base de dados foi criada a partir de dois *datasets* não originários de *Fake News*. O primeiro repositório foi obtido a partir do *Sina Weibo* contendo 4.749 notícias com rótulos adaptados para *fake* e 4.779 para real, além de 9.528 imagens. O segundo repositório foi obtido a partir do *Twitter* contendo 7.898 notícias com rótulos adaptados para *fake* e 6.026 para real, além de 514 imagens;

D26) *Twitter e Sina Weibo para Early Detection Through Propagation Path* [Liu and BrookWu 2018]: Este repositório foi criado a partir de três *datasets* usados para detecção de rumores. O primeiro, oriundo da rede social *Weibo*, com os rótulos *"rumor (fake)" e "otherwise (true)"*. Já os outros dois *datasets*, obtidos do *Twitter*, são rotulados como *"fake", "true", "unverified" e "non-rumor (debunking of fake)"*. As características dos usuários foram obtidas por meio de pesquisas realizadas nas respectivas redes sociais.

[12] https://www.alexa.com/topsites/category/News

[13] https://www.npr.org/bluff-the-listener

D27) *Twitter e Sina Weibo para TCNN-URG* [Qian et al. 2018]: Base de dados que utilizou dois *datasets*. O primeiro *dataset* foi obtido, automaticamente, a partir do *Sina Weibo*. Já o segundo *dataset* foi gerado por um processo manual de coleta de dados. Para tal, foram selecionadas notícias em sites avaliados como confiáveis (*The Guardian* [14]) e notoriamente falsos. Com as URLs de todas as notícias coletadas, pesquisas foram realizadas no *Twitter* para cada uma das notícias classificadas como falsas ou reais.

D28) *Twitter para Automatically Identifying Fake News*
[Buntain and Golbeck 2017]: Base de dados que utilizou os *datasets* PHEME (rumor no *Twitter*), CredBank (credibilidade no *Twitter*) e *BuzzFeed News Fact-Checking Dataset* (Checagem de fatos no *Facebook*). Os três *datasets* precisaram ser alinhados com as mesmas características e rótulos;

D29) *Twitter para Content Polluters* [Nasim et al. 2018]: Repositório de dados criado para detecção de *bots*. Este *dataset*, obtido a partir do *Twitter*, foi rotulado manualmente como "Bot"ou "Não Bot";

D30) *Twitter para Mitigation via Point Process* [Farajtabar et al. 2017]: Este trabalho realizou experimentos com contas reais no Twitter e com uma base de dados sintética onde, entre *N* nós, foi assumido que 20 nós criaram *Fake News* e outros 20 nós divulgaram notícias verdadeiras;

D31) *Twitter para TraceMiner* [Wu and Liu 2018]: Conjunto de dados gerado pela coleta de informações do *Twitter* com rotulação a partir do site de checagem de fatos *Snopes* [15]. Nesta base, os rótulos atribuidos são *"Real news"ou "Fake news"*;

D32) *Twitter Trec* [Srivastava et al. 2018]: Conjunto de dados gerado pela coleta de informações do *Twitter*, sem rotulação, contendo uma rede de relacionamentos formada por 3.919.215 nós e 5.399.949 arestas.

Tabela 1.4: Trabalhos x Datasets

Id	DataSet
T1	FakeNewsAMT (D11), Celebrity (D5) e PoliticalNews (D19)
T2	FakeNewsAMT (D11) e Celebrity (D5)
T3	BuzzFeedNews (2016-10-facebookfact-check modificado) (D3)
T4	Twitter para Automatically Identifying Fake News (D28)
T5	FakeNewsNet1 (D13)
T6	FakeNewsNet1 (D13)
T7	DataSet Emergent (D7)
T8	Twitter e Sina Weibo para CSI (D24)
T9	DistrustRank Datasets (D8)
T10	Twitter e Sina Weibo para EANN (D25)
T11	Twitter e Sina Weibo para Early Detection Through Propagation Path (D26)
T12	Signal Media para Evaluating Machine Learning Algorithms for Fake News Detection (D22)
T13	Soc-LiveJournal (D23) e Twitter Trec (D32)
T14	Facebook para Detective (D9)
T15	Twitter para Mitigation via Point Process (D30)
T16	FakeNewsNet1 (D13)
T17	LIAR (D18)
T18	Twitter e Sina Weibo para TCNN-URG (D27)
T19	DataSet Emergent (D7)
T20	Twitter para Content Polluters (D29)
T21	Pheme (dataset para Rumor)
T22	FakeNewsData1 (D12)
T23	RST-SVM Dataset (D21)
T24	Twitter para TraceMiner (D31)
T25	Large-scale Training Dataset e Small-scale Evaluation Dataset (D17)
T26	PolitiFact para XFake (D20)

[14]https://www.theguardian.com/

[15]https://www.snopes.com

Tabela 1.5: Comparação entre Datasets

Id	Dados									URL
	Publicação				Propagação					
	Notícia			Usuário	Contribuição			Usuário	Rede	
	Texto	Áudio	Imagem		Texto	Áudio	Imagem			
D1	Link para notícia	Link para notícia	Link para notícia	No Dataset	Link para notícia	Link para notícia	Link para notícia	Link para notícia	Link para notícia	https://github.com/thiagovas/bs-detector-dataset
D2	Link para notícia	Link para notícia	Link para notícia	No Dataset	Link para notícia	Link para notícia	Link para notícia	No Dataset	Link para notícia	https://github.com/gsantia/BuzzFace
D3	No Dataset		Link para imagem	No Dataset	No Dataset			Link para notícia	Link para notícia	
D4	Link para notícia	Link para notícia	Link para notícia	No Dataset	Link para notícia	Link para notícia	Link para notícia	Link para notícia	Link para notícia	https://github.com/BuzzFeedNews/2016-10-facebook-fact-check
D5	No Dataset			No Dataset						http://lit.eecs.umich.edu/downloads.html#undefined
D6	No Dataset			No Dataset				No Dataset	No Dataset	http://compsocial.github.io/CREDBANK-data/
D7	No Dataset			No Dataset						https://github.com/FakeNewsChallenge/fnc-1
D8	Link para notícia	Link para notícia	Link para notícia	No Dataset						
D9				No Dataset				No Dataset		
D10	No Dataset	Link para notícia	Link para notícia	No Dataset						https://github.com/jgolbeck/fakenews
D11	No Dataset			No Dataset						http://lit.eecs.umich.edu/downloads.html#undefined
D12	No Dataset									https://github.com/BenjaminDHorne/fakenewsdata1/blob/master/Horne2017_FakeNewsData.zip
D13	No Dataset		Link para imagem	No Dataset				No Dataset	No Dataset	https://github.com/KaiDMML/FakeNewsNet
D14	No Dataset		Link para notícia	No Dataset				No Dataset	No Dataset	https://github.com/KaiDMML/FakeNewsNet
D15	No Dataset		Link para imagem	No Dataset						https://www.kaggle.com/mrisdal/fake-news/data
D16	No Dataset			No Dataset						
D17	No Dataset			No Dataset						http://dws.informatik.uni-mannheim.de/en/research/twitter-fake-news-detection
D18	No Dataset			No Dataset						https://github.com/nishitpatel01/Fake_News_Detection/tree/master/liar_dataset ou https://www.cs.ucsb.edu/~william/software.html
D19	No Dataset			No Dataset						https://osf.io/ez5q4/
D20	No Dataset			No Dataset						
D21	No Dataset			No Dataset						
D22	No Dataset			No Dataset						
D23									No Dataset	https://snap.stanford.edu/data/soc-LiveJournal1.html
D24	No Dataset			No Dataset	No Dataset			No Dataset	No Dataset	https://github.com/majingCUHK/Rumor_RvNN ou http://alt.qcri.org/~wgao/data/rumdect.zip
D25	No Dataset		No Dataset	No Dataset						
D26				No Dataset				No Dataset	No Dataset	Twitter 15 e 16 (https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0) e Weibo(http://alt.qcri.org/~wgao/data/rumdect.zip)
D27	No Dataset			No Dataset	No Dataset					False (https://drive.google.com/open?id=1WRoRVV9j4CStMFKDwP7DVGAHFJZX4t5a) e True(https://drive.google.com/open?id=1JgbW4suN2yWHx65P4QU8HkrrB3OMHsuo)
D28	No Dataset			No Dataset				No Dataset	No Dataset	
D29				No Dataset				No Dataset	No Dataset	
D30									No Dataset	
D31				No Dataset				No Dataset	No Dataset	
D32	No Dataset			No Dataset				No Dataset	No Dataset	https://trec.nist.gov/data/tweets/

1.4. Estudo de Caso em Detecção Automática de *Fake News*

Dentre as abordagens de detecção de *Fake News* apresentadas na Subseção 1.3.2, destacam-se as baseadas na reputação do usuário. Uma das razões para tal destaque é a não necessidade da utilização do conteúdo das notícias. Haja vista que a atual similaridade entre as notícias *fake* e não *fake* [Liu and BrookWu 2018] tem dificultado a detecção de *Fake News* por conteúdo.

Um dos principais trabalhos que utilizam a reputação dos usuários para detectar *Fake News* é [Tschiatschek et al. 2018]. Este trabalho se sobressai, pois a reputação do usuário não é obtida por meio do perfil do usuário na rede social, tendo como base, a dificuldade em se obter tais informações, normalmente de cunho sigiloso [Shu et al. 2017b]. Assim, [Tschiatschek et al. 2018] obtém a reputação do usuário a partir da sua capacidade em sinalizar as notícias. Em resumo, por meio de uma funcionalidade disponível na rede social, o usuário pode opinar se as notícias visualizadas são *fake* ou não. Assim, a reputação do usuário é expressa em função do seu histórico de acertos e erros em suas opiniões. Desta forma, [Tschiatschek et al. 2018] propõe um método chamado *Detective* que classifica uma notícia, como *fake* ou não, a partir de *Crowd Signals*. Este *Crowd* é formado pelas opiniões dos usuários, juntamente com as suas respectivas reputações. Este método, baseado em *Crowd Signals*, em essência, utiliza um classificador bayesiano binário, cujos conceitos básicos e fundamentos são descritos abaixo.

Dada uma rede social $\mathscr{G}$, a entrada do *Detective* contém os seguintes elementos:

um intervalo de tempo t (por exemplo, um dia), um conjunto de usuários U de $\mathscr{G}$, um *dataset* D com notícias rotuladas e uma notícia específica a ser analisada a.

D contém notícias com dois tipos de rótulos: o real e o sinalizado pelo usuário. O rótulo real e o seu valor são indicados pelas variáveis $Y^*(x)$ e $y^*(x)$, respectivamente, onde $y^*(x)$ pertence a $\{f, \overline{f}\}$ onde $y^*(x) = f$ (resp. $y^*(x) = \overline{f}$) significa que uma notícia x é *fake* (resp. não *fake*). Denotado por uma variável $Y_u(x)$, o rótulo sinalizado é aquele atribuído por um usuário u para uma notícia x. Seu valor $y_u(x)$ pertence a $\{f, \overline{f}\}$ onde $y_u(x) = f$ (resp. $y_u(x) = \overline{f}$) significa que u sinalizou x como *fake* (resp. não *fake*). É importante notar que, diferente de outras notícias, $y^*(a)$ é desconhecido e deve ser previsto pelo *Detective*.

Inicialmente, *Detective* aplica as funções $\pi^t(a)$ e $\psi^t(a)$ ao D. Enquanto o primeiro retorna o conjunto de usuários que viram a notícia a no final da época t, o último retorna o conjunto completo de usuários que sinalizaram a como *fake* no final de t.

O *Detective* pode assumir que não há abstinência na sinalização e, para cada usuário $u \in \pi^t(a)$, calcula $\theta_{u,\overline{f}}$ e $\theta_{u,f}$, considerando as notícias sinalizadas por u antes de t. Assim, $\theta_{u,\overline{f}}$ (resp. $\theta_{u,f}$) é a probabilidade de u sinalizar uma notícia x como não *fake* (resp. *fake*), dado que x é realmente não *fake* (resp. fake). Em ambos os casos, o cálculo da probabilidade é limitado ao conjunto de notícias revisadas por u antes de t. Assim, para cada usuário u, *Detective* representa a observada atividade de sinalização de u pela correspondente matriz $\mathscr{M}_u$, genericamente definida como segue.

$$\begin{vmatrix} \theta_{u,\overline{f}} & 1 - \theta_{u,f} \\ 1 - \theta_{u,\overline{f}} & \theta_{u,f} \end{vmatrix}$$

onde:

- $\theta_{u,\overline{f}} = P(Y_u(x) = \overline{f} \mid Y^*(x) = \overline{f})$
- 1 - $\theta_{u,\overline{f}} = P(Y_u(x) = f \mid Y^*(x) = \overline{f})$
- $\theta_{u,f} = P(Y_u(x) = f \mid Y^*(x) = f)$
- 1 - $\theta_{u,f} = P(Y_u(x) = \overline{f} \mid Y^*(x) = f)$

Por fim, seguindo uma abordagem Bayesiana, o *Detective* usa as equações 2 e 3 para calcular as probabilidades de a ser *fake* e não *fake*, respectivamente. Sendo que ω (resp. $1 - \omega$) é a probabilidade a priori de que qualquer notícia seja *fake* (resp. não *fake*). Ambas as equações consideram a capacidade dos usuários de acertar, assim como de errar suas opiniões de acordo com seu voto. Assim, o *Detective* pode se beneficiar quando os usuários acertarem ou errarem, mesmo que eles mostrem incapacidade [Freeman 2017] ou má intenção ao avaliar as notícias. A classe correspondente à maior probabilidade é a opinião do *Detective* sobre a e, portanto, sua saída.

$$P(Y^*(a) = f) = \omega \cdot \prod_{u \in \psi^t(a)} \theta_{u,f} \cdot \prod_{u \in \pi^t(a) \setminus \psi^t(a)} (1 - \theta_{u,f}) \tag{2}$$

$$P(Y^*(a) = \overline{f}) = (1 - \omega) \cdot \prod_{u \in \psi^t(a)} (1 - \theta_{u,\overline{f}}) \cdot \prod_{u \in \pi^t(a) \setminus \psi^t(a)} \theta_{u,\overline{f}} \tag{3}$$

Para este estudo de caso, a metodologia experimental utilizada foi semelhante à metodologia ótima seguida por [Tschiatschek et al. 2018]. As probabilidades θ foram aleatoriamente designadas aos usuários, criando três grupos: *bom* ($\theta_{u,\overline{f}} = \theta_{u,f} = 0.9$), *indiferente* ($\theta_{u,\overline{f}} = \theta_{u,f} = 0.5$) e *spammer* ($\theta_{u,\overline{f}} = \theta_{u,f} = 0,1$). Assim, como realizado em [Tschiatschek et al. 2018], foi assumido que nenhum usuário se absteve de dar sua opinião sobre as notícias a serem analisadas. Embora não esteja claramente indicado em [Tschiatschek et al. 2018], também foi assumido que cada usuário deveria sinalizar aleatoriamente uma notícia de acordo com a probabilidade atribuída ao seu grupo. Por exemplo: dada uma notícia *a* para ser analisada por *u*, um *bom* usuário. De acordo com a configuração definida para *bons* usuários, *u* deve acertar ou errar para *a* com probabilidades de 90% e 10%, respectivamente. Além disso, foi usado o método da roleta para decidir se cada usuário deve acertar ou errar o rótulo real de uma notícia. Embora pouco realista, essa metodologia leva aos resultados de maior precisão produzidos pelo *Detective*.

Para a execução dos nossos experimentos, foram escolhidos os *datasets* BuzzFeed e PolitiFact, ambos pertencentes ao repositório FakeNewsNet1 (D13), descrito na Subseção 1.3.3. Nossa escolha foi guiada por três razões principais. Primeiro, estes *datasets* foram criados para o específico propósito de detecção de *Fake News* e contêm, para cada notícia, seu rótulo real, ou seja, a indicação de que a notícia é *fake* ou não. Em segundo lugar, eles descrevem a propagação das notícias nas redes sociais. Por fim, eles foram usados e disponibilizados por publicações recentes e relevantes [Sharma et al. 2019] [Shu et al. 2017a] [Shu et al. 2019b] [Shu et al. 2019a] [Gupta et al. 2018]. A Tabela 1.6 fornece uma visão estatística geral dos *datasets* escolhidos.

Tabela 1.6: *Datasets* usados nos Experimentos

Dataset	Não *Fake News*	*Fake News*	Usuários	Média de usuários por notícia
BuzzFeed	91	91	15257	125,16
PolitiFact	120	120	23865	136,63

Para se obter os resultados do método *Detective* foi utilizada a acurácia como métrica de desempenho. A Tabela 1.7 ilustra o resultado de uma rodada de execução do experimento.

Tabela 1.7: Acurácia do método de detecção de *Fake News*

Método	BuzzFeed	PolitiFact
Detective	**0.9890**	**0.9791**

1.5. Problemas em Aberto

O combate automático às *Fake News* em redes sociais é uma nova e emergente área de pesquisa que, mesmo com estudos já realizados, ainda carece de maior aprofundamento científico. Desta forma, os seguintes problemas são descritos como áreas ainda fertéis para o desenvolvimento de novos trabalhos:

• Carência de *datasets* que forneçam, de forma suficiente, os diferentes dados necessários para combater as *Fake News* em redes sociais;

• Trabalhos que levem em consideração aspectos temporais do ciclo de vida da *Fake News* e que, consequentemente, possam intervir mais rapidamente;

• Estudos que analisem o aspecto intencional, assim não se limitam a verificar a autenticidade (veracidade) das notícias;

• Extração de características a partir de imagem e/ou áudio, portanto não se limitando as análises de texto;

• Métodos que abordem características baseadas na rede que representa a propagação da notícia. Neste caso, inclusive, podem ser aplicadas técnicas baseadas em grafos;

• Pesquisas que, ao invés de realizarem uma classificação binária, utilizem probabilidades e/ou pertinências na detecção. Esta linha de trabalho se baseia no fato de que, normalmente, a *Fake News* é uma mistura de afirmações falsas e verdadeiras;

• Utilização de um comitê de classificadores para determinar se uma notícia é *fake*. Desta forma, pode-se agregar diferentes técnicas de classificação na detecção;

• Utilização de modelos não supervisionados ou semi-supervisionados devido à carência de *datasets* rotulados que possuam variedade de dados;

• Estudo sobre o comportamento distinto da *Fake News* em diferentes comunidades (escolar, trabalho e etc) e/ou redes sociais *(Weibo, WhatsApp e etc)*. Isto se deve pela possível mudança de comportamento das notícias de acordo com o meio;

• Classificar os usuários de *Fake News* com o objetivo de identificar o seu tipo *(humanos, bots e cyborgs)*. Isto se deve pela possível alteração de comportamento das notícias propositalmente falsas de acordo com o seu tipo de usuário divulgador;

• Trabalhos relacionados à intervenção de *Fake News*, tanto para bloqueio quanto para mitigação. Haja vista que o combate às *Fake News* não se limita à detecção, sendo necessária, também, a intervenção sobre a mesma;

• Abordagens que atuem descentralizadas na rede. Esta atuação se destaca, pois quanto mais rápido e extensivo for o combate, menor serão os efeitos nocivos da notícia;

• Abordagens que utilizem o assunto para a análise da notícia, pois assuntos relevantes, normalmente, motivam a criação de notícias intencionalmente falsas;

• Pesquisas que levem em consideração a reputação dos usuários, pois usuários com baixa reputação tendem a ser potenciais divulgadores de *Fake News*.

1.6. Considerações Finais

Cada vez mais pessoas estão consumindo notícias das redes sociais, ao invés dos canais tradicionais. Tal tendência amplificou a disseminação de *Fake News*, isto é, as notícias falsas publicadas intencionalmente. Este tipo de notícia pode ter significativos impactos sociais negativos, por exemplo, a manipulação da opinião em larga escala.

Tendo como base os riscos que as *Fake News* trazem para a sociedade, tanto a academia quanto a indústria buscam por soluções que viabilizem o combate a este tipo de notícia nas redes sociais. Ademais, devido tanto ao volume como à velocidade de divulgação das *Fake News*, faz-se necessário o emprego de abordagens computacionais no combate às notícias intencionalmente falsas nas redes sociais.

Portanto, neste capítulo, nós exploramos o combate automático às *Fake News* em redes sociais. Para tal, revisamos a literatura existente, visando realizar um levantamento das abordagens computacionais, tendo como base a proposta de um modelo comparativo. Em seguida, um estudo de caso foi realizado, objetivando uma introdução, não só teórica, como também prática. Por fim, foram apresentados alguns problemas sobre o referido combate que ainda carecem de maior aprofundamento científico.

Referências

[Ajao et al. 2019] Ajao, O., Bhowmik, D., and Zargari, S. (2019). Sentiment aware fake news detection on online social networks. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2507–2511.

[Bhatt et al. 2018] Bhatt, G., Sharma, A., Sharma, S., Nagpal, A., Raman, B., and Mittal, A. (2018). Combining neural, statistical and external features for fake news stance identification. In Companion Proceedings of the The Web Con 2018, WWW '18, pages 1353–1357, Republic and Canton of Geneva, Switzerland. International World Wide Web Con Steering Committee.

[Braz and Goldschmidt 2017] Braz, P. and Goldschmidt, R. (2017). Um método para detecção de bots sociais baseado em redes neurais convolucionais aplicadas em mensagens textuais. In SBSeg 2017, pages 501–508. 10/11/2017.

[Buntain and Golbeck 2017] Buntain, C. and Golbeck, J. (2017). Automatically identifying fake news in popular twitter threads. In 2017 IEEE International Con on Smart Cloud (SmartCloud), pages 208–215.

[Campan et al. 2017] Campan, A., Cuzzocrea, A., and Truta, T. M. (2017). Fighting fake news spread in online social networks: Actual trends and future research directions. In 2017 IEEE International Con on Big Data (Big Data), pages 4453–4457.

[Castelo et al. 2019] Castelo, S., Almeida, T., Elghafari, A., Santos, A., Pham, K., Nakamura, E., and Freire, J. (2019). A topic-agnostic approach for identifying fake news pages. In Companion Proceedings of The 2019 World Wide Web Conference, WWW '19, pages 975–980, New York, NY, USA. ACM.

[Cazalens et al. 2018] Cazalens, S., Lamarre, P., Leblay, J., Manolescu, I., and Tannier, X. (2018). A content management perspective on fact-checking. In Companion

Proceedings of the The Web Con 2018, WWW '18, pages 565–574, Republic and Canton of Geneva, Switzerland. International World Wide Web Con Steering Committee.

[Ciampaglia et al. 2015] Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. (2015). Computational fact checking from knowledge networks. PLOS ONE, 1:1–13.

[Conroy et al. 2015] Conroy, N., Rubin, V., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Association for Information Science and Technology, 52:1–4.

[Deng et al. 2014] Deng, S., Huang, L., and Xu, G. (2014). Social network-based service recommendation with trust enhancement. Expert Systems with Applications, 41(18):8075 – 8084.

[Dong et al. 2014] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In ACM SIGKDD international Con on Knowledge discovery and data mining, pages 601–610.

[Farajtabar et al. 2017] Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., Li, S., Song, L., and Zha, H. (2017). Fake news mitigation via point process based intervention. In Proceedings of the 34th International Con on Machine Learning - Volume 70, ICML'17, pages 1097–1106. JMLR.org.

[Ferrara et al. 2016] Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. Commun. ACM, 59(7):96–104.

[Flintham et al. 2018] Flintham, M., Karner, C., Bachour, K., Creswick, H., Gupta, N., and Moran, S. (2018). Falling for fake news: Investigating the consumption of news via social media. In Proceedings of the 2018 CHI Con on Human Factors in Computing Systems, CHI '18, pages 376:1–376:10, New York, NY, USA. ACM.

[Freeman 2017] Freeman, D. M. (2017). Can you spot the fakes?: On the limitations of user feedback in online social networks. In Proceedings of the 26th International Conference on World Wide Web, WWW '17, pages 1093–1102, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[Gilda 2017] Gilda, S. (2017). Evaluating machine learning algorithms for fake news detection. In 2017 IEEE 15th Student Con on Research and Development (SCOReD), pages 110–115.

[Golbeck et al. 2018] Golbeck, J., Mauriello, M., Auxier, B., Bhanushali, K. H., Bonk, C., Bouzaghrane, M. A., Buntain, C., Chanduka, R., Cheakalos, P., Everett, J. B., Falak, W., Gieringer, C., Graney, J., Hoffman, K. M., Huth, L., Ma, Z., Jha, M., Khan, M., Kori, V., Lewis, E., Mirano, G., Mohn IV, W. T., Mussenden, S., Nelson, T. M., Mcwillie, S., Pant, A., Shetye, P., Shrestha, R., Steinheimer, A., Subramanian, A., and Visnansky, G. (2018). Fake news vs satire: A dataset and analysis. In Proceedings of

the 10th ACM Con on Web Science, WebSci '18, pages 17–21, New York, NY, USA. ACM.

[Gupta et al. 2018] Gupta, S., Thirukovalluru, R., Sinha, M., and Mannarswamy, S. (2018). Cimtdetect: A community infused matrix-tensor coupled factorization based method for fake news detection. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 278–281.

[Helmstetter and Paulheim 2018] Helmstetter, S. and Paulheim, H. (2018). Weakly supervised learning for fake news detection on twitter. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 274–277.

[Hendrikx et al. 2015] Hendrikx, F., Bubendorfer, K., and Chard, R. (2015). Reputation systems: A survey and taxonomy. Journal of Parallel and Distributed Computing, pages 184–197.

[Horne and Adali 2017] Horne, B. D. and Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In Association for the Advancement of Artificial Intelligence.

[Janze and Risius 2017] Janze, C. and Risius, M. (2017). Automatic detection of fake news on social media platforms. In PACIS 2017.

[Kim et al. 2018] Kim, J., Tabibian, B., Oh, A., Schölkopf, B., and Gomez-Rodriguez, M. (2018). Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In Proceedings of the Eleventh ACM International Con on Web Search and Data Mining, WSDM '18, pages 324–332, New York, NY, USA. ACM.

[Kshetri and Voas 2017] Kshetri, N. and Voas, J. (2017). The economics of fake news. IT Professional, 19(06):8–12.

[Li et al. 2015] Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., and Han, J. (2015). A survey on truth discovery. ACM SIGKDD Explorations Newsletter, 17:1–16.

[Liu and BrookWu 2018] Liu, Y. and BrookWu, Y. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In AAAI Con on Artificial Intelligence, pages 354–361.

[Liu and Xu 2016] Liu, Y. and Xu, S. (2016). Detecting rumors through modeling information propagation networks in a social media environment. IEEE Transactions on Computational Social Systems, 3(2):46–62.

[Ma et al. 2016] Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K., and Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. In International Joint Con on Artificial Intelligence.

[Ma et al. 2015] Ma, J., Gao, W., Wei, Z., Lu, Y., and Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. In Proceedings of the 24th ACM International on Con on Information and Knowledge Management, CIKM '15, pages 1751–1754, New York, NY, USA. ACM.

[Mustafaraj and Metaxas 2017] Mustafaraj, E. and Metaxas, P. T. (2017). The fake news spreading plague:was it preventable? In Web Science Con, pages 236–239.

[Nasim et al. 2018] Nasim, M., Nguyen, A., Lothian, N., Cope, R., and Mitchell, L. (2018). Real-time detection of content polluters in partially observable twitter networks. In Companion Proceedings of the The Web Con 2018, WWW '18, pages 1331–1339, Republic and Canton of Geneva, Switzerland. International World Wide Web Con Steering Committee.

[Pérez-Rosas et al. 2018] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic detection of fake news. In International Conference on Computational Linguistics, pages 3391–3401.

[Qian et al. 2018] Qian, F., Gong, C., Sharma, K., and Liu, Y. (2018). Neural user response generator: Fake news detection with collective user intelligence. In International Joint Con on Artificial Intelligence, pages 3834–3840.

[Reis et al. 2019] Reis, J. C. S., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019). Explainable machine learning for fake news detection. In Proceedings of the 10th ACM Conference on Web Science, WebSci '19, pages 17–26, New York, NY, USA. ACM.

[Reis et al. 2019] Reis, J. C. S., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019). Supervised learning for fake news detection. IEEE Intelligent Systems, 34(2):76–81.

[Rubin et al. 2015] Rubin, V. L., Conroy, N. J., and Chen, Y. (2015). Towards news verification: Deception detection methods for news discourse.

[Ruchansky et al. 2017] Ruchansky, N., Seo, S., and Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM on Con on Information and Knowledge Management, CIKM '17, pages 797–806, New York, NY, USA. ACM.

[Santia and Williams 2018] Santia, G. C. and Williams, J. R. (2018). Buzzface: A news veracity dataset with facebook user commentary and egos. In AAAI Con on Web and Social Media, pages 531–540.

[Seo J. 2013] Seo J., Choi S., H. S. (2013). The method of trust and reputation systems based on link prediction and clustering. In IFIP International Con on Trust Management, pages 223–230.

[Sethi 2017] Sethi, R. J. (2017). Crowdsourcing the verification of fake news and alternative facts. In Proceedings of the 28th ACM Con on Hypertext and Social Media, HT '17, pages 315–316, New York, NY, USA. ACM.

[Sharma et al. 2019] Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. ACM Trans. Intell. Syst. Technol., 10(3):21:1–21:42.

[Sherchan et al. 2013] Sherchan, W., Nepal, S., and Paris, C. (2013). A survey of trust in social networks. ACM Comput. Surv., 45(4):47:1–47:33.

[Shu et al. 2019a] Shu, K., Mahudeswaran, D., and Liu, H. (2019a). Fakenewstracker: A tool for fake news collection, detection, and visualization. Comput. Math. Organ. Theory, 25(1):60–71.

[Shu et al. 2018] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2018). Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. In arXiv.

[Shu et al. 2017a] Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017a). Fake news detection on social media: A data mining perspective. SIGKDD Explor. Newsl., 19(1):22–36.

[Shu et al. 2017b] Shu, K., Wang, S., and Liu, H. (2017b). Exploiting tri-relationship for fake news detection. In arXiv.

[Shu et al. 2019b] Shu, K., Wang, S., and Liu, H. (2019b). Beyond news contents: The role of social context for fake news detection. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM '19, pages 312–320, New York, NY, USA. ACM.

[Srivastava et al. 2018] Srivastava, A., Kannan, R., Chelmis, C., and Prasanna, V. K. (2018). Factcheck: Keeping activation of fake news at check. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '18, pages 2079–2081, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[Tschiatschek et al. 2018] Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., and Krause, A. (2018). Fake news detection in social networks via crowd signals. In Companion Proceedings of the The Web Con 2018, WWW '18, pages 517–524, Republic and Canton of Geneva, Switzerland. International World Wide Web Con Steering Committee.

[Vavilis et al. 2014] Vavilis, S., PetkoviÄ‡, M., and Zannone, N. (2014). A reference model for reputation systems. Decision Support Systems, 61:147 – 154.

[Vo and Lee 2018] Vo, N. and Lee, K. (2018). The rise of guardians: Fact-checking url recommendation to combat fake news. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, pages 275–284, New York, NY, USA. ACM.

[Vosoughi et al. 2017] Vosoughi, S., Mohsenvand, M. N., and Roy, D. (2017). Rumor gauge: Predicting the veracity of rumors on twitter. ACM Trans. Knowl. Discov. Data, 11(4):50:1–50:36.

[Wang et al. 2018a] Wang, P., Angarita, R., and Renna, I. (2018a). Is this the era of misinformation yet: Combining social bots and fake news to deceive the masses. In Companion Proceedings of the The Web Con 2018, WWW '18, pages 1557–1561, Republic and Canton of Geneva, Switzerland. International World Wide Web Con Steering Committee.

[Wang 2017] Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

[Wang et al. 2018b] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. (2018b). Eann: Event adversarial neural networks for multi-modal fake news detection. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, pages 849–857, New York, NY, USA. ACM.

[Woloszyn and Nejdl 2018] Woloszyn, V. and Nejdl, W. (2018). Distrustrank: Spotting false news domains. In Proceedings of the 10th ACM Con on Web Science, WebSci '18, pages 221–228, New York, NY, USA. ACM.

[Wu and Liu 2018] Wu, L. and Liu, H. (2018). Tracing fake-news footprints: Characterizing social media messages by how they propagate. In Proceedings of the Eleventh ACM International Con on Web Search and Data Mining, WSDM '18, pages 637–645, New York, NY, USA. ACM.

[Yang et al. 2019] Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E. D., Ji, S., and Hu, X. B. (2019). Xfake: Explainable fake news detector with visualizations. In The World Wide Web Conference, WWW '19, pages 3600–3604, New York, NY, USA. ACM.

[Zhang et al. 2018] Zhang, Q., Yilmaz, E., and Liang, S. (2018). Ranking-based method for news stance detection. In Companion Proceedings of the The Web Con 2018, WWW '18, pages 41–42, Republic and Canton of Geneva, Switzerland. International World Wide Web Con Steering Committee.

[Zhou and Zafarani 2018] Zhou, X. and Zafarani, R. (2018). Fake news: A survey of research, detection methods, and opportunities. In arXvi.

[Zhou and Zafarani 2019] Zhou, X. and Zafarani, R. (2019). Fake news detection: An interdisciplinary research. In Companion Proceedings of The 2019 World Wide Web Conference, WWW '19, pages 1292–1292, New York, NY, USA. ACM.

[Zhou et al. 2019] Zhou, X., Zafarani, R., Shu, K., and Liu, H. (2019). Fake news: Fundamental theories, detection strategies and challenges. In Proceedings of the Twelfth ACM International Con on Web Search and Data Mining, WSDM '19, pages 836–837, New York, NY, USA. ACM.

Capítulo 1

Ecossistemas de Dados na Web: da teoria aos desafios

Marcelo Iury S. Oliveira, Bernadette Farias Lóscio

Abstract

A number of initiatives have been developed to share and consume data on the Web. The growing interest in these initiatives drove the emergence of Data on Web Ecosystems, which provide an environment for the creation and management of initiatives of data sharing, as well as providing the necessary support to ensure the sustainability of such initiatives. These ecosystems promise a range of benefits for their participants, such as creating new business opportunities, generating innovation and creating value from the publication and consumption of data on the Web. In this course the main concepts related to this new environment are discussed, including theoretical relevant aspects as well as challenges and research opportunities in this area.

Resumo

Uma série de iniciativas vem sendo desenvolvidas em todo o mundo objetivando o compartilhamento e consumo dos dados na Web. O crescente interesse nessas iniciativas motivou o surgimento dos Ecossistemas de Dados na Web, os quais fornecem um ambiente propício para a criação e o gerenciamento de iniciativas de compartilhamento de dados, bem como oferecem o suporte necessário para garantir a sustentabilidade de tais iniciativas. Esses ecossistemas promovem uma série de benefícios para seus participantes, tais como a criação de novas oportunidades de negócios, a geração de inovação e a criação de valor a partir da publicação e do consumo dos dados na Web. Neste contexto, este minicurso discute os principais conceitos relacionados a este novo ambiente, abordando aspectos relevantes, tanto do ponto de vista teórico quanto de desafios e oportunidades de pesquisa nesta área.

1.1. Introdução

Uma série de iniciativas vem sendo desenvolvidas em todo o mundo objetivando o compartilhamento e consumo dos dados na Web. O crescente interesse nessas iniciativas

motivou o surgimento dos Ecossistemas de Dados na Web, os quais fornecem um ambiente propício para a criação e o gerenciamento de iniciativas de compartilhamento de dados, bem como oferecem o suporte necessário para garantir a sustentabilidade de tais iniciativas [Oliveira and Lóscio, 2018; Zuiderwijk et al., 2014].

Como visualizado na Figura 1.1, Ecossistema de Dados na Web são compostos por redes de atores autônomos que consomem, produzem ou fornecem, direta ou indiretamente, dados e outros recursos relacionados a dados publicados na Web (*e.g.*, software, serviços e infraestrutura) [Oliveira and Lóscio, 2018]. Em um Ecossistema de Dados, um ator pode desempenhar um ou mais papéis e pode estar conectado a outros atores por meio de relacionamentos, de modo que a colaboração e a competição dos atores promovem a autorregulação do ecossistema [Oliveira and Lóscio, 2018].

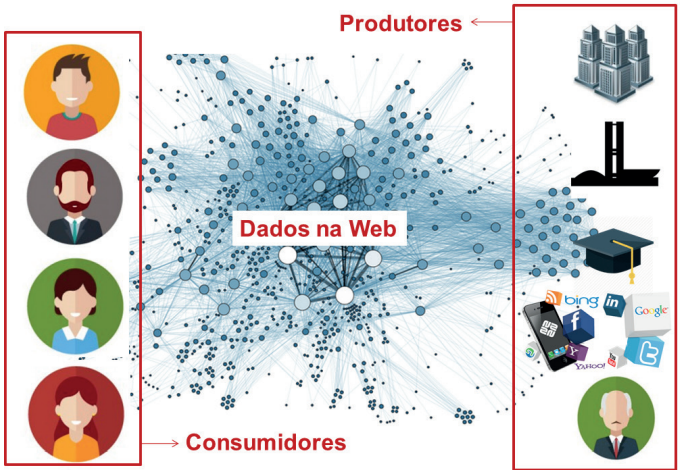

Figura 1.1. Ecossistema de Dados na Web. Fonte:Autores

A pesquisa em Ecossistemas de Dados na Web tem apresentado um crescente interesse, seja por parte da academia, seja pela indústria [Oliveira et al., 2019]. Os governos estão abrindo seus conjuntos de dados com objetivo seja para atender a preceitos democráticos (*e.g.*, transparência e prestação de contas) ou para permitir a melhoria nos seus serviços. Muitas empresas privadas tem liberado dados com objetivo de melhorar sua imagem e até aumentar os lucros, permitindo que grupos de pessoas analisem informações e obtenham resultados valiosos. Outras empresas tem devotado esforços na criação de mercados de dados para facilitar o comércio e o compartilhamento de dados, serviços e outros recursos. Um exemplo é o Bloomberg Marketplace[1] que reúne dados de uma variedade de fontes e provedores, organizando-os e disponibilizando-os para clientes que pagam por serviços de curadoria de dados .

Ecossistemas de Dados na Web não devem ser confundidos com a simples publicação de dados. De acordo com Pollock [2011], usualmente, o modelo básico atual para o fornecimento e uso de dados é uma via de mão única. Não há um retorno por dos parte usuários de dados e consumidores de dados, *i.e.*, os usuários não compartilham dados e conhecimento com os produtores de dados. Em um cenário ideal, os usuários deveriam contribuir com os produtores de dados. Essa contribuição pode incluir um simples

[1]https://www.bloomberg.com/professional/product/market-data/

feedback de avaliação dos dados, sinalização de erros, envio de correções ou mesmo a contribuição com conjuntos de dados limpos e integrados.

De fato, todos podem e devem colaborar, quer sejam usuários ou produtores de dados. Inclusive, Ubaldi [2013] argumenta que os benefícios esperados do compartilhamento de dados somente serão alcançados por meio da criação de um Ecossistema de Dados nos quais produtores e consumidores de dados interagem entre si.

Um exemplo de um Ecossistema de Dados é uma comunidade de instituições relacionada a serviços e pesquisa de saúde. No setor de saúde, um grande volume de conjuntos de dados é produzido [Murdoch and Detsky, 2013]. Por exemplo, sistemas de prontuários eletrônicos coletam uma grande quantidade de dados [Murdoch and Detsky, 2013]. Cada paciente tem seu próprio registro digital, que inclui dados demográficos, histórico médico, alergias, resultados de exames laboratoriais, dentre outros. Os registros são compartilhados por meio de sistemas de informação e estão disponíveis para prestadores de serviços do setor público e privado. Na última década, empresas farmacêuticas e instituições de pesquisa também vêm acumulando dados de pesquisa e desenvolvimento em bancos de dados médicos. Pesquisadores e médicos podem usar essas grandes quantidades de dados a fim de auxiliar diagnósticos ou mesmo encontrar tendências e tratamentos que tenham as maiores taxas de sucesso no mundo real [Lebied, 2018]. Similarmente, hospitais podem usar dados de várias fontes para obter previsões diárias e horárias de quantos pacientes são esperados em um intervalo de tempo específico. Neste exemplo, os sistemas de prontuários eletrônicos, as empresas farmacêuticas e as instituições de pesquisa atuam como produtores de dados. Além disso, os pesquisadores, médicos e hospitais atuam como consumidores de dados. Os dados são, então, fornecidos por organizações públicas e privadas envolvidas em tratamento de saúde, planejamento urbano, monitoramento de tráfego e fiscalização de segurança. Cada um atuando de acordo com os seus objetivos e também esperando alcançar suas próprias metas.

Ecossistemas de Dados na Web promovem uma série de benefícios para os seus participantes, tais como a criação de novas oportunidades de negócios, inovação e a criação de valor a partir da publicação e do consumo dos dados na Web. No entanto, enquanto o potencial desses ecossistemas é real, a realização é malsucedida em muitos casos. Vários ecossistemas não são sustentáveis e, consequentemente, o esforço desempenhado pelos seus atores não é utilizado de forma adequada ou é desperdiçado. A falta de comunicação e cooperação entre os produtores de dados e os consumidores é um dos principais obstáculos para a obtenção de Ecossistemas de Dados sustentáveis. Além disso, projetar, desenvolver e manter sistemas adicionais para apoiar os ecossistemas de coleta de dados se tornam tarefas cada vez mais desafiadoras.

Neste contexto, este minicurso discute os principais conceitos relacionados a Ecossistemas de Dados na Web, abordando aspectos relevantes, tanto do ponto de vista teórico quanto de desafios e oportunidades de pesquisa nesta área. De maneira mais específica, serão discutidos os principais conceitos que descrevem um Ecossistemas de Dados na Web, soluções para modelagem dos ecossistemas, arquiteturas e modelos de negócio para Ecossistemas de Dados na Web, o papel de boas práticas para na sustentabilidade dos ecossistemas e os principais desafios de pesquisa existentes na área.

1.2. *Background* Teórico

A noção de ecossistemas origina-se nas áreas de Ecologia e Biologia. Os ecologistas usam o termo ecossistema (biológico) como uma unidade natural que consiste em todas as plantas, animais e micro-organismos em uma área que funciona em conjunto com todos os recursos físicos não vivos do ambiente [Wikipedia, 2001]. O conceito de Ecossistema de Dados também abrange ideias de outros ecossistemas, tais como Ecossistema de Negócios, Ecossistema Digital e Ecossistema de Software.

Moore [1999] define um Ecossistema de Negócios como uma comunidade econômica apoiada por um ator base que interage com organizações e indivíduos, incluindo clientes, produtores líderes, concorrentes e outras partes interessadas. Essas comunidades podem se unir de uma maneira parcialmente intencional, altamente auto-organizada e até acidental. Bem como, podem ser criadas em torno de atores-chave que são empresas dominantes com forte influência sobre os processos co-evolucionários. Os membros das comunidades também incluem fornecedores, produtores, concorrentes e outras partes interessadas, como organizações de financiamento, associações comerciais, órgãos normativos, sindicatos trabalhistas e instituições governamentais. Além disso, os Ecossistemas de Negócios são baseados em recursos essenciais que são explorados para produzir inovação, negócios disruptivos ou produtos essenciais.

Iansiti and Levien [2004] apresentam outra visão para Ecossistema de Negócios inspirada em Ecossistemas Biológicos. Segundo eles, um Ecossistema de Negócios é caracterizado por uma rede de negócios na qual vários atores interconectados dependem uns dos outros para a sobrevivência e eficiência mútua. Um Ecossistema de Negócios compreende todos os atores que contribuem para o desenvolvimento de processos que influenciarão a integridade de longo prazo da rede [Iansiti and Levien, 2004]. Se o ecossistema é saudável, atores individuais prosperam. Se o ecossistema não é saudável, os atores individuais sofrem profundamente.

O termo Ecossistema (de Negócios) Digitais foi proposto pela primeira vez em [Nachira, 2002; Nachira et al., 2007]. Tendo sido concebido principalmente como uma estratégia ou política de adoção efetiva de Tecnologias de Informação e Comunicação (TICs) na União Européia. Um Ecossistema Digital faria com que indústrias, setores ou regiões se tornassem mais inovadores e competitivos nos mercados globais. Um Ecossistema Digital é considerado uma evolução natural dos Ecossistemas de Negócios.

No entanto, o termo Ecossistema Digital tem sido usado para descrever uma variedade de conceitos. Por exemplo, Fiorina [2000] refere-se à infra-estrutura de rede existente relacionada à Internet, que envolve várias empresas que oferecem serviços digitais, bem como clientes para usar os serviços digitais oferecidos. O termo também está sendo cada vez mais associado à adoção de *e-business* e *e-commerce* para proporcionar maior crescimento, empregos cada vez mais qualificados e maior inclusão social. Outra perspectiva é apresentada em [Sondergaard, 2017] que define um Ecossistema Digital como uma comunidade de empresas e indivíduos interdependentes que compartilham plataformas digitais padronizadas para um propósito mutuamente benéfico (*e.g.*, inovação ou criação de valor).

Ecossistemas de Software também são um termo recente, referindo-se a organiza-

ções de rede ou indivíduos que baseiam suas relações no desenvolvimento, comércio e uso de uma tecnologia de software. Os Ecossistemas de Software quebram as fronteiras internas da linha de produção das organizações, permitindo contribuições de desenvolvedores externos, fornecedores e outras partes externas. Como consequência, os Ecossistemas de Software criam dependências que antes não existiam entre os componentes e as organizações associadas.

Nesse sentido, Jansen et al. [2009] definem Ecossistemas de Software como um conjunto de atores de negócios interagindo com um mercado compartilhado de software e serviços, juntamente com as relações entre eles. Na maioria dos casos, os Ecossistemas de Software são sustentados por plataformas tecnológicas ou mercado comuns com as relações sendo realizadas pela troca de informações, recursos e artefatos. Outra definição proposta por Bosch and Bosch-Sijtsema [2010] define um Ecossistema de Software como "um conjunto de soluções de software que permitem, suportam e automatizam as atividades e transações dos atores nos ecossistemas sociais ou de negócios associados e as organizações que fornecem essas soluções". Essa definição enfatiza mais a automação da atividade, bem como o interesse comum em software e seu uso.

Assim como os ecossistemas acima apresentados, um Ecossistema de Dados envolve múltiplos elementos, incluindo dados, software e outros recursos computacionais, fluxos de trabalho, pessoas, mercado, governo e infraestrutura. Esses elementos sugerem que um Ecossistema de Dados precisa combinar componentes de diferentes ecossistemas. Desta forma, um Ecossistema de Dados pode ser visto como um ecossistema híbrido, ou seja, que contempla características dos diferentes ecossistemas mencionados.

É preciso ressaltar que, apesar de compartilhar as características de rede e coevolução, os Ecossistemas de Dados também diferem dos ecossistemas anteriores. Ao contrário de outros ecossistemas, os Ecossistemas de Dados não dependem de uma plataforma comum explícita na qual diferentes atores possam colaborar. A plataforma comum é, na verdade, os vários conjuntos de dados disponibilizados, trocados e consumidos pelos atores. Em particular, os dados não precisam necessariamente ser fornecidos por um único ator. A falta de uma plataforma comum cria uma rede de oferta e demanda mais difusa. Outra diferença está relacionada a como os produtos são negociados entre os atores. Em Ecossistemas de Negócios, operações de negócios e atores são *per se* os produtos [Manikas and Hansen, 2013a]. Em Ecossistemas de Software, os produtos são componentes ou serviços de software. Em Ecossistemas de Dados, o principal produto são os dados e seus derivados.

1.3. Definições para Ecossistemas de Dados

Há pouco consenso sobre a nomenclatura e a definição de Ecossistemas de Dados. Embora uma discussão mais aprofundada sobre a terminologia esteja além do escopo deste minicurso, a fim de poder analisar o campo do Ecossistema de Dados, bem como guiar o processo de estudo, revisaremos algumas das definições existentes.

Primeiramente, é preciso destacar que um grande número de estudos não definem o termo Ecossistema de Dados. Segundo Oliveira et al. [2019], 13 dos 29 trabalhos encontrados na literatura não apresentam nenhuma definição para Ecossistemas de Dados. No entanto, alguns desses estudos fazem referência a estudos anteriores. Na maioria dos

casos, isso acontece porque estes estudos focam, de maneira secundária, em Ecossistema de Dados. Em outros casos, os autores consideram um estudo prévio (escrito por eles mesmos ou por outros pesquisadores) como a base para as definições necessárias.

Pollock [2011] fornece a definição mais antiga para um ecossistema de dados.. De acordo com ele, *"Um ecossistema tem ciclos de dados, nos quais consumidores e intermediários de dados (e.g., desenvolvedores de aplicativos) podem compartilhar seus dados limpos, integrados e empacotados no ecossistema de forma reutilizável. Geralmente, esses dados limpos e integrados são mais valiosos do que a fonte original"*. Essa definição enfatiza a necessidade de ciclos de dados para criar Ecossistemas de Dados. Além do ciclo, a visão de Pollock exige que os atores desempenhem papéis, como produtores, intermediários e consumidores.

Similarmente, de acordo com o Zubcoff et al. [2016], um Ecossistema de Dados é composto de muitos atores e pequenas estruturas organizacionais que devem reconhecer dados como a matéria-prima. Os atores devem formar um ciclo a fim de "alimentar" o ecossistema, proporcionando benefícios a todas as partes. Essa visão do Ecossistema de Dados também defende um ciclo, bem como aponta para a existência de múltiplos atores, cada um com suas próprias expectativas.

Por sua vez, Harrison et al. [2012] vislumbram a ideia de um Ecossistema Governamental, que é uma espécie de Ecossistema de Dados Abertos. Segundo eles, *"Um Ecossistema Governamental prevê organizações governamentais como atores centrais, tomando a iniciativa em redes organizadas para atingir objetivos específicos relacionados à inovação e governança"*. Eles também complementam essa definição, afirmando que a metáfora do ecossistema *é frequentemente usada por formuladores de políticas e acadêmicos [...] para transmitir um senso de um sistema social interdependente de atores, organizações, infraestruturas materiais e recursos simbólicos, todos suportados através do uso intensivo de tecnologia e informação"*. Assim como Pollock [2011], Harrison et al. [2012] defendem que os papéis devem ser definidos, mas também enfatizam a ideia de um papel fundamental que controla e coordena o ecossistema. Além disso, eles também reconhecem um conjunto de fatores contextuais (*e.g.*, aspecto social) como um elemento-chave de um Ecossistemas de Dados. Além disso, Harrison et al. [2012] combina a metáfora do ecossistema com o conceito de múltiplas e variadas inter-relações entre produtores, usuários, dados, infraestrutura material e instituições.

Uma perspectiva diferente é apresentada por Zuiderwijk et al. [2016], que define Ecossistemas de Dados como *"todas as atividades para liberar e publicar dados na Internet, para quais usuários de dados podem conduzir atividades como pesquisar, localizar, avaliar e visualizar dados e suas licenças relacionadas, limpeza, análise, enriquecimento, combinação, vinculação e visualização de dados e interpretação e discussão de dados e fornecimento de feedback ao produtor de dados e outras partes interessadas"*. Similarmente, o Ding et al. [2011] define um Ecossistemas de Dados como *"um sistema baseado em dados no qual stakeholders de diferentes tamanhos e funções encontram, gerenciam, arquivam, publicam, reutilizam, integram e consomem dados em conexão com ferramentas, serviços e ferramentas online"*. Ambas as definições apresentam a ideia de atividades que desenvolveriam algum valor ou benefício para os atores que usam dados. Essas atividades podem ser atribuídas a funções específicas que serão desempenhadas pelos atores

(*i.e.*, *stakeholders*).

Assim, de acordo com essas definições, os Ecossistemas de Dados contam com um vasto e heterogêneo conjunto de atores, cada um com propriedades, capacidades e expectativas diferentes. Os atores podem produzir e consumir recursos usando atividades diferentes e sob diferentes condições. Além disso, muitos desses elementos são dinâmicos e evoluem com o tempo. Podemos concluir que o panorama do Ecossistema de Dados é composto por um conjunto de atores que relacionam-se entre si por meio da troca de um conjunto de recursos distribuídos, heterogêneos, dinâmicos e em evolução.

1.4. Principais Componentes de Ecossistemas de Dados na Web

De acordo com Oliveira and Lóscio [2018]; Oliveira et al. [2018], apesar da abstração das características apresentadas na literatura, quatro elementos principais destacam-se (como apresentado na Figura 1.2), sendo eles: (1) atores, (2) papéis, (3) relacionamentos e (4) recursos.

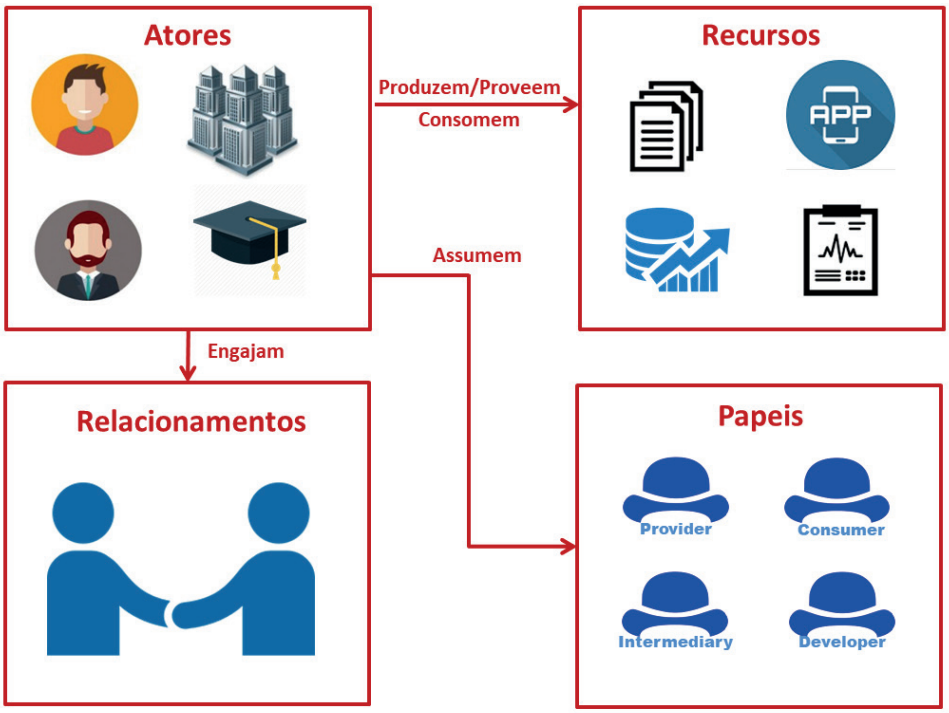

Figura 1.2. Elementos Básicos de Ecossistema de Dados na Web. Fonte:Autores

Um *Ator* é uma entidade autônoma, como uma empresa, instituição ou indivíduo, que desempenha um ou mais papéis específicos em um Ecossistema de Dados. Um conjunto de interesses motiva os atores e cada um deles tem diferentes expectativas e capacidades. Os atores geralmente se comprometem voluntariamente com o ecossistema, mas receber incentivos para ser atuarem no ecossistema.

Um *Papel* é uma função desempenhada por um ator em um Ecossistema de Dados. Está relacionado a um conjunto de deveres e atividades. Diversos papéis podem ser identificados no Ecossistema de Dados. Em geral, pelo menos os consumidores de dados

e os produtores de dados são identificados nos Ecossistemas de Dados. Porém, existem vários papéis adicionais responsáveis por diferentes tarefas e atividades. Além disso, os papéis podem ou não sobrepor suas responsabilidades.

Atores vinculados a um papel devem possuir a capacidade de cumprir os compromissos que um papel lhes impõe. Além das capacidades, os atores também exigem recursos adequados para fornecer e consumir dados. Exemplos comuns de recursos são conjuntos de dados, serviços, ferramentas, capital financeiro, bem como capital humano, equipamentos, materiais e tecnologia proprietária.

Um *Relacionamento* representa interações entre os atores do Ecossistema de Dados. Os relacionamentos, geralmente, são baseados em um interesse comum ou também estão relacionados ao papel que cada ator desempenha no ecossistema. No mais, eles podem variar de acordo com aspectos econômicos, políticos, culturais e/ou tecnológicos. Finalmente, os relacionamentos são restritos por fatores físicos ou não físicos (*e.g.*, recursos) [Bosch, 2009], assim como produzem valor e envolvem custos.

Um *Recurso* é um produto útil e valioso ou uma capacidade produzida, fornecida, gerenciada ou consumida pelos atores. Em Ecossistemas de Dados, os recursos variam desde conjuntos de dados e software baseado em dados até a infraestrutura. Em particular, o software baseado em dados inclui ativos reutilizáveis (componentes e serviços) ou ativos de software (aplicativos) usados para consumir, produzir ou fornecer dados. Os recursos podem ser trocados individualmente ou em combinação por meio de transações de relacionamentos. Os recursos geralmente estão em conformidade com um conjunto de padrões que também são limitados por um conjunto de licenças, que definem como o recurso pode ser explorado. Além disso, os recursos são geralmente avaliados de acordo com diferentes propriedades de qualidade.

Em relação a recursos, Zuiderwijk et al. [2015] os distinguem entre três categorias: recursos humanos, recursos de dados e recursos de TI. Recursos humanos referem-se a indivíduos que usam seus recursos para explorar dados. Os recursos de dados se referem aos ativos baseados em dados estáticos e dinâmicos, como bancos de dados, bases de conhecimento ou simplesmente conjuntos de dados. Os recursos de TI referem-se ao hardware (*e.g.*, infraestruturas, redes e computadores), plataformas e aplicativos (software). Na verdade, os atores não precisam necessariamente possuir, gerenciar ou operar os recursos subjacentes, mas podem consumir ou contratar esses recursos por meio de outros atores, como prestadores de serviços ou outros tipos de atores intermediários.

Ainda em relação aos elementos que compõem um Ecossistema de Dados, Mercado Lara and Gil-Garcia [2014] agrupam os elementos em três domínios: (i) políticas e práticas governamentais, (ii) inovadores; uma combinação de tecnologia, negócios e governo e (iii) usuários, sociedade civil e negócios. Shin and Choi [2015] também identificam como elementos-chave de um Ecossistema de Dados: (i) infra-estrutura, (ii) software e tecnologias, (iii) serviço e aplicações, (iv) normas, (v) usuários, (vi) fatores sociais e culturais, (vii) governo e (viii) indústria.

De uma forma geral, todos os elementos do Ecossistema de Dados estão interconectados de forma que quando um elemento é alterado, os efeitos podem ser sentidos em todo o sistema [Immonen et al., 2014]. De fato, os atores afetam e são afetados pela

criação e entrega de recursos executados pelos outros atores. Além disso, os interesses dos atores podem levar a conflitos. Por exemplo, os consumidores de dados são fortemente influenciados pela decisão de um produtor de dados de não publicar ou atualizar um determinado dado, alterar o formato em que os dados são publicados, comprometer a qualidade dos dados ou alterar como ele pode ser usado.

1.5. Modelagem em Ecossistemas de Dados na Web

As técnicas de modelagem desempenham um papel importante no suporte ao design e ao desenvolvimento de sistemas complexos. Em geral, os modelos permitem compartilhar uma visão e conhecimento comuns entre as partes interessadas técnicas e não técnicas, facilitando e promovendo a comunicação entre elas. Além disso, os modelos tornam o planejamento do projeto mais eficaz e eficiente, proporcionando uma visão mais apropriada do sistema a ser desenvolvido e permitindo que o controle do projeto seja alcançado de acordo com critérios objetivos Brambilla et al. [2012]; OMG [2016].

Na área de Ecossistemas de Dados, modelos conceituais tanto permitem práticas de gerenciamento mais eficientes quanto auxiliam na descrição do conhecimento sobre os recursos e outras características de um Ecossistema de Dados. Além disso, modelos conceituais podem ser usados como linguagem de metamodelos para desenvolver ferramentas CASE (engenharia de software auxiliada por computador) para ajudar os profissionais a construir modelos conceituais derivados que representariam Ecossistemas de Dados específicos (*e.g.*, Ecossistemas de Dados Biomédicos e Ecossistemas de Dados Financeiros).

Diante deste contexto, em Oliveira et al. [2018], é apresentado uma proposta de meta-modelo que define os conceitos fundamentais de Ecossistema de Dados e seus relacionamentos para permitir a análise e descrição destes ecossistemas.

A Figura 1.3 apresenta os principais elementos do meta-modelo proposto por Oliveira et al. [2018]. Este meta-modelo foi formalizado por meio do Eclipse Modeling Framework (EMF) [2] e na linguagem de meta-modelagem ECore fornecida pelo EMF. O meta-modelo formaliza os conceitos Ator, Relacionamento, Papel e Recursos apresentados na seção anterior. Além disso, ele introduz novos conceitos como Contexto, Elementos Contextuais, Modelos de Negócios, Transações e entre outros.

O meta-modelo proposto por Oliveira et al. [2018] foi o primeiro voltado especificamente para Ecossistemas de Dados. O mesmo, apesar de formalizar conceitos importantes, ainda carece de diagramas que permitam a praticantes não-técnicos visualizarem de forma gráfica Ecossistemas de Dados. Diante deste contexto, trabalhos de outras áreas poderiam ser adaptados para o contexto de Ecossistemas de Dados de forma a permitir o desenho gráfico do estado de um ecossistemas. Por exemplo, os trabalhos [Brinkkemper et al., 2009; Yu and Deng, 2011] propõem soluções para modelagem de Ecossistemas de Software. Estas soluções permitem identificar os atores e suas relações, assim como fornecem uma visão rápida dos ecossistemas modelados.

Brinkkemper et al. [2009] apresenta um modelo que consiste em dois diagramas: PCD e SSN. O *Product Context Diagram* (PCD) descreve o contexto de um produto de software. Já o diagrama Supply Network (SSN) descreve as diferentes partes envolvi-

[2] http://www.eclipse.org/modeling/emf/

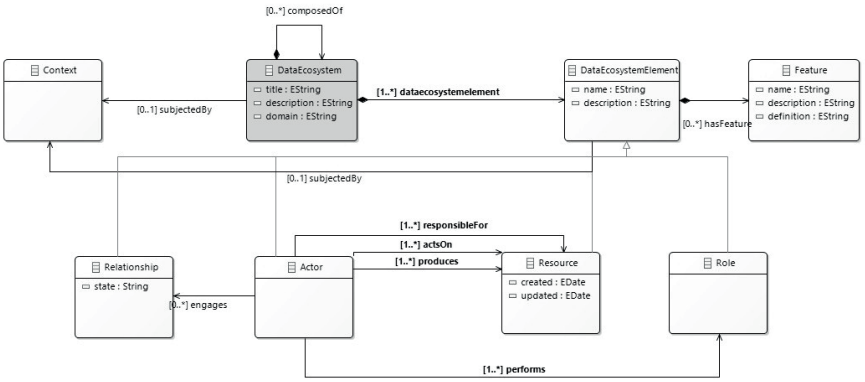

Figura 1.3. Meta-Modelo para descrição de Ecossistemas de Dados

das na entrega e implementação de um produto de software ou serviço. Por sua vez, Yu and Deng [2011] apresenta uma abordagem de modelagem estratégica baseada na estrutura de modelagem i* para ajudar a entender os ecossistemas de software. Os modelos i* são usados para descrever dependências estratégicas entre fornecedores de software, desenvolvedores de terceiros e usuários finais, além de ajudar a explorar e raciocinar sobre formas alternativas de atingir metas estratégicas para cada ator. Ambos os trabalhos podem ser adaptados para mapeamento dos atores de um Ecossistema de Dados na Web.

1.6. Papéis em Ecossistemas de Dados na Web

Diversos papéis podem ser identificados em Ecossistemas de Dados. Inclusive, é possível identificar dois ou mais papéis compartilhando as mesmas tarefas. Além disso, as características do Ecossistema de Dados podem levar à necessidade de definição de papéis mais específicos. Por exemplo, em Ecossistemas de Dados baseados em dados médicos/de saúde, geralmente há papéis responsáveis por avaliar questões éticas.

O papel mais frequentemente identificado é o usuário de dados, responsável pelo consumo direto ou indireto de dados. Esse papel é apresentado com uma miríade de nomes nos estudos, por exemplo, usuários finais, consumidores de dados, beneficiários de dados, dentre outros. Os usuários de dados não têm necessariamente a capacidade de consumir dados diretamente dos produtores de dados. Eles geralmente dependem de serviços fornecidos por *Re-Users* (usuários que criam valor a partir dos dados disponíveis na Web), intermediários de dados ou provedores de serviços. Além disso, os usuários de dados geralmente representam os usuários finais de um ecossistema.

O segundo papel mais destacado é o provedor de dados, responsável pela publicação ou fornecimento de dados. Existem também alguns estudos que apresentam papéis menores relacionados ao fornecimento de dados. Por exemplo, o Immonen et al. [2014] apresentam papéis secundários relacionados ao fornecimento de dados, como "Armazenador para coletar e salvar dados, um Desenvolvedor para gerenciar e processar dados, um Agregador para combinar e editar dados de diferentes fontes, um Harmonizador para padronizar e homogeneizar dados de diferentes fontes, um Atualizador para atualizar informações, um Editor para publicar os dados e um Registro para manter a administração

de recursos de dados".

É importante notar que os provedores de dados não são necessariamente responsáveis pela geração de dados. Essa responsabilidade pode ser atribuída a outro papel, chamado Produtor ou Criador de Dados, responsável pela captura ou geração de dados. Esse papel também pode compilar, agregar e empacotar dados.

Outro papel identificado é um Re-User responsável por agregar valor aos dados a serem reutilizados. De acordo com o Köster and Suárez [2016], o Re-User é responsável pelo uso de dados para desenvolver aplicativos ou serviços destinados a usuários de dados.

O papel de ator-chave (do inglês, *Keystone Actor*) é responsável por impulsionar as forças por trás do ecossistema, além de fornecer estabilidade em ambientes instáveis [Iansiti and Levien, 2004]. Esse papel é muito comum em Ecossistemas de Software. Em Dawes et al. [2016], os pesquisadores afirmaram que, em Ecossistemas de Dados, os atores-chave são responsáveis por fornecer a maioria dos dados, bem como por promover o ecossistema. Lee [2014] afirmam que esse papel deve ser atribuído a atores que lideram os programas de Dados Abertos Governamentais.

Embora tenhamos apresentado alguns papéis, ainda há uma miríade de diferentes papéis na literatura. Oliveira et al. [2019] realizaram um levantamento na literatura de papéis relacionados a Ecossistemas de Dados. No total, foram identificados 13 papéis primários e mais 22 papéis secundários (*i.e.*, um papel especializado que é responsável por alguns dos deveres e atividades de um papel primário). No entanto, os papéis são definidos de forma superficial. Na maioria dos estudos, os papeis sequer são especificados. Vários dos estudos apenas listam os atores do Ecossistema de Dados, deixando para leitor a responsabilidade de identificar e classificar suas funções.

1.7. Estruturas de Organização de Ecossistemas de Dados na Web

Em um Ecossistema de Dados, cada ator é conectado a outros atores por um conjunto de interesses ou modelos de negócios. Toda a rede de relacionamentos pode seguir uma estrutura organizacional, variando de uma abordagem difusa para uma abordagem centrada em um ator-chave. A organização de um Ecossistema de Dados leva em conta tanto a maneira como os atores estão conectados e as propriedades de seus relacionamentos (por exemplo, dependência) [Manikas and Hansen, 2013b]. Estudar a forma de organização dos Ecossistemas de Dados é importante para entender e governar a interação e participação dos atores [Christensen et al., 2014].

De acordo com Oliveira et al. [2019], é possível identificar na literatura 5 diferentes abordagens de organização, sendo elas: centrada em ator-chave, baseada em intermediários, centradas em plataforma, baseada em marketplace e orientada a modelos de negócio.

Na organização centrada em ator-chave, os atores são organizados em torno de um ator-chave, que é direta ou indiretamente responsável por fornecer grande parte dos dados. Contudo, o ator-chave não tem controle completo sobre os outros atores. Eles podem sair (ou entrar) no ecossistema a qualquer momento [Heimstädt et al., 2014]. Já a organização baseada em intermediários depende da presença de atores que atuem como intermediários de dados para gerar valor a partir dos dados.

Na organização centrada em plataforma, uma plataforma fornece infraestrutura e serviços para suportar o fornecimento e o consumo de dados. Dawes et al. [2016]; Ding et al. [2011] enfatizam que os custos de fornecimento de dados são reduzidos quando os dados são liberados por meio da plataforma. A plataforma também pode atenuar problemas de interoperabilidade e usabilidade. Ferramentas de catalogação de dados abertos (*e.g*, CKAN [3]) tem sido comumente utilizadas para criar Ecossistemas de Dados na Web.

Em organizações baseadas em marketplaces, os marketplaces fornecem infraestrutura, modelos de negócios, regras e serviços necessários para transações de dados e software entre os atores [Smith et al., 2016]. Em geral, os marketplaces abrangem uma plataforma técnica com capacidade de vincular produtores de dados e usuários de dados. Eles também permitem a venda de dados, serviços e aplicativos.

Apesar de não definir como os atores devem ser organizados, alguns estudos apresentam modelos de negócios, que descrevem a lógica de como um ator cria, entrega e captura valor. Em particular, valor refere-se a qualquer benefício que um agente obtém do Ecossistema de Dados, como satisfação, utilidade, solução de problemas ou receita.

1.8. Gerenciamento de Ecossistemas de Dados na Web

Enquanto o potencial dos Ecossistemas de Dados é real, a sua realização é malsucedida em muitos casos [Heimstädt et al., 2014; Mercado Lara and Gil-Garcia, 2014; Zuiderwijk et al., 2012]. De acordo com o Dawes et al. [2016], como consequência de inúmeras barreiras e limitações, o desempenho das iniciativas de Ecossistema de Dados tende a ser simplista. Várias iniciativas se concentram em lançar e promover concursos de curta duração, como *hackathons* e *code fests*. Até os aplicativos desenvolvidos para esses concursos apresentam resultados insatisfatórios [Gama and Lóscio, 2014]. A maioria das aplicações nesses cenários acaba sendo rapidamente abandonada.

O estabelecimento correto de Ecossistema de Dados significa a coordenação adequada de várias categorias de atores, a provisão de apoio às empresas, o estímulo ao desenvolvimento e uso/reúso de recursos [Koznov et al., 2016]. Outros elementos essenciais para um Ecossistema de Dados de sucesso são a colaboração entre atores, a integração de informações, a preservação dos processos e o gerenciamento adaptativo [Zuiderwijk et al., 2014].

De fato, o gerenciamento de Ecossistemas de Dados é importante para facilitar ativamente o funcionamento efetivo e cumprir as metas dos ecossistemas [Mercado Lara and Gil-Garcia, 2014]. Além disso, se um Ecossistema de Dados não possui uma estrutura de gerenciamento, torna-se difícil impulsionar o ecossistema, construir e aprender com experiências passadas [Lee, 2014].

Os Ecossistemas de Dados funcionarão bem somente se forem projetados considerando sua complexidade por completo. De acordo com Mercado Lara and Gil-Garcia [2014], o gerenciamento de um Ecossistema de Dados requer o esboço de alguns tópicos básicos, que se concentram em (i) identificar os atores mais ativos que atuam como componentes essenciais do ecossistema; (ii) analisar a natureza das transações que ocorrem entre esses atores; (iii) reconhecer quais recursos são necessários para cada ator e como

[3] https://ckan.org/

eles envolvem as transações; e (iv) estudar os indicadores que sinalizam o status da atividade do ecossistema [Mercado Lara and Gil-Garcia, 2014]. Portanto, essas considerações exigem uma abordagem mais sistêmica para planejar, projetar e coordenar Ecossistemas de Dados.

No entanto, até o momento, as iniciativas de gerenciamento de Ecossistemas de Dados são simplistas. Por exemplo, vários governos desenvolveram programas e políticas que visam promover o fornecimento e o uso de dados do setor público [Koznov et al., 2016; Mercado Lara and Gil-Garcia, 2014]. Tais políticas muitas vezes se concentram em garantir a disponibilidade e a qualidade dos recursos de dados. Essas políticas de gerenciamento ajudaram a expandir os Ecossistemas de Dados e melhorar o fornecimento de dados. No entanto, essas políticas não incluem outros atores-chave, como consumidores de dados e intermediários de dados, que realmente demandam o fornecimento. Por isso, é crucial incluir, desde o início, o ponto de vista de todos os atores do ecossistema. Uma gestão integrada e colaborativa deve garantir que as metas incluídas em uma agenda de Ecossistema de Dados atendam às necessidades, direitos e interesses de todos os atores que fazem parte do ecossistema [Köster and Suárez, 2016].

Soluções como modelos de maturidade, *frameworks* e metodologias propostas para outras áreas podem ser adaptadas para o contexto de Ecossistemas de Dados. Tais soluções propõem diversos instrumentos de gerenciamento que poderiam ser usados para um melhor monitoramento e sistematização das atividades de Ecossistemas de Dados. Entre essas soluções destacamos: DMBOK, ISO 8000 e DMM.

O DMBOK (do inglês, *Data Management Body of Knowledge*) consiste de um corpo de conhecimento que compila um conjunto de processos e práticas para servir como um guia abrangente para as atividades de gerenciamento de dados. O DMBOK foi desenvolvido em 2009 com a colaboração de mais de 120 profissionais. Ele fornece uma visão geral do gerenciamento de dados, além de fornecer definições de processos de gerenciamento de dados, funções e resultados de entrega e sua terminologia padrão.

O DMBOK possui 10 funções principais de gerenciamento de dados, conforme apresentado na Figura 1.4. Tomando a governança de dados como o núcleo, ela narra o escopo de cada função pela rotação no sentido horário. Cada uma das funções é detalhada como [Mosley et al., 2010]:

- Data Governance: é responsável pelo planejamento, supervisão e controle de alto nível do gerenciamento de ativos de dados;
- Data Architecture Management: é responsável por definir o plano de gerenciamento de dados para atender às necessidades de dados corporativos. Essa função inclui o desenvolvimento e a manutenção da arquitetura de dados corporativos, dentro do contexto de toda a arquitetura corporativa;
- Data Development: é responsável por projetar, implementar e manter soluções para atender às necessidades de dados da empresa. Inclui análise de demanda de dados, implementação, teste, manutenção e outras soluções;
- Data Operation Management: é responsável pelo planejamento, controle e suporte

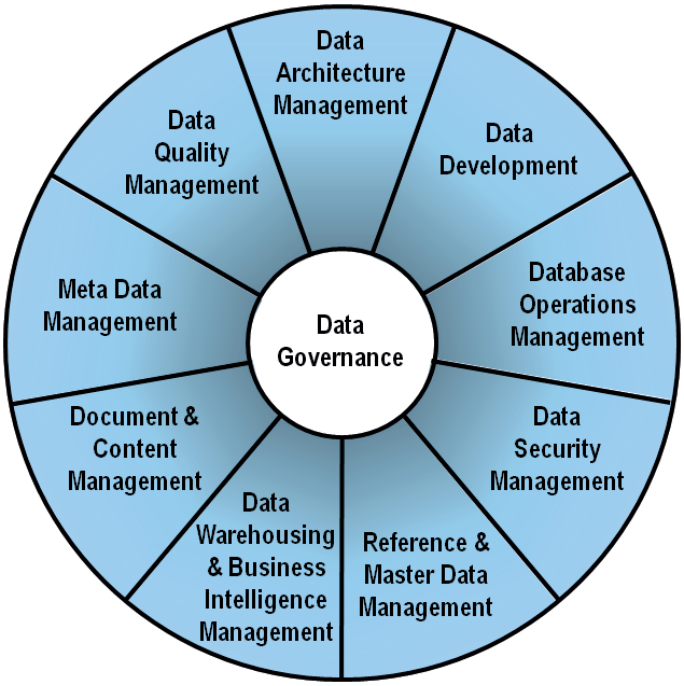

Figura 1.4. Mandala de Funções do DMBOK. Fonte:[Mosley et al., 2010]

do ciclo de vida dos dados, desde a aquisição de dados ao arquivamento e eliminação de dados;

- Data Security Management: é responsável por planejar políticas e medidas de segurança para garantir a confidencialidade dos dados e os direitos de acesso hierárquico.

- Reference and Master Data Management: é responsável pelas atividades de planejamento, implementação e controle para garantir a consistência dos valores dos dados.

- Data Warehouse e o Business Intelligence Management: é responsável pelo planejamento, implementação e controle de processos para fornecer dados de suporte a decisões e suporte para trabalhadores do conhecimento envolvidos em relatórios, consultas e análises;

- Document and Content Management: é responsável pelo gerenciamento de arquivos eletrônicos e registros físicos (incluindo texto, gráficos, imagens, áudio e vídeo);

- Metadata Management: é responsável por integrar, controlar e fornecer metadados de alta qualidade;

- Data Quality Management: é responsável pelo planejamento, implementação e controle de atividades que aplicam técnicas de gerenciamento de qualidade para medir, avaliar, melhorar e garantir a adequação dos dados para uso.

Por usa vez, a família de padrões ISO 8000 fornece um conjunto de estruturas para melhorar a qualidade dos dados para tipos específicos de dados [ISO, 2011]. A ISO 8000 está sendo aplicada em vários setores industriais e países em todo o mundo [4]. A ISO 8000 abrange características de qualidade de dados industriais durante todo o ciclo de vida do produto, desde a concepção até o descarte [ISO, 2011]. Ele também descreve o vocabulário e os recursos, bem como define os requisitos para troca padrão e garantia de qualidade dos dados. A força da ISO 8000 reside no fato de que suas definições sobre dados de qualidade são baseadas em acordos internacionais.

Outra solução de gerenciamento que pode ser utilizada para a criação de métodos e processos de gerenciamento é o modelo DMM (do inglês, *Data Management Maturity*) Institute [2014]. O DMM é uma estrutura abrangente de práticas de gerenciamento de dados em seis categorias principais que ajuda as organizações a avaliar suas capacidades, identificar pontos fortes e lacunas e alavancar seus ativos de dados para melhorar o desempenho dos negócios. O DMM se concentra em 5 áreas de processo: Estratégia de Gerenciamento de Dados, Governança de Dados, Qualidade de Dados, Operações de Dados e Arquitetura de Dados.

Usando como base soluções desenvolvidas para outros tipos de ecossistemas, Baars and Jansen [2012] e Albert [2014] propuseram *frameworks* voltados para habilitação de governança em Ecossistemas de Software. Baars and Jansen [2012] descrevem um *framework* para a análise da governança de Ecossistemas de Software. O *framework* permite que organizações possam analisar e melhorar sua governança de Ecossistemas de Software de maneira estruturada, levando a um melhor desempenho e saúde do ecossistema. Albert [2014] propõem a abordagem SECOGov (Software Ecosystems Governance) que permite gerir e analisar informações sob a visão de Ecossistemas de Software. Apesar destas solução não terem sido concebidas especificamente para gestão de Ecossistemas de Dados, parte de suas práticas podem ser adaptadas para o contexto de Ecossistemas de Dados.

1.9. Criação de Valor e Modelos de Negócio em Ecossistemas de Dados

Em Ecossistemas de Dados, os atores são obrigados a empregar um conjunto de capacidades e recursos para gerar valor. De acordo com o Magalhaes et al. [2014], muitas vezes o ônus é reservado aos consumidores que devem extrair valor dos recursos disponíveis. Isso cria um problema, já que o consumidor médio usualmente não possui as habilidades necessárias [Zuiderwijk et al., 2012]. Devido a essas barreiras, o valor não deve ser criado apenas por um único ator, mas sim por uma cadeia de valor (*i.e.*, em uma rede de atores). Uma cadeia de valor é um conjunto de atividades independentes de valor agregado que é usada para explorar um conjunto de recursos. Além disso, uma cadeia de valor consiste em diferentes atores conduzindo uma ou mais atividades (*e.g.*, provisão de dados, curadoria de dados, análise de dados), e cada atividade pode consistir em um número de ações ou técnicas de criação de valor (*e.g.*, coleta, visualização, criação de serviço).

Em Ecossistemas de Dados, a cadeia de valor mínima consiste em produtores de dados, intermediários de dados e consumidores de dados [Heimstädt et al., 2014]. Como as atividades de valor agregado oferecem complexidade diferente, é possível que cada

[4]https://www.dataqualitypro.com/iso-8000-data-quality-certification-options/

ação possa consistir em uma ou mais cadeias de valor [Attard et al., 2016].

A introdução de incentivos e recompensas também pode estimular o fluxo de recursos e geração de valor em um Ecossistema de Dados [Lindman et al., 2016; Moiso and Minerva, 2012]. De fato, a produção, provisão e exploração de recursos do Ecossistema de Dados precisam de investimentos [Lindman et al., 2016; Moiso and Minerva, 2012]. Moiso and Minerva [2012] afirmam que, sem recursos financeiros, torna-se muito difícil sustentar iniciativas de Ecossistema de Dados. No entanto, há pouco incentivo para investir em recursos e capacidades. A falta de conhecimento sobre os benefícios do compartilhamento de dados e a falta de novos modelos de operação são os principais impedimentos que explicam porque os atores, principalmente as empresas privadas, não estão motivados a se engajar em Ecossistemas de Dados [Immonen et al., 2014].

Portanto, é importante desenvolver modelos de negócios sustentáveis que proporcionem um incentivo para manter os dados atualizados e acessíveis e, além disso, criar aplicativos e ferramentas comerciais sustentáveis [Immonen et al., 2014]. Modelos de negócios apoiam a proposição de valor para atores em um ecossistema. Como apresentado na Figura 1.5, um modelo de negócio envolve diversos aspectos, tais como estrutura de custo, capacidade, modelo de lucro, dentre outros aspectos.

Diversos modelos de negócios aplicáveis a dados foram descritos na literatura [Teece, 2010; Zott and Amit, 2010], tais como cobrança por suporte, modelo de assinatura e oferta de serviços/consultoria. Serviços e aplicativos podem ser cobrados com base em funcionalidades, custo ou por meio do modelo *pay-per-use* [Immonen et al., 2014]. Os recursos de dados também podem ser precificados usando o modelo de assinatura, no qual o consumidor paga um preço fixo por um determinado período de tempo. Outro modelo adequado é o fluxo de receitas múltiplas do Flickr [Teece, 2010], pois envolve a coleta de taxas de assinatura, cobrança de publicidade contextual para os anunciantes e o recebimento de patrocínio e taxas de participação nos lucros das parcerias.

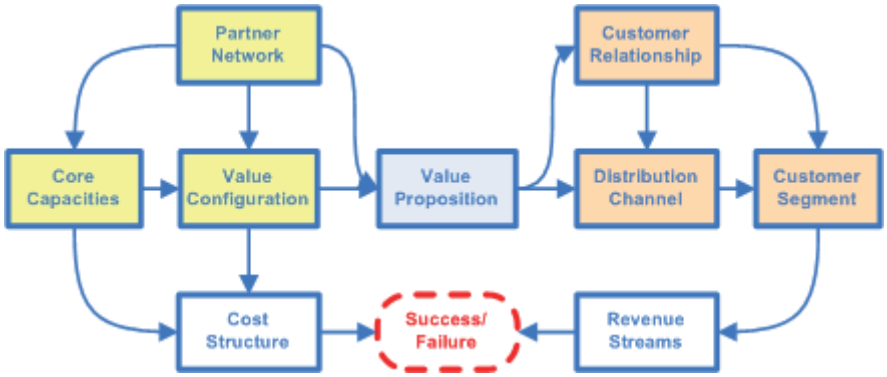

Figura 1.5. Modelo de Negócio. Fonte:https://bit.ly/2U89mLM

Outro exemplo é o modelo de negócios em nuvem, que é adequado no caso de grandes conjuntos de dados pelo fato do armazenamento, processamento e análise usualmente exigir uma grande quantidade de recursos. Além disso, como Ecossistemas de Dados são dinâmicos, os modelos de negócios devem tolerar a imprevisibilidade. Isso se deve ao fato de que os Ecossistemas de Dados ainda são incertos por natureza e não

é possível prever ou garantir que alguns recursos estarão disponíveis por tempo longo o suficiente [Zuiderwijk et al., 2015].

1.10. Boas Práticas para Publicação de Dados na Web

O sucesso de Ecossistemas de Dados, entre outros fatores, dependem diretamente da qualidade dos dados que estão providos e consumidos pelos atores. Além disso, segundo Lóscio et al. [2015], a heterogeneidade dos dados e a falta de padrões para descrição e acesso aos conjuntos de dados tornam o processo de publicação, compartilhamento e consumo uma tarefa complexa. De maneira geral, torna-se necessário publicar dados de forma que possam ser facilmente compreendidos e utilizados por consumidores, além da disponibilização dos dados em formatos que possam ser facilmente processados por aplicações.

Desse modo, em busca de alternativas que possibilitem um entendimento comum entre os atores desse contexto, o W3C criou um grupo de trabalho denominado *Data on the Web Best Practices*. Esse grupo teve como objetivo propor uma recomendação que servisse como um guia para a publicação e o consumo dos dados na Web.

Tabela 1.1. Desafios da publicação de dados na Web. Fonte: Derilinx et al. [2015]

Desafio	Descrição
Metadados	Permitir que os seres humanos entendam os metadados, interpretando a natureza e a estrutura dos dados, e que as máquinas também possam processá-los
Licença	Permitir que os seres humanos compreendam as informações da licença e que as máquinas possam detectar automaticamente
Proveniência	Permitir que os seres humanos conheçam a origem ou o histórico do conjunto de dados e que as máquinas possam processar automaticamente tais informações
Qualidade	Documentar a qualidade dos dados, para facilitar o processo de seleção dos conjuntos de dados e chances de reutilização
Versionamento	Permitir que versões dos dados sejam geradas e seja possível o acesso a cada versão
Identificação	Fornecer identificadores únicos para os conjuntos de dados e distribuições
Formato	Escolher formatos que permitam o uso e o reuso
Vocabulários	A fim de melhorar a interoperabilidade e manter terminologia comum entre os produtores e consumidores
Acesso	Permitir o fácil acesso aos dados usando a infraestrutura da Web tanto para seres humanos quanto para máquinas
Preservação	A fim de indicar corretamente se os dados foram removidos ou arquivados
Feedback	Receber feedback dos consumidores e assegurar que os dados atendam as suas necessidades
Enriquecimento	Enriquecer, melhorar ou refinar os dados brutos agregando valor
Republicação	Permitir que os dados utilizados possam ser republicados

Para alcançar esse objetivo, o grupo de trabalho do W3C, selecionou um conjunto de casos de uso[5] que representam cenários de como os dados são publicados e consumidos na Web. Com esses casos de uso, foi possível identificar os principais desafios enfrentados por produtores e consumidores de dados (ver Quadro 1.1), assim como um conjunto de requisitos necessários para a publicação. A partir dos desafios e requisitos encontrados nesses casos de usos, foi desenvolvido o documento de Boas Práticas para Publicação de Dados na Web (*Data on the Web Best Practices* - DWBP).

De acordo com Lóscio et al. [2016], as Boas Práticas para Dados na Web foram

[5]https://www.w3.org/TR/dwbp-ucr

desenvolvidas para oferecer orientação técnica para a publicação de dados na Web, contribuindo para melhorar a relação entre publicadores e consumidores de dados. Além disso, elas são independentes de domínio e aplicação, ou seja, é aplicável a todos os domínios, podendo ainda ser estendidas ou complementadas com outros documentos ou normas mais especializadas. Para cada desafio apresentado no Quadro 1.1 foram propostas uma ou mais práticas. No total, são 35 boas práticas que discursam sobre diferentes aspectos relacionados à publicação e consumo de dados, como acesso aos dados, identificadores, metadados, formatos, dentre outros. O Quadro 1.2 apresenta as boas práticas estipuladas para cada desafio encontrado.

Tabela 1.2. Boas Práticas para publicação de dados na Web. Fonte: Lóscio et al. [2017]

Desafio	Boas Práticas
Metadados	BP1 - Fornecer metadados BP2 - Fornecer metadados descritivos BP3 - Fornecer metadados estruturais
Licença	BP4 - Fornecer informações de licenciamento de dados
Proveniência e Qualidade	BP5 - Fornecer informações sobre a proveniência dos dados BP6 - Fornecer informações sobre a qualidade dos dados
Versionamento	BP7 - Fornecer um indicador de versão BP8 - Fornecer histórico de versão
Identificação	BP9 - Utilizar URIs constantes como identificadores de conjuntos de dados BP10 - Utilizar URIs constantes como identificadores dentro dos conjuntos de dados BP11 - Designar URIs para versões e séries de conjuntos de dados
Formato	BP12 - Utilizar formatos de dados padronizados inteligíveis por máquinas BP13 - Utilizar representações de dados de localidade neutra BP14 - Fornecer dados em formatos múltiplos
Vocabulários	BP15 - Reutilizar vocabulários preferencialmente padronizados BP16 - Escolher o nível correto de formalização
Acesso	BP17 - Fornecer download em massa BP18 - Fornecer subconjuntos para conjuntos de dados extensos BP19 - Utilizar a negociação de conteúdo para disponibilizar dados em formatos múltiplos BP20 - Fornecer acesso em tempo real BP21 - Fornecer dados atualizados BP22 - Fornecer uma justificativa para dados não disponíveis BP23 - Disponibilizar dados por meio de uma API BP24 - Utilizar padrões da Web como base para as APIs BP25 - Fornecer a documentação completa para sua API BP26 - Evitar modificações que quebrem sua API
Preservação	BP27 - Preservar os identificadores BP28 - Avaliar a cobertura do conjunto de dados
Feedback	BP29 - Coletar feedback de consumidores de dados BP30 - Disponibilizar feedback
Enriquecimento	BP31 - Enriquecer dados por meio da geração de novos dados BP32 - Fornecer apresentações complementares
Republicação	BP33 - Fornecer feedback ao editor original BP34 - Seguir os termos de licenciamento BP35 - Citar a publicação original

Conforme apresentado por Lóscio et al. [2017], cada boa prática (BP) tem um resultado esperado com sua aplicação e possíveis formas de implantação da prática. Além disso, são descritas a motivação para o seu uso e quais testes podem ser realizados para verificar se a prática foi implementada de forma adequada. Ao final, ainda apresenta as evidências que comprovam a relevância da prática e os benefícios que serão alcançados com o seu uso.

1.11. Desafios em Ecossistemas de Dados na Web

O objetivo deste minicurso foi apresentar pesquisas relevantes e atuais sobre Ecossistemas de Dados, bem como fornecer uma visão geral do campo. Além de fornecer uma visão geral do campo Ecossistema de dados, também consideramos importante apresentar vários aspectos que ainda não são abordados na literatura.

Por exemplo, há poucos estudos apresentando e analisando possíveis modelos de negócios que permitam a criação e agregação de valor em Ecossistemas de Dados. Também há uma falta de linguagens de modelagem ou diagramas para representar os Ecossistemas de Dados. Atores de negócios e usuários não-técnicos podem enfrentar algumas dificuldades ao avaliar um modelo de negócios ou a organização estrutural de um Ecossistema de Dados. Uma nova linguagem de modelagem baseada em uma notação gráfica pode fornecer a capacidade de entender aspectos e processos importantes de um Ecossistema de Dados, além de oferecer aos atores a capacidade de comunicar esses aspectos e processos de maneira padrão. Por enquanto, as únicas opções existentes foram projetadas para outros tipos de ecossistemas, o que pode dificultar o uso pleno.

Modelos e soluções de avaliação para validar a saúde dos Ecossistemas de Dados são outra lacuna de pesquisa. Esses modelos devem fornecer os meios para avaliar a funcionalidade e o status dos elementos em um Ecossistema de Dados. A saúde de um ecossistema depende em parte de uma variedade de fatores, incluindo os atores e como eles agem, relacionamentos, políticas e a infraestrutura disponível. De fato, parte da definição de eficácia e sucesso de um Ecossistema de Dados está na determinação e utilização de métricas para medir sua saúde.

O desenvolvimento de métodos para governança e controle mais eficazes de Ecossistemas de Dados é outra lacuna na pesquisa. O funcionamento de um Ecossistema de Dados depende da atividade e interação de um conjunto de diferentes atores. Esse comportamento difuso tem como consequência a diminuição do controle e o aumento resultante dos desafios associados ao planejamento e manutenção. Portanto, os métodos de gerencia de Ecossistema de Dados podem fornecer uma estrutura comum na forma de regras, procedimentos, protocolos e processos bem definidos para desenvolver, coordenar e evoluir os Ecossistemas de Dados.

Existem ainda vários desafios técnicos/não técnicos relacionados ao uso de dados em um Ecossistema de Dados, incluindo a complexidade das atividades necessárias para identificar, entender e usar dados, falta de recursos e conhecimento técnico entre os atores Janssen et al. [2012]; Zuiderwijk and Janssen [2014]. Desafios adicionais incluem proveniência de dados, qualidade de dados (*e.g.*, validade, integridade e pontualidade), fornecimento de metadados e interoperabilidade, bem como preocupações com privacidade e confidencialidade Janssen et al. [2012]; Zuiderwijk and Janssen [2014]. De um modo geral, a proposta de soluções para enfrentar alguns desses desafios pode aliviar o ônus para os atores, principalmente os que consomem dados, e, consequentemente, promover sua participação nos Ecossistemas de Dados.

1.12. Conclusão

Este minicurso teve como principal objetivo apresentar e contribuir para disseminação do conceito de Ecossistemas de Dados. De forma a atender este objetivo, apresentamos tanto pesquisas relevantes sobre Ecossistemas de Dados, bem como fornecemos uma visão geral do campo Oliveira et al. [2019]. A pesquisa em Ecossistemas de Dados, de uma forma geral, ainda está em estágio inicial, tendo recebido mais atenção em blogs, relatórios, poucas dezenas de artigos. Além disso, a literatura em Ecossistemas de Dados é focada em certos tipos de dados (por exemplo, dados governamentais ou dados científicos), ou não fornece uma descrição clara sobre como um Ecossistema de Dados deve ser.

Como já observado, o campo de Ecossistemas de Dados é inspirado no trabalho realizado em outros ecossistemas, como os Ecossistemas de Negócios e Software. De fato, é difícil definir os limites entre esses diferentes tipos de ecossistemas. Apesar da relação, a literatura em Ecossistema de Dados não parece examinar o trabalho realizado em ecossistemas de outras áreas. Por exemplo, em [Oliveira et al., 2019], foram identificados apenas 2 trabalhos analisando conceitos dos Ecossistemas de Software. Nesse sentido, a análise de outros campos permitiria o uso de suas teorias e soluções que poderiam, por sua vez, alavancar as pesquisas no Ecossistema de Dados.

O correto estudo de Ecossistemas de Dados abrange uma ampla variedade de disciplinas, incluindo projeto, técnica, orquestração, gerenciamento e avaliação. Neste sentido, identificamos vários aspectos que não são abordados na literatura, entre eles destacamos: esquemas para classificação de papéis dos atores, métodos de governança, soluções de engenharia e gerenciamento de Ecossistemas de Dados, plataformas e outras ferramentas de suporte, dentre outras lacunas. É preciso destacar que não existem muitos trabalhos acadêmicos relacionados aos Ecossistemas de Dados. Na maioria dos casos, eles estão focados em algum aspecto específico, como modelos de negócios ou uma solução que reflete apenas um pequeno fragmento de toda a área de pesquisa.

Referências

B Albert. *SECOGov: Um Modelo de Governança de Ecossistemas de Software para Apoiar Atividades de Arquitetura de TI*. PhD thesis, Dissertação. COPPE/UFRJ, Rio de Janeiro, Brasil, 2014.

Judie Attard, Fabrizio Orlandi, and Sören Auer. Data value networks: Enabling a new data ecosystem. In *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*, pages 453–456. IEEE, 2016.

Alfred Baars and Slinger Jansen. A framework for software ecosystem governance. In *International conference of software business*, pages 168–180. Springer, 2012.

Jan Bosch. From software product lines to software ecosystems. In *Proceedings of the 13th international software product line conference*, pages 111–119. Carnegie Mellon University, 2009.

Jan Bosch and Petra Bosch-Sijtsema. From integration to composition: On the impact of software product lines, global development and ecosystems. *Journal of Systems and Software*, 83(1):67–76, 2010.

Marco Brambilla, Jordi Cabot, and Manuel Wimmer. Model-driven software engineering in practice. *Synthesis Lectures on Software Engineering*, 1(1):1–182, 2012.

Sjaak Brinkkemper, Ivo Van Soest, and Slinger Jansen. Modeling of product software businesses: Investigation into industry product and channel typologies. In *Information Systems Development*, pages 307–325. Springer, 2009.

Henrik Bærbak Christensen, Klaus Marius Hansen, Morten Kyng, and Konstantinos Manikas. Analysis and design of software ecosystem architectures–towards the 4S telemedicine ecosystem. *Information and Software Technology*, 56(11):1476–1492, 2014.

Sharon S Dawes, Lyudmila Vidiasova, and Olga Parkhimovich. Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly*, 33(1):15–27, 2016.

Deirdre Lee Derilinx, Bernadette Farias Lóscio, and Phil Archer. Data on the web best practices use cases and requirements. https://www.w3.org/TR/dwbp-ucr/, feb 2015. Acessado em 31 de agosto de 2019.

Li Ding, Timothy Lebo, John S Erickson, Dominic DiFranzo, Gregory Todd Williams, Xian Li, James Michaelis, Alvaro Graves, Jin Guang Zheng, Zhenning Shangguan, et al. Twc logd: A portal for linked open government data ecosystems. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(3):325–333, 2011.

Carleton Fiorina. The digital ecosystem. https://goo.gl/eSWuxN, 2000. Acessado em 31 de agosto de 2019.

Kiev Gama and Bernadette Farias Lóscio. Towards ecosystems based on open data as a service. In *International Conference on Information Systems*, pages 659–664, 2014.

Teresa M Harrison, Theresa A Pardo, and Meghan Cook. Creating open government ecosystems: A research and development agenda. *Future Internet*, 4(4):900, 2012.

Maximilian Heimstädt, Fredric Saunderson, and Tom Heath. Conceptualizing open data ecosystems: A timeline analysis of open data development in the UK. In *CeDEM14: Conference for E-Democracy an Open Government*, page 245. MV-Verlag, 2014.

Marco Iansiti and Roy Levien. *The keystone advantage: what the new dynamics of business ecosystems mean for strategy, innovation, and sustainability*. Harvard Business Press, 2004.

Anne Immonen, Marko Palviainen, and Eila Ovaska. Requirements of an open data based business ecosystem. *IEEE Access*, 2:88–103, 2014.

CMMI Institute. Data management maturity, 2014. URL `https://cmmiinstitute.com/data-management-maturity`. Acessado em 31 de agosto de 2019.

ISO. Iso 8000 – the international standard for data quality, 2011. URL `https://www.iso.org/standard/50798.html`. Acessado em 31 de agosto de 2019.

Slinger Jansen, Sjaak Brinkkemper, and Anthony Finkelstein. Business network management as a survival strategy: A tale of two software ecosystems. In *Proceedings of the 1st International Workshop on Software Ecosystems*, page 34, 2009.

Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. Benefits, adoption barriers and myths of open data and open government. *Information systems management*, 29 (4):258–268, 2012.

Victoria Köster and Gustavo Suárez. Open data for development: Experience of uruguay. In *Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance*, pages 207–210. ACM, 2016.

D. Koznov, O. Andreeva, U. Nikula, A. Maglyas, D. Muromtsev, and I. Radchenko. A survey of open government data in Russian Federation. In *IC3K 2016 - Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, volume 3, pages 173–180, 2016.

Mona Lebied. 12 examples of big data analytics in healthcare that can save people. https://www.datapine.com/blog/big-data-examples-in-healthcare/, 2018. Acessado em 31 de agosto de 2019.

Deirdre Lee. Building an open data ecosystem: an Irish experience. In *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*, pages 351–360. ACM, 2014.

Juho Lindman, Tomi Kinnari, and Matti Rossi. Business roles in the emerging open-data ecosystem. *IEEE Software*, 33(5):54–59, 2016.

Bernadette Lóscio, Caroline Guimarães, and Newton Calegari. Data on the web best practices: Challenges and benefits. *Open Data Reserach Symposium (ODRS 2016)*, 2016.

Bernadette Farias Lóscio, Marcelo Iury S Oliveira, and Ig Ibert Bittencourt. Publicação e Consumo de Dados na Web: Conceitos e Desafios. *Tópicos em Gerenciamento de Dados e Informações (Mini Cursos - SBBD 2015)*, pages 39–69, 2015. URL `http://dexl.lncc.br/sbbd2015/anais/ShortCourses.pdf`.

Bernadette Farias Lóscio, C Burle, and N Calegari. Data on the Web Best Practices. https://www.w3.org/TR/dwbp/, Janeiro 2017. Acessado em 31 de agosto de 2019.

Gustavo Magalhaes, Catarina Roseira, and Laura Manley. Business models for open government data. In *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*, pages 365–370. ACM, 2014.

Konstantinos Manikas and Klaus Marius Hansen. Reviewing the health of software ecosystems–a conceptual framework proposal. In *Proceedings of the 5th International Workshop on Software Ecosystems (IWSECO)*, pages 33–44, 2013a.

Konstantinos Manikas and Klaus Marius Hansen. Software ecosystems–a systematic literature review. *Journal of Systems and Software*, 86(5):1294–1306, 2013b.

Eunice Mercado Lara and J Ramon Gil-Garcia. Open government and data intermediaries: the case of AidData. In *Proceedings of the 15th Annual International Conference on Digital Government Research*, pages 335–336. ACM, 2014.

Corrado Moiso and Roberto Minerva. Towards a user-centric personal data ecosystem the role of the bank of individuals' data. In *Intelligence in Next Generation Networks (ICIN), 2012 16th International Conference on*, pages 202–209. IEEE, 2012.

James F. Moore. *Creating Value in the Network Economy*, chapter Predators and Prey: A New Ecology of Competition, pages 121–141. Harvard Business School Press, Boston, MA, USA, 1999. ISBN 0-87584-911-3.

Mark Mosley, Michael H Brackett, Susan Earley, and Deborah Henderson. *DAMA guide to the data management body of knowledge*. Technics Publications, 2010.

Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. *Jama*, 309(13):1351–1352, 2013.

Francesco Nachira. Towards a network of digital business ecosystems. `www.digital-ecosystems.org/doc/discussionpaper.pdf`, 2002. Acessado em 31 de agosto de 2019.

Francesco Nachira, Paolo Dini, and Andrea Nicolai. *A network of digital business ecosystems for Europe: roots, processes and perspectives*. European Commission, Bruxelles, 1 edition, 2007. ISBN 92-79-01817-5. Acessível em http://www.digital-ecosystems.org/book/.

Marcelo Iury S. Oliveira and Bernadette Farias Lóscio. What is a data ecosystem? In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, dg.o '18, pages 74:1–74:9, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-6526-0.

Marcelo Iury S. Oliveira, Lairson Emanuel R. A. Oliveira, Marlos G. Ribeiro Batista, and Bernadette Farias Lóscio. Towards a meta-model for data ecosystems. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, dg.o '18, pages 72:1–72:10, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-6526-0.

Marcelo Iury S Oliveira, Glória de Fátima Barros Lima, and Bernadette Farias Lóscio. Investigations into data ecosystems: a systematic mapping study. *Knowledge and Information Systems*, pages 1–42, 2019.

OMG. Meta object facility (mof) core specification v2.5.1. `http://www.omg.org/spec/MOF/2.5.1/PDF`, 2016. Acessado em 31 de agosto de 2019.

Rufus Pollock. Building the (open) data ecosystem. https://bit.ly/2DddJ1k, 2011. Acessado em 31 de agosto de 2019.

Dong-Hee Shin and Min Jae Choi. Ecological views of big data: Perspectives and issues. *Telematics and Informatics*, 32(2):311–320, 2015.

Göran Smith, Hosea Ayaba Ofe, and Johan Sandberg. Digital service innovation from open data: exploring the value proposition of an open data marketplace. In *System Sciences (HICSS), 2016 49th Hawaii International Conference on*, pages 1277–1286. IEEE, 2016.

Peter Sondergaard. Ecosystems are the future of digital. http://www.gartner.com/technology/topics/digital-ecosystems.jsp, 2017. Acessado em 31 de agosto de 2019.

David J Teece. Business models, business strategy and innovation. *Long Range Planning*, 43(2-3):172–194, 2010.

Barbara Ubaldi. Open government data: Towards empirical analysis of open government data initiatives. Technical report, Organisation for Economic Cooperation and Development (OECD), 2013. Acessado em 31 de agosto de 2019.

Wikipedia. Ecosystem. https://en.wikipedia.org/wiki/Ecosystem, 2001. Acessado em 31 de agosto de 2019.

Eric Yu and Stephanie Deng. Understanding software ecosystems: A strategic modeling approach. In *IWSECO-2011 Software Ecosystems 2011. Proceedings of the Third International Workshop on Software Ecosystems. Brussels, Belgium*, pages 65–76, 2011.

Christoph Zott and Raphael Amit. Business model design: an activity system perspective. *Long Range Planning*, 43(2-3):216–226, 2010.

Jose Jacobo Zubcoff, Llorenç Vaquer, Jose-Norberto Mazón, Francisco Maciá, Irene Garrigós, Andrés Fuster, and Jose Vicente Carcel. The university as an open data ecosystem. *International Journal of Design & Nature and Ecodynamics*, 11(3):250–257, 2016.

Anneke Zuiderwijk and Marijn Janssen. Barriers and development directions for the publication and usage of open data: A socio-technical view. In *Open Government*, pages 115–135. Springer, 2014.

Anneke Zuiderwijk, Marijn Janssen, Sunil Choenni, Ronald Meijer, R Sheikh Alibaks, and R Sheikh_Alibaks. Socio-technical impediments of open data. *Electronic Journal of e-Government*, 10(2):156–172, 2012.

Anneke Zuiderwijk, Marijn Janssen, and Chris Davis. Innovation with open data: Essential elements of open data ecosystems. *Information Polity*, 19(1, 2):17–33, 2014.

Anneke Zuiderwijk, Marijn Janssen, Kostas Poulis, and Geerten van de Kaa. Open data for competitive advantage: insights from open data use by companies. In *Proceedings of the 16th Annual International Conference on Digital Government Research*, pages 79–88. ACM, 2015.

Anneke Zuiderwijk, Marijn Janssen, Geerten van de Kaa, and Kostas Poulis. The wicked problem of commercial value creation in open data ecosystems: Policy guidelines for governments. *Information Polity*, 21:223–236, 2016.

Capítulo 1

Aprendizado de máquina e inferência em Grafos de Conhecimento

Daniel N. R. da Silva (LNCC), Artur Ziviani (LNCC) e Fabio Porto (LNCC)

Abstract

The increasing production and availability of massive and heterogeneous data bring forward challenging opportunities. Among them, the development of computing systems capable of learning, reasoning, and inferring facts based on prior knowledge. In this scenario, knowledge bases are valuable assets for the knowledge representation and automated reasoning of diverse application domains. Especially, inference tasks on knowledge graphs (knowledge bases' graphical representations) are increasingly important in academia and industry. In this short course, we introduce machine learning methods and techniques employed in knowledge graph inference tasks as well as discuss the technical and scientific challenges and opportunities associated with those tasks.

Resumo

A crescente produção e disponibilização de dados caracterizados por heterogeneidade e larga escala apresentam oportunidades desafiadoras à nossa sociedade. Dentre elas, como construir sistemas computacionais capazes de aprender, raciocinar e realizar inferências sobre fatos a partir de conhecimento prévio é uma tarefa relevante. Nesse cenário, bases de conhecimento são ativos importantes na representação e raciocínio automatizado do conhecimento de diversos domínios de aplicação. Em especial, a inferência de informação a partir de sua representação em rede — grafos de conhecimento — ganhou notoriedade na academia e indústria nos últimos anos. Em face ao exposto, neste curso, é apresentada uma introdução aos métodos e técnicas de aprendizado de máquina utilizadas em tarefas de inferência em grafos de conhecimento, discutindo-se os desafios e oportunidades tecnológicas e científicas desse tipo de tarefa.

1.1. Introdução

A representação computacional de conhecimento remonta ao nascimento da área de Inteligência Artificial. Ela é motivada pela necessidade de que a informação sobre o mundo

esteja descrita em uma forma processável e compreensível aos sistemas artificiais inteligentes [van Harmelen et al. 2008]. Nesse contexto, a representação de conhecimento na forma de uma rede tem atraído interesse da academia e indústria recentemente [Bonatti et al. 2019, Noy et al. 2019]. Esse tipo de representação, que remete ao surgimento de redes semânticas na década de 1960 [Lehmann 1992], ganhou novo fôlego no início dos anos 2010 na forma de grafos de conhecimento (*knowledge graphs*).[1]

Grafos de conhecimento têm se estabelecido como um arcabouço relevante para representação de conhecimento [Noy et al. 2019]. Eles fornecem uma estrutura semântica adequada para que sistemas computacionais sejam capazes de processar o conhecimento, assim como proveem uma representação próxima à linguagem natural. Esses grafos representam o conhecimento por meio da descrição de objetos (nós) e conexões (arestas) entre eles, sendo frequentemente imposto um esquema ou ontologia a esses objetos e conexões. Em geral, os nós desses grafos simbolizam entidades e classes do domínio de interesse, isto é, objetos do mundo real e categorias a que eles pertencem. Por sua vez, as arestas representam asserções sobre as entidades e classes de interesse. Em particular, uma aresta é usualmente disposta na forma de uma tripla (s, r, o), a qual indica que um tipo de relação r existe entre as entidades (e/ou classes) s e o.

Há várias maneiras de construir grafos de conhecimento. Eles podem ser produto de um processo de curadoria [Lenat 1995, Baker et al. 1998], de iniciativas de *crowdsourcing* [Vrandečić and Krötzsch 2014], extraídos a partir de bases contendo informação semiestruturada [Lehmann et al. 2015, Vrandečić and Krötzsch 2014], ou mesmo informação não estruturada [Dong et al. 2014]. Seja qual for a metodologia utilizada, o resultado da construção frequentemente está longe de ser perfeito [Paulheim 2017, Ratner et al. 2018]. Isso se deve a diversos fatores, incluindo a falta de informação digital sobre entidades de interesse e o processo, sujeito a falhas, empregado na construção desses grafos. A imperfeição inerente ao seu processo de construção implica diretamente na qualidade e utilidade de um grafo de conhecimento.

A qualidade e utilidade de grafos de conhecimento é atrelada a no mínimo três características: recentidade (*freshness*), exatidão (*correctness*) e completude (*coverage* ou *completeness*) [Paulheim 2017, Noy et al. 2019]. Recentidade diz respeito a se o conhecimento do grafo é atual, i.e., quão atualizada é a informação que ele contém. Já exatidão tange a se o grafo contém informação acurada, i.e., se essa informação retrata aquilo que é verdade. Por fim, completude tange a quanto do conhecimento de interesse está expresso no grafo. A dificuldade em se atender de forma abrangente a cada uma dessas três características promove a realização de tarefas de melhoria de grafos de conhecimento, isto é, de seu refino.

Tarefas de refino visam melhorar a qualidade de grafos de conhecimento ao inferir e adicionar conhecimento faltante ou ao identificar e remover erros. Nos últimos anos, essa tarefa tem sido abordada de forma desacoplada da construção de grafos de conhecimento. Por um lado, a construção é vista como um conjunto de operações (e.g., um *pipeline* analítico), realizadas sobre fontes de dados, que produzem um grafo de conhecimento. Por outro lado, o refino assume que os métodos de correção e/ou complementação serão aplicados em um grafo já existente. Perceba que o desacoplamento dessas duas ta-

[1] https://www.blog.google/products/search/introducing-knowledge-graph-things-not/

refas tem permitido o desenvolvimento de métodos de refino independentes de grafos de conhecimento [Paulheim 2017].

Técnicas de aprendizado de máquina são cada vez mais utilizadas no processo de refino de grafos de conhecimento, em particular na complementação desses grafos [Wang et al. 2017]. Em geral, essas abordagens mapeiam a complementação de um grafo de conhecimento em tarefas de aprendizado (supervisionado). Em outras palavras, um método de aprendizado é ajustado ao grafo de conhecimento a fim de realizar inferências de acordo com a tarefa mapeada. Por exemplo, se o objetivo da complementação é a adição de relacionamentos ao grafo, um classificador binário pode ser ajustado ao grafo de modo que na presença de uma tripla não observada, o classificador a atribua um escore de plausibilidade ou probabilidade relativo ao seu valor verdade. De acordo com esse escore, a tripla é ou não adicionada ao grafo de conhecimento.

Neste capítulo é apresentado sucintamente o emprego de técnicas de aprendizado de máquina em tarefas relacionadas a grafos de conhecimento; sobretudo na sua complementação. O capítulo está organizado da seguinte forma. Na Seção 1.2 é introduzido o emprego de grafos de conhecimento como forma de representação de conhecimento. Na Seção 1.3 são apresentados modelos de dados e sistemas para organização de grafos de conhecimento. Na Seção 1.4 são apresentadas algumas aplicações em que grafos de conhecimento têm sido empregados. Na Seção 1.5 são apresentadas tarefas em grafos de conhecimento, em particular, direcionadas a construção e complementação de grafos de conhecimento. Na Seção 1.6 discute-se como técnicas de aprendizado de máquina relacional são utilizadas em tarefas de complementação de grafos de conhecimento; em particular, técnicas baseadas no aprendizado de representações vetoriais. Por fim, na Seção 1.7 são feitas as considerações finais, ressaltando-se algumas oportunidades de pesquisa.

1.2. Contextualização

A popularidade recente de grafos de conhecimento tem trazido consigo ambiguidade ao termo. Tanto na indústria quanto na academia, o termo grafo de conhecimento é empregado, ora similarmente, ora distintamente de outros termos tais como base orientadas a grafos, base de conhecimento, ontologia e sistemas baseados em conhecimento [Ehrlinger and Wöß 2016]. Disto isto, a seguir são traçados alguns paralelos entre esses termos a fim de mostrar suas similaridades e diferenças.

Tradicionalmente, o termo base de conhecimento refere-se a estruturas informacionais destinadas à representação explícita de um domínio de conhecimento [Brodie and Mylopoulos 1986]. Além de conter asserções relativas às entidades desse domínio, essa classe de base é provida de uma semântica formal expressa em elementos como axiomas, definições e regras. Por exemplo, no domínio da biologia, uma base de conhecimento poderia conter um conjunto de fatos como *"O espécime Bo01 é uma borboleta Morpho Menelaus"* e axiomas como *"Toda borboleta é um inseto"*. Por causa da teoria que a base abarca (*"Toda borboleta é um inseto"*), mesmo que ela não contenha explicitamente o fato *"Bo01 é um inseto"*, ela o representa implicitamente.

A representação do conhecimento em uma base pode fazer uso de uma ontologia. Ontologias podem ser definidas como uma descrição formal dos conceitos (e.g., pessoa, localização e campo de estudo) — também chamados de classes — de um domínio de dis-

curso [Noy and Mcguinness 2001]. Em particular, conceitos contêm propriedades (e.g., a idade de uma pessoa) cujo domínio de valores é usualmente restrito. Conceitos também são arranjados em uma hierarquia taxonômica (e.g., insetos são animais) e participam de relações (e.g., pessoas podem trabalhar em empresas). Além dos conceitos, para que se permita processos de raciocínio e inferência, descreve-se regras e axiomas em ontologias. Esses processos são executados por um motor de raciocínio, acoplado à base, capaz de compreender a representação de conhecimento adotada.

Frequentemente, é possível dividir bases de conhecimento em dois componentes, um terminológico e um assertivo. Considerando que as entidades que instanciam os conceitos não são partes da ontologia, ontologias constituem o componente terminológico da base de conhecimento. Por sua vez, os fatos sobre as entidades do domínio compõem o componente assertivo da base.

Bases de conhecimento e banco de dados são estruturas distintas [Brodie and Mylopoulos 1986]. Como foi dito, bases de conhecimento devem estar associadas a uma teoria semântica que reflete o conhecimento do domínio de aplicação. Por sua vez, bancos de dados demandam uma teoria computacional concreta para o armazenamento e organização dos dados. A fim de exemplificar essa distinção, se remete ao exemplo biológico apresentado anteriormente. Nesse contexto, um banco de dados orientado a grafos conteria arestas relacionando espécimes a suas respectivas espécies (e.g., Bo01 é uma borboleta), assim como arestas relacionando taxonomicamente as espécies (e.g., Borboleta é um lepidóptero, que por sua vez, é um inseto). Perceba que o banco "desconhece" o fato *"Bo01 é um inseto"*, mesmo contendo informação suficiente para inferi-lo. Para realizar essa inferência seria necessário descrever, na forma de consulta, o caminho no grafo que explicite o encadeamento de raciocínio entre o espécime e conceito inseto.

Grafos de conhecimento representam o conhecimento de um domínio na forma de rede. Ainda não há uma definição formal para grafos de conhecimento [Ehrlinger and Wöß 2016]. Apesar disso, o termo grafo de conhecimento é utilizado frequentemente para se referir a uma estrutura (i) que representa o conhecimento de maneira similar à linguagem natural, isto é, em rede; (ii) que é útil na integração de dados de origens heterogêneas haja vista seu esquema flexível; (iii) que é frequentemente restrita por uma ontologia ou esquema de dados; e (iv) que está associada a aplicações e técnicas de inteligência artificial. Se o grafo contém uma formalização do conhecimento (e.g., uma ontologia), ele pode ser considerado um tipo de base de conhecimento no sentido tradicional do termo. Nessa perspectiva, grafos são bases de conhecimento estruturadas em grafo que armazenam informação factual na forma de relacionamentos. Para melhor compreensão daquilo a que grafos de conhecimento se referem, apresenta-se a seguir uma definição para eles. Ela se baseia nas definições apresentadas em trabalhos que propõem métodos de aprendizado para complementação de grafos de conhecimento.

É possível definir um grafo de conhecimento como o par $K = (\Delta, \Sigma)$ onde Σ e Δ denotam respectivamente os componentes terminológico (ontológico) e assertivo (de entidades) do grafo. O componente terminológico $\Sigma = (A, C, R_C, T_C, T_{C \to A}, V)$ é formado por (i) um conjunto C de conceitos; (ii) um conjunto R_C de (meta)-relações entre conceitos; (iii) um conjunto $A = \{A_j\}_{j=1}^{|A|}$ de atributos associados aos conceitos, onde cada atributo A_j toma valores em um conjunto de valores $V_j \in V$; (iv) um conjunto $T_C \subseteq C \times R_C \times C$

de relacionamentos entre conceitos; e (v) um conjunto $T_{C\to A} \subseteq C \times \{has\} \times A$ que associa atributos a conceitos, onde a constante *has* é utilizada para denotar que um conceito possuí determinado atributo (propriedade). Observe na Figura 1.1 que os nós *Pessoa* e *Organização* são conceitos, enquanto os nós *população* e *resumo* são atributos.

O componente assertivo $\Delta = (E, R_E, T_E, T_{E\to C}, T_{E\to V})$ é formado por (i) um conjunto E de entidades; (ii) um conjunto R_E de tipos de relações entre entidades; (iii) um conjunto de triplas $T_E \subseteq E \times R_E \times E$ contendo relacionamentos entre entidades; (iv) um conjunto $T_{E\to V} \subseteq E \times \cup_{j=1}^{|A|} \left(\{A_j\} \times V_j\right)$ contemplando relacionamentos atributivos, isto é, entre entidades e literais; e (v) um conjunto $T_{E\to C} \subseteq E \times \{isA\} \times C$ abarcando os relacionamentos de instanciação de entidades, onde a constante *isA* é utilizada para denotar que uma entidade instancia um conceito. Observe na Figura 1.1 que os nós *Elis* e *Maria* denotam entidades, a aresta (Elis, mãe-de, Maria) denota que Elis é mãe de Maria, enquanto a aresta (São Paulo, população, "11,967,825") denota que São Paulo possuí aproximadamente 12 milhões de habitantes.

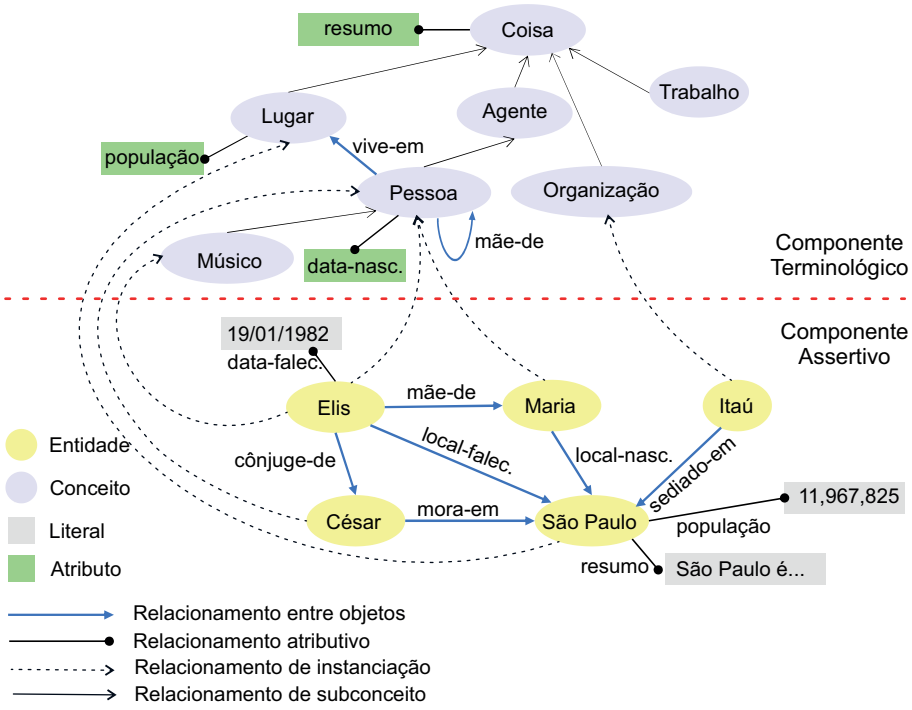

Figura 1.1. Exemplo de extrato de grafo de conhecimento.

Por fim, perceba que o conjunto de entidades e conceitos são disjuntos ($E \cap C = \emptyset$) assim como os conjuntos R_E, R_C e A também são disjuntos entre si ($(R_E \cup A) \cap R_C = \emptyset$, $R_E \cap A = \emptyset$). Além disso, o conjunto de triplas (arestas) T do grafo de conhecimento é tal que $T = T_E \cup T_{E\to C} \cup T_{E\to V} \cup T_C \cup T_{C\to A}$. Note que cada tripla $t \in T$ é da forma (s, r, o) onde s é a cabeça (sujeito), o a cauda (objeto) e r o tipo de relação.

1.3. Modelos de dados e sistemas

Para estruturar a informação referente a um grafo de conhecimento se recorre usualmente a um modelo de dados. Sobretudo, dois modelos (e suas ramificações) são usualmente

empregados na organização desse tipo de grafo: o modelo de dados do RDF (*Resource Description Framework*) e o grafo rotulado de propriedades (*Labeled Property Graph*).

O modelo do RDF[2] é o padrão W3C[3] de modelo de dados para a *Web* Semântica. Ele representa um grafo rotulado direcionado por meio de expressões da forma *sujeito-predicado-objeto*, conhecidas como triplas. Cada tripla (aresta) — identificada por um IRI (*Internationalized Resource Identifier*)[4] — representa uma asserção que relaciona dois nós do grafo. Cada nó é de um dos três tipos: recurso (*resource*), literal ou em branco (*blank*). Um nó recurso é identificado por um IRI e representa um elemento do domínio de interesse (entidades e conceitos). Por sua vez, um nó literal representa propriedades dos elementos do domínio. Para isso, ele possui um tipo de dados que define o intervalo de valores possíveis, e.g., cadeias de caracteres, números e datas. Por fim, um nó em branco representa um recurso para qual um IRI não foi dado. Note que em grafos RDF os nós e arestas não possuem uma estrutura interna, lhes distinguindo de grafos de propriedade.

Um grafo rotulado de propriedades é representado por nós, relações, propriedades e rótulos. Nesse grafo, cada nó possui um identificador único e um conjunto de propriedades (pares chave-valor) que os caracteriza. Além disso, um nó pode estar associado a zero ou mais rótulos, os quais podem representar classes, por exemplo. Por sua vez, os relacionamentos (arestas direcionadas) entre os nós devem estar associadas a um único tipo de relação. Mais ainda, de modo análogo aos nós do grafo, pode-se associar cada aresta a um conjunto de propriedades, o qual é definido pelo tipo de relação correspondente.

Nesse cenário, alguns tipos de sistemas são utilizados para o armazenamento e gerência de grafos. Em particular, de forma natural, grafos de conhecimento são usualmente armazenados e geridos em *triplestores* e sistemas de bancos de dados orientados a grafos (BDG). Dito isso, são citadas três ferramentas desenvolvidas com foco em grafos de conhecimento: Ontotext, Grakn e Amazon Neptune.

Ontotext GraphDB[5] é um triplestore com suporte a RDF e SPARQL[6]. A versão gratuita do sistema, implementada em Java, possui duas camadas: uma de inferência e outra de armazenamento, as quais empregam o *framework* RDF4J[7] de análise e consulta de dados RDF. O modelo de dados de Ontotext é baseado no RDFS (RDF Schema), o qual estende o vocabulário RDF ao permitir a descrição de taxonomias de classes e propriedades. Além disso, ele estende as definições de alguns dos elementos RDF, como o domínio e intervalo de propriedades.

Grakn[8] é um sistema de bancos de dados hiper-relacional dedutivo orientado ao armazenamento de grafos de conhecimento. O sistema lança mão de várias plataformas de computação distribuída e orientada a grafos, em especial, a JanusGraph[9], uma base de dados que implementa a API do Apache TinkerPop[10]. Além disso, Grakn provê um

[2]https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-mt/index.html
[3]https://www.w3.org/.
[4]https://www.w3.org/International/iri-edit/draft-duerst-iri-05.txt
[5]https://www.ontotext.com/products/graphdb/
[6]https://www.w3.org/TR/rdf-sparql-query/
[7]https://rdf4j.eclipse.org/
[8]https://grakn.ai/
[9]https://janusgraph.org/
[10]https://tinkerpop.apache.org/

sistema de representação do conhecimento baseado em hipergrafos e no modelo de entidades e relacionamentos. Por fim, o sistema provê uma linguagem de consulta chamada Graql. Por meio dessa linguagem, o usuário define a ontologia do grafo na forma de um esquema e regras, realiza consultas declarativas (*Online Transaction Processing*) capazes de inferência e executa tarefas analíticas (*Online Analytical Processing)* como o cômputo de centralidade em grafos.

Amazon Neptune[11] é um serviço e banco de dados orientado a grafos disponibilizado pela *Amazon Web Services*. O sistema suporta os modelos de dados grafo rotulado de propriedades e RDF, respectivamente, por meio da linguagem Gremlin do Apache TinkerPop e dos padrões W3C de Web Semântica RDF 1.1 e SPARQL 1.1. Especificamente, a unidade básica de dados é uma quádrupla (sujeito, objeto, predicado, grafo) — chamada de *quad* — baseada na tripla do RDF. Cada *quad* expressa a existência de um relacionamento entre dois recursos ou anexa um par chave-valor a um recurso. Além disso, segundo a representação adotada, RDF ou grafo de propriedades, o elemento *grafo* em cada quádrupla refere-se respectivamente a um *named graph identifier* ou identificador de aresta. Por fim, tanto os dados na forma de grafos de propriedades quanto RDF são armazenados no serviço Amazon S3; especificamente, em um volume virtual único que consiste de cópias dos dados ao longo de uma região AWS única.

1.4. Aplicações

Grafos de conhecimento têm se tornado uma tecnologia cada vez mais presente na indústria e academia, obtendo um papel de destaque em diversas aplicações. Na indústria, grafos de conhecimentos são adotados em aplicações como motores de busca, mecanismos de resposta a perguntas, sistemas de recomendação e agentes conversacionais. Na academia, espera-se que esses grafos promovam aplicações científicas — por exemplo, em biologia e medicina — por meio da integração de conhecimento acadêmico, assim como aplicações de grande impacto social como o combate à difusão de notícias falsas.

Atualmente os principais motores de busca — e.g., Baidu, Bing e Google – lançam mão de grafos de conhecimento na tarefa de resposta a consultas. Por exemplo, Freebase[12] foi utilizado na construção do grafo de conhecimento do Google. Esses motores de busca recorrem ao conhecimento enciclopédico e factual, expresso nesses grafos, sobre as entidades mais diversas, incluindo pessoas, localizações e instituições. Por exemplo, dada a consulta *Altura do Monte Everest*, além de apresentarem ao usuário documentos relacionados à consulta, esses motores também exibem painéis de informação, e.g., com a altura do monte e entidades relacionadas a ele.

Grafos de conhecimento também são empregados em sistemas computacionais como assistentes virtuais (e.g., Amazon Alexa, Google Assistant e Microsoft Cortanta), robôs conversacionais (e.g., ebay ShopBot e SalesForce Einstein Bot) e de respostas a perguntas (e.g., IBM Watson). Por exemplo, IBM Watson emprega grafos de conhecimento — e.g., DBpedia [Lehmann et al. 2015], Freebase e Yago [Rebele et al. 2016] — como fonte de informação estruturada [Nickel et al. 2016]. No contexto de comércio ele-

[11] https://aws.amazon.com/neptune/

[12] https://developers.google.com/freebase/data

trônico, ShopBot[13] foi um robô que recorria a grafos de conhecimento, contendo dados comportamentais e informação enciclopédica, a fim de compreender e refinar os pedidos de usuários em compras virtuais.

Cada vez mais sistemas de recomendação fazem parte da vida das pessoas, as sugerindo itens de interesse, por exemplo, filmes e músicas em serviços de *streaming*. Para realizar as recomendações é preciso que se modele nesses sistemas as interações entre os usuários e os itens de interesse. Isso é tradicionalmente feito ao recorrer-se a métodos de filtro colaborativo. Entretanto, o desempenho desses métodos sofre com a esparsidade dos relacionamentos entre os usuários e itens, além da falta de informação a respeito de usuários e itens recentes no sistema. Para enfrentar esses problemas, pode-se lançar mão da informação relativa aos usuários e itens. Em especial, grafos de conhecimento podem ser utilizados na estruturação dessa informação, promovendo recomendações mais adequadas [Wang et al. 2019a].

Grafos de conhecimento possuem grande potencial de aplicação nas áreas médicas e biológicas. Nessas áreas, grafos de conhecimento podem ser empregados na integração de conhecimento e informação biomédica. Por exemplo, a análise do grande e heterogêneo volume de literatura biomédica pode alavancar a descoberta de novos medicamentos. Nesse sentido, o método chamado GrEDel (*Graph Embedding based Deep Learning Method*) pode ser utilizado no processo de descoberta de fármacos [Sang et al. 2019]. Em resumo, esse método constrói um grafo de conhecimento a partir de resumos de artigos da literatura biomédica e aplica técnicas de aprendizado na descoberta de possíveis medicamentos.

Espera-se que grafos de conhecimento possam promover aplicações em medicina personalizada, a qual leva em conta a informação específica de cada paciente (e.g., variabilidade genética, ambiente e estilo de vida) na prevenção e tratamento de doenças. Essa abordagem médica depende da integração de um conjunto heterogêneo de informação sobre o paciente, o que pode incluir informação genética, além de dados sobre a administração de medicamentos e sobre o monitoramento das funções biológicas. Nesse contexto, a base (grafo) de conhecimento *Precision Medicine Knowledge Base* (PredMedKB) é uma iniciativa de integração. Em específico, esse grafo de conhecimento visa integrar informação e conhecimento sobre os quatro componentes fundamentais da medicina de precisão: doenças, genes, variantes genéticas e drogas [Yu et al. 2018].

Grafos de conhecimento também podem impulsionar o processo de checagem de fatos. Por causa dos danos sociais que a prática de propagação de notícias falsas incorre, iniciativas para checagem de fatos — e.g., o sítio *web* Snopes[14] — se fazem relevantes; em particular, iniciativas que realizam essa checagem de forma automatizada. Nesse contexto, grafos de conhecimento podem alavancar esse tipo de iniciativa ao serem empregados em métodos de detecção de notícias falsas baseada em conteúdo (*content based fake news detection*) [Pan et al. 2018].

[13] https://www.ebayinc.com/stories/news/say-hello-to-ebay-shopbot-beta/
[14] https://www.snopes.com/

1.5. Tarefas em grafos de conhecimento

Grafos de conhecimento estão associados a um conjunto de tarefas computacionais; desde a extração de informação até o uso do grafo na aplicação fim. Dentre essas, são ressaltadas, e apresentadas a seguir, a construção e o refino de grafos de conhecimento. Como o maior interesse deste capítulo é no uso de aprendizado de máquina no processo de refino do grafo, em particular, na inferência de elementos do grafo de conhecimento, o processo de construção é exposto de forma conceitual.

1.5.1. Construção automatizada de bases e grafos de conhecimento

O processo de construção, isto é, povoamento de bases e grafos de conhecimento com informação de interesse, tem se tornado uma tarefa cada vez mais automatizada. Em especial, a construção semiautomática desse tipo de base a partir da integração de dados estruturados, semiestruturados e não estruturados tem se tornado factível. Isso se deve em parte considerável a técnicas de extração de conhecimento baseadas em aprendizado de máquina. Em particular, aquelas que lançam mão de modelos de aprendizado profundo; estas têm obtido nos últimos anos desempenho estado da arte em subtarefas da construção de bases de conhecimento, operando diretamente nos dados de entrada como texto e imagem [Ratner et al. 2018]. Nesse contexto, descreve-se a seguir brevemente as tarefas de extração de entidades e relacionamentos, úteis no povoamento de bases de conhecimento; elas são descritas de forma conceitual e tomando em conta dados não estruturados (textuais). Em seguida, são apresentados os sistemas DeepDive e Fonduer, destinados à construção automatizada de bases e grafos de conhecimento.

1.5.1.1. Extração de entidades e relacionamentos

A tarefa de extração de entidades (*entity extraction*) visa obter entidades de interesse a partir de dados semiestruturados ou não estruturados [Yan et al. 2016]. Nesse contexto, o reconhecimento de entidades nomeadas e a ligação de entidades são duas subtarefas importantes desse tipo de extração. O objetivo do reconhecimento de entidades nomeadas (*Named Entity Recognition* - NER) é a identificação e classificação de entidades nomeadas (objetos do mundo real, e.g., pessoas e localizações) em documentos textuais. Por exemplo, ao ser considerado o fragmento textual *Einstein nasceu na Alemanha*, o resultado desejado do reconhecimento é a identificação e classificação dos termos *Einstein* e *Alemanha*, presentes no fragmento, como uma pessoa e país respectivamente. Em suma, dada uma sentença $x = (w_1, w_2, ..., w_n)$ o reconhecimento deve gerar como saída tuplas da forma (i_s, i_e, t) onde $i_s, i_e \in \{1, 2, ..., n\}$ são os índices inicial e final respectivamente de cada entidade nomeada, enquanto t é a classe/tipo a ela associada.

O objetivo da tarefa de ligação de entidades (*entity linking* ou *named entity desambiguation*) é vincular menções textuais as suas respectivas representações em um grafo de conhecimento de interesse [Yan et al. 2016]. Geralmente, essa tarefa está associada ao reconhecimento de entidades. Especificamente, ela realiza o processo de ligação a partir das menções (entidades nomeadas) produzidas durante o processo de NER. Por exemplo, o termo *Apple* nos trechos *Apple significa maçã em inglês* e *Apple é uma empresa de tecnologia* refere-se respectivamente a uma fruta e a uma empresa. Um método de ligação

de entidades deve associar o termo *Apple* no primeiro e segundo fragmento a entidades distintas no grafo de conhecimento. Em outras palavras, deve-se associar, se possível, cada entidade nomeada x a uma entidade $e \in E$ no grafo de conhecimento.

O objetivo da tarefa de extração de relacionamentos é obter fatos sobre as entidades de interesse a partir dos dados; por exemplo, o fato (Barack Obama, casado-com, Michelle Obama) a partir do fragmento textual *Barack Obama é casado com Michelle Obama* [Yan et al. 2016]. Usualmente, a tarefa de extração de relacionamentos é tomada como um problema de classificação binário. Por exemplo, dada uma sentença $x = (w_1, ..., e_1, ..., w_i, .., .e_2., .., w_n)$, onde e_1 e e_2 são entidades nomeadas e ϕ_x é o conjunto de características associado a x, deseja-se aprender um classificador f_r tal que $f_r(\phi_x) = 1$ se e_1 e e_2 são relacionadas pela relação r e $f_r(\phi_x) = 0$, caso contrário.

1.5.1.2. DeepDive

DeepDive [Zhang et al. 2016] é um sistema destinado à construção semiautomática de bases de conhecimento. A partir de uma coleção de dados estruturados, semiestruturados e não estruturados, o sistema extrai fatos, povoando uma base relacional. A principal motivação de DeepDive é aliviar o fardo de engenharia de características vinculado ao emprego de técnicas de aprendizado de máquina na construção desse tipo de base. Para isso, DeepDive implementa um conjunto de funcionalidades para extração de relacionamentos e emprega um modelo probabilístico na inferência do valor verdade dos elementos extraídos. Deve-se ressaltar que o sistema promoveu o desenvolvimento de aplicações em diversos domínios, incluindo no combate ao tráfico humano e em paleontologia.[15] Dito isso, o processo de construção adotado por DeepDive é descrito em termos gerais a seguir.

Em primeiro lugar, a coleção de documentos provida pelo usuário é armazenada em um banco de dados relacional. Por padrão, cada documento dessa coleção é processado e armazenado no formato uma linha por sentença de texto. Nesse processo, são anexadas aos textos marcações produzidas por ferramentas de pré-processamento de linguagem natural disponibilizadas no sistema. Após a ingestão dos documentos, DeepDive executa dois tipos de consultas: mapeamentos de candidatos a relacionamento e associação de características. O primeiro tipo produz menções textuais, entidades e relacionamentos possíveis e o segundo associa características aos candidatos a relacionamentos.

Posteriormente, o usuário elabora, de forma assistida, o conjunto de treinamento empregado no ajuste do modelo probabilístico. Em particular, DeepDive associa a cada relação da base de conhecimento uma relação evidência de mesmo esquema, salvo um campo adicional que indica se uma tupla na relação é falsa ou verdadeira. O povoamento da relação evidência é feita por meio de rotulagem manual ou supervisão distante.

A fim de estimar a probabilidade de os candidatos serem verdade, o sistema adota um grafo de fatores (*factor graph*) como modelo probabilístico, similar a Redes Lógicas de Markov [Richardson and Domingos 2006], além de usar técnicas do sistema Tuffy [Niu et al. 2011]. Candidatos cujas estimativas são maiores do que um limiar estabelecido pelo usuário são promovidos a relacionamentos.

[15] http://deepdive.stanford.edu/showcase/apps

1.5.1.3. Fonduer

Fonduer [Wu et al. 2018] é um sistema para construção de bases de conhecimento a partir de documentos formatados de forma complexa. Em geral, os sistemas destinados a construção de bases de conhecimento realizam o processo de extração a partir de dados textuais semiestruturados e tabulares. De forma distinta, Fonduer visa efetuar a construção de bases de conhecimento levando em conta informação multimodal. Por exemplo, no contexto de relatório técnicos, o sistema pode extrair o fato *O lucro líquido no quarto bimestre foi de $100* a partir de uma tabela e suas evidências textuais.

O processo de extração em Fonduer é realizado de forma semisupervisionada com base em heurísticas do usuário, assim como métodos de supervisão fraca (*weak supervison*) e modelos de aprendizado profundo. Em primeiro lugar, o usuário estabelece um conjunto de documento de interesse (e.g., PDFs e páginas em HTML) e o esquema alvo (tipo de relação), por exemplo, triplas do tipo (Cônjuge A, casado-com, Cônjuge B). O sistema processa cada documento de entrada em um modelo de dados que associa características aos elementos de informação, e.g., a altura relativa à página de uma tabela em um documento PDF. Além disso, o usuário escreve um conjunto de funções arbitrárias para extrair menções a entidades, por exemplo, a partir de fragmentos textuais ou tabelas HTML. O produto cartesiano entre essas menções forma o conjunto de candidatas a relacionamentos.

Como o número de candidatos pode ser grande, um conjunto de funções heurísticas, escritas pelos usuários, é utilizado para eliminar parte dos candidatos a relacionamentos. O usuário ainda descreve um conjunto de funções rotuladoras que associam a cada candidato não eliminado um rótulo de crença: verdadeiro, falso ou abstenção. Note que para um mesmo candidato, uma rotuladora pode associar um rótulo verdadeiro, enquanto outra um rótulo falso. Com base nos rótulos produzidos para os candidatos, um modelo generativo, baseado em *Data Programming* [Ratner et al. 2016], é aprendido a fim de estimar o erro associado aos rótulos e produzir um rótulo único (estocástico) para cada candidato. Por fim, os candidatos e seus respectivos rótulos são passadas a uma rede neural BiLSTM (*Bidirectional Long Short Term Memory*) multimodal que lança mão das características associadas aos candidatos para os classificar como verdadeiros ou falsos. Os candidatos possuindo rótulos verdadeiros são adicionados à base.

1.5.2. Refino de Grafos de Conhecimento

Por causa da natureza de seu processo de construção, grafos de conhecimento frequentemente contêm informação faltante ou ruidosa. Por exemplo, relacionamentos entre entidades do grafo que existem na realidade e não estão expressos no grafo, ou que não existem e estão. Com isso, o refino (*refinement*) de grafos de conhecimento é um processo natural. Nesse sentido, tarefas de refino de grafos de conhecimento podem ser divididas em no mínimo três maneiras distintas: (i) *objetivo geral da tarefa*: complementação ou correção; (ii) *alvo do refino*: por exemplo, entidades, relacionamentos, atributos; e (iii) *uso de informação lateral*, por exemplo, emprego de fontes de informação externas ao grafo de conhecimento na execução da tarefa [Paulheim 2017]. Como o processo de complementação é aquele de maior interesse deste capítulo, não são abordadas tarefas de correção. Interessados na correção de grafos de conhecimento podem recorrer a [Paulheim 2017].

1.5.2.1. Complementação de Grafos de Conhecimento

O objetivo da complementação de grafos de conhecimento é a adição de informação faltante, isto é, nós ou arestas ao grafo. Dentre as tarefas de complementação, a inferência de fatos não observados a partir do grafo se destaca. Esse tipo de inferência se traduz na predição de arestas do grafo de conhecimento, portanto, na inferência de informação relativa ao seu conjunto de triplas. A seguir são apresentadas as principais tarefas relacionadas a inferências de fatos em grafos de conhecimento. Posteriormente, na Seção 1.6, são apresentadas de maneira explícita, algumas técnicas para resolver essas tarefas.

Há no mínimo cinco tarefas associadas à complementação de fatos: inferência do valor verdade de triplas, predição de ligações, predição de atributos, predição de relações e classificação de entidades (ver Tabela 1.1). No âmbito de aprendizado de máquina, cada uma dessas tarefas é resolvida por meio do ajuste de um modelo à informação expressa no grafo de conhecimento. Em particular, de acordo com a tarefa de interesse, esse modelo toma como entrada uma tripla do grafo e produz um escore de plausibilidade.

O objetivo da tarefa de classificação de triplas (*triple classification*) é inferir o valor verdade de triplas não observadas. Em outras palavras, deseja-se corretamente inferir se triplas de consulta $(s,r,o) \notin T$ pertencem ou não ao grafo de conhecimento; por exemplo, o valor verdade da tripla (Einstein, morreu-em, EUA). Essa tarefa pode ser tratada como um problema de classificação binário, onde uma classe indica a veracidade de uma tripla, enquanto outra sua falsidade.

Tabela 1.1. Exemplos ilustrativos de complementação de fatos.

Tarefa	Exemplo de tripla de consulta	Exemplo de resultado
Classificação de tripla	(Einstein, morreu-em, EUA)	(Sim, 90%)
Pred. de ligação (cauda)	(Elvis Presley, estrelou-em, ?)	(Feitiço Havaiano, ...)
Pred. de ligação (cabeça)	(?, estrelou-em, Casablanca)	(Humphrey Bogart, ...)
Predição de relação	(Einstein, ?, Alemanha)	(nasceu-em, ...)
Predição de atributo	(B. Obama, nacionalidade, ?)	(americano, queniano, ...)
Classificação de entidade	(Michael Jackson, isA,?)	(cantor, compositor, ...)

Tipicamente, o objetivo da tarefa de predição de ligações (*link prediction*) é predizer se uma entidade se relaciona com outra, ou se um conceito está associado a outro. Em particular, no caso das entidades, deseja-se saber quais entidades $e \in E$ satisfazem determinada tripla de consulta incompleta na forma $(?,r,o)$ (predição de sujeito/cabeça) ou $(s,r,?)$ (predição de objeto/cauda), onde o símbolo "?" denota o alvo de inferência. Por exemplo, as consultas podem ter como objetivo o conhecimento sobre o filme Casablanca — (?, estrelou-em, Casablanca) — e o cantor Elvis Presley — (Elvis Presley, estrelou-em, ?). Note que o resultado desta tarefa é uma lista ranqueada de entidades (e.g., começando com filmes na segunda consulta) de maneira decrescente pelo escore de plausibilidade. Note que quanto maior o escore de um item, mais o modelo acredita que ele é verdadeiro.

O objetivo da predição de relações é inferir os tipos de relação existentes entre entidades ou conceitos, isto é, inferir que elementos satisfazem triplas de consulta $(s,?,o)$. Essa tarefa pode ser abordada a partir da classificação de triplas ou de forma análoga à

predição de ligações. Na primeira abordagem, para uma consulta $(s, ?, o)$, avalia-se o rótulo de classificação de todas as triplas (s, r, o), isto é, para todos os tipos de relação ($r \in R_E$ ou $r \in R_C$). Por outro lado, de maneira análoga a predição de ligações, é possível avaliar a classificação dos tipos de relação para aquela consulta.

Por fim, as tarefas de classificação de entidades e predição de atributos podem ser abordadas como especializações da predição de ligações. O objetivo da tarefa de classificação de entidades é associar classes às entidades do grafo. Se as classes estiverem expressas no grafo (conceitos), essa tarefa pode ser simplesmente tratada como um problema de predição de ligações do tipo $(s, isA, ?)$, onde $s \in E$ e "?" $\in C$. Caso as classes não estejam expressas no grafo, o problema pode ser visto como uma tarefa de aprendizado multiclasse, caso apenas uma classe deva ser associada a cada entidade, ou multirrótulo, caso mais de uma classe possa ser associada a cada entidade. Por sua vez, a predição de atributos visa inferir relacionamentos atributivos, isto é, o valor de um atributo associado a determinada entidade. Por exemplo, se o domínio do atributo for finito (ou considerado como finito), esse tipo de predição pode ser traduzida na predição de ligações da forma $(s, a, ?)$ onde $s \in E$, $a \in A$ e "?" $\in V_a$.

1.6. Aprendizado de Máquina Relacional

Aprendizado de máquina relacional (AMR) destina-se à criação de modelos estatísticos para dados relacionais, isto é, dados cuja a informação relacional é tão ou mais importante que a informação individual de cada elemento. Essa classe de aprendizado tem sido utilizada em diversas aplicações, por exemplo, na extração de informação de dados não estruturados [Zhang et al. 2016] e na modelagem de linguagem natural [Vu et al. 2018]. Em particular, técnicas AMR têm sido amplamente empregadas em tarefas associadas a grafos de conhecimento, sobretudo na sua complementação [Nickel et al. 2016].

A adoção de técnicas de aprendizado de máquina relacional em tarefas de complementação se baseia na ideia de existência de regularidades semânticas presentes no grafo de conhecimento. Essas regularidades, produto de padrões universais ou estatísticos, fazem com que o valor verdade de um relacionamento seja correlacionado com o valor verdade de outros relacionamentos. Por exemplo, em grafos de diversos domínios há uma tendência de entidades similares — i.e., que compartilham atributos comuns como faixa etária e crenças — se inter-relacionarem [Nickel et al. 2016]. Nesse caso, dadas duas entidades similares, se uma delas participa de um determinado tipo de relação, a chance da outra participar no mesmo tipo de relação aumenta.

Assumindo que os relacionamentos de interesse se deem apenas entre as entidades observadas no grafo de conhecimento, técnicas de AMR adotam três metodologias principais para abordar a existência e interdependência das triplas possíveis [Nickel et al. 2016]: (i) *modelos gráficos probabilísticos* assumem que a existência de cada tripla possível dependa da existência de um conjunto local de triplas; (ii) *modelos de características de grafo* assumem que a existência de cada tripla possível seja condicionalmente independente das demais, dadas as características **observadas** do grafo (e.g., caminhos) e parâmetros adicionais do modelo; e (iii) *modelos de características latentes* assumem que a existência de cada tripla possível seja condicionalmente independente das demais tripas dados os parâmetros do modelo e as características **não observadas** das entidades

s, o e relação r.

A seguir, essas metodologias de modelagem são apresentadas, sendo dada maior ênfase à apresentação de modelos de características latentes. A fim de simplificar a discussão, considere que, durante essa apresentação, apenas os relacionamentos entre entidades sejam de interesse, isto é, o domínio de triplas possíveis D seja tal que $D = E \times R_E \times E$. Ao discutir-se os aspectos de modelos de características latentes, são realizadas as devidas considerações sobre os demais tipos de relacionamento, e.g., ontológico e atributivo.

1.6.1. Modelos gráficos probabilísticos

No contexto de complementação, modelos gráficos probabilísticos assumem que a existência de uma tripla — isto é, ela representar uma proposição verdadeira — possa estar relacionada com as demais triplas [Raedt et al. 2016]. Em particular, para capturar a interdependência entre a existência de triplas, adota-se um grafo de dependências. Cada nó desse grafo representa uma variável estocástica $Y_{(s,r,o)} \in \{0,1\}$, a qual indica a existência de uma tripla possível $(s,r,o) \in D$. Por sua vez, cada aresta desse grafo de dependências modela a interdependência entre duas triplas. Uma vez que é impraticável considerar todas as $|D| \times (|D|-1)$ possíveis interdependências, é necessário que apenas aquelas mais relevantes sejam consideradas. Nesse contexto, usualmente emprega-se o modelo gráfico probabilístico não direcionado Campos Aleatórios de Markov como ferramenta de representação dessas interdependências. Particularmente no contexto de complementação são adotadas Redes Lógicas de Markov [Richardson and Domingos 2006], uma extensão desse modelo.

Redes Lógicas de Markov combinam Campos Aleatórios de Markov e lógica de primeira ordem. Nesse sentido, além do conjunto de triplas, emprega-se um conjunto de fórmulas lógicas que expressam regras e heurísticas do domínio do grafo de conhecimento, sendo cada fórmula associada a um peso real. Por exemplo, a fórmula (X, cônjuge_de, Y), (Y, mãe_de, Z) $\rightarrow$ (X,pai_de, Y) indica que usualmente o esposo da mãe de um indivíduo é seu pai. Essas fórmulas são utilizadas na definição de quais interdependências entre triplas devem ser consideradas. Em um processo chamado de instanciação, essas fórmulas são instanciadas (e.g., (João, cônjuge_de, Maria), (Maria, mãe_de, Lúcio) $\rightarrow$ (João, pai_de, Lúcio)) de forma coerente, isto é, obedecendo as restrições. Com base nesse processo, a probabilidade conjunta da existência de triplas é modelada por

$$P\left(\bigcap_{(s,r,o)\in D} Y_{(s,r,o)} \mid \theta\right) = \frac{1}{Z}\prod_i \exp\left(\theta_i \cdot x_i\right) \tag{1}$$

onde x_i e θ_i denotam respectivamente a quantidade de instanciações válidas e peso associados à fórmula f_i. Além disso, Z é uma função de partição que assegura que P é uma distribuição de probabilidade.

Como o processo de inferência — estimativa da atribuição mais provável para os $Y_{(s,r,o)}$ — é um problema computacionalmente intratável, emprega-se abordagens heurísticas, por exemplo, amostragem de Gibbs e MC-SAT. Além disso, como o aprendizado de parâmetros θ por maximização de verossimilhança ou probabilidade a *posteriori* recorre a etapa de inferência, são empregadas aproximações como pseudo-verossimilhança.

1.6.2. Modelos de características de grafo

Modelos de características de grafo lançam mão de representações baseadas em elementos observáveis na estrutura do grafo, por exemplo, caminhos e vizinhanças. Esse tipo de método parte da premissa de que existem padrões expressos no grafo que possuem poder preditivo. Por exemplo, a quantidade de caminhos entre duas entidades pode ser um indicador da existência de determinado relacionamento entre elas. Nesse contexto, algumas abordagens para inferência de triplas incluem o uso de índices de similaridade, mineração de regras e programação lógica indutiva [Nickel et al. 2016]. Dentre essas, destaca-se o método *Path Ranking Algorithm*.

Path Ranking Algorithm [Lao et al. 2011] é um algoritmo para produção de modelos de características de grafo. Ele emprega a exploração aleatória de caminhos de comprimento limitado no grafo de conhecimento a fim de construir representações vetoriais (vetores de características) para suas triplas. A construção dessas representações é dividida em duas etapas, extração de características e treinamento. Na etapa de extração de características, um conjunto de caminhos de relação é selecionado; por exemplo, um conjunto $P = \{p_i\}_{i=1}^{|P|}$ de caminhos de comprimento n. Cada um desses caminhos segue a forma $p = (r_1, r_2, ..., r_n)$ onde cada r_i é um tipo de relação. Por exemplo, p = (cônjuge_de, mãe_de) é um caminho de relação de comprimento dois.

Após serem extraídos os caminhos de relação, um conjunto de treinamento é selecionado a partir do conjunto de triplas. Para cada tripla (s, r^*, o) no conjunto de treinamento e cada caminho de relação $p \in P$ computa-se a probabilidade que ao se iniciar o caminho p em s se chegue a o de forma consistente, isto é, seguindo os tipos de relação expressos em p. Note que o cômputo dessa probabilidade é feito de forma uniforme, isto é, a probabilidade de navegar-se "para fora" de um nó s através de um determinado tipo de relação r' é proporcional a quantidade de vizinhos associados a s por r'.

Após computado, o conjunto de probabilidades é empilhado em um vetor de características $f^{\text{PRA}}_{s,r^*,o} \in \mathbb{R}^{|P|}$ e associado à tripla (s, r^*, o). Computadas as representações vetoriais para as triplas de um conjunto de treinamento, um modelo de aprendizado "de prateleira" é ajustado. Por exemplo, ao ser empregado um modelo de regressão logística, define-se o escore dado a uma tripla (s, r, o) como

$$\phi^{\text{PRA}}_{(s,r,o)} := \sigma\left(\mathbf{v}_r^\top f^{\text{PRA}}_{s,r,o}\right) \tag{2}$$

onde $\mathbf{v}_r \in \mathbb{R}^{|P|}$ denota o vetor de pesos (a ser aprendido) associado ao tipo de relação r e $\sigma(x) = 1/(1+\exp^{-x})$ é a função sigmoide. Note que a cada nova consulta sobre a existência de uma tripla é necessário computar o vetor de características a ela associado.

1.6.3. Modelos de características latentes

Nos últimos anos, o desenvolvimento de modelos de características latentes tem se tornado a linha de pesquisa dominante na tarefa de complementação [Kejriwal 2019], sendo possível enunciar alguns fatores para isso. Em primeiro lugar, há o sucesso recente da área de pesquisa de aprendizado de representações (*embeddings*) [Bengio et al. 2013, Hamilton et al. 2017]. No caso particular de grafos de conhecimento, a ideia é que as representações das entidades e relacionamentos, necessárias para o melhor desempenho

de um modelo, precisam ser aprendidas. Em outras palavras, elas devem ser produzidas durante o processo de aprendizado de um modelo e não engendradas minuciosamente a priori [Hamilton et al. 2017]. Isso se contrapõe às abordagens de características de grafo, as quais definem a priori vetores de características com base em propriedades do grafo (e.g., estatísticas sumarizantes). Em segundo lugar, modelos de características latentes não pressupõem uma representação simbólica mais formal do conhecimento (e.g., definição de regras) e têm demonstrado serem escaláveis a grafos com milhões de entidades [Nickel et al. 2016]. Isso vai de encontro a grande parte dos modelos gráficos probabilísticos usados em complementação, como as Redes Lógicas de Markov, apresentadas anteriormente. Essa última classe de modelos, apesar de ter sido dominante na tarefa de complementação no passado, perdeu popularidade por causa de dificuldades tangentes à escalabilidade dos processos de inferência [Kejriwal 2019].

De forma geral, modelos de características latentes, também chamados de modelos de *embedding*, embutem entidades e relações em espaços vetoriais reais e complexos [Wang et al. 2017]. O modelo é ajustado para que a estrutura do espaço de *embedding* reflita a estrutura do grafo de conhecimento; por exemplo, mantendo uma certa similaridade entre os relacionamentos geométricos das representações vetoriais e seus correspondentes expressos simbolicamente no grafo de conhecimento. Além disso, a dimensão desse espaço escolhido precisa ser bem menor do que a quantidade de entidades presentes no grafo. Desse modo, uma maior quantidade de regularidades presentes no grafo pode ser capturada. Entretanto, esse número não deve ser muito baixo a ponto de as representações vetoriais não serem capazes de modelar a semântica do grafo.

As técnicas baseadas em *embeddings* podem ser categorizadas em dois grupos: modelos de distância translacional (*translational distance models*) e modelos de correspondência semântica (*semantic matching models*) [Wang et al. 2017]. Modelos de distância translacional exploram funções de escore baseadas em distância. Isto é, eles medem a plausibilidade de um fato como algum tipo de distância entre as representações vetoriais das entidades envolvidas nesse fato, usualmente após a translação pelo tipo de relação correspondente. Por sua vez, modelos de combinação exploram funções de escore baseadas em similaridade. Eles medem a plausibilidade de um fato ao combinar a semântica latente de entidades e relacionamentos. A seguir são apresentados alguns desses modelos. Posteriormente, o processo de treinamento utilizado no aprendizado desses modelos é discutido.

1.6.3.1. Modelos de distância translacional

TransE [Bordes et al. 2013] foi um dos primeiros modelos de *embedding* propostos para grafos de conhecimento; sendo ele de certo modo o "pai" dos modelos translacionais. Entretanto, apesar de sua idade, ele continua sendo relevante, tanto como medida de comparação quanto base para novos modelos. Por exemplo, TransH [Wang et al. 2014], TransR [Lin et al. 2015] e TransA [Jia et al. 2016] estendem as ideias de TransE como é disposto na Figura 1.2 e Tabela 1.2. Em particular, em TransE os relacionamentos são representados como translações em um espaço de *embedding*. Uma das motivações para esse tipo de abordagem vem do uso de aprendizado de representações no processamento

de linguagem natural. Nesse contexto, observou-se que alguns modelos de *embedding* representavam as palavras referentes a relacionamentos (e.g., *capital-de*) como translações [Bouraoui et al. 2018].

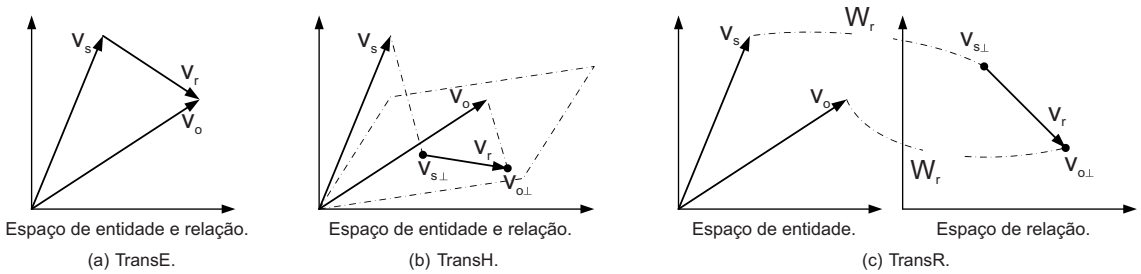

Figura 1.2. Ilustrações de modelos translacionais. Adaptado de [Wang et al. 2017, p.3]. Note que enquanto TransE aplica a ideia de translação de forma direta, TransH [Wang et al. 2014] e TransR [Lin et al. 2015] primeiro projetam as entidades em um hiperplano e espaço específico de relação respectivamente.

TransE gera *embeddings* para as entidades e relações de tal forma que a representação vetorial do objeto de uma tripla seja aproximadamente igual à translação da representação do sujeito. Em outras palavras, para cada tripla $(s,r,o) \in T_E$, $\mathbf{v}_s + \mathbf{v}_r \approx \mathbf{v}_o$ onde $\mathbf{v}_s, \mathbf{v}_r, \mathbf{v}_o \in \mathbb{R}^d$ ($d \in \mathbb{N}$) são as representações vetoriais, respectivamente, de s, r e o. Em particular, o modelo é definido pela função de escore:

$$\phi^{\text{TransE}}_{(s,r,o)} := -\|\mathbf{v}_s + \mathbf{v}_r - \mathbf{v}_o\|_{1\vee 2} \tag{3}$$

onde $\|\cdot\|_{1\vee 2}$ é a norma L_1 ou L_2. Perceba que apesar de TransE ser capaz de abarcar relações 1:1, ele apresenta dificuldades ao lidar com relações do tipo 1:N, N:1 ou M:N. Tome como exemplo o tipo de relação M:N *atua-em*; ela indica que um ator atua em um filme. Se houver duas triplas no grafo (a, atua-em, f1) e (a, atua-em, f2), o modelo poderá aprender representações similares para f1 e f2 ($\mathbf{v}_{f1} \approx \mathbf{v}_{f2}$), mesmo se f1 e f2 forem elementos muito distintos.

Por causa das dificuldades apresentadas, alguns modelos têm sido propostos, dentre os quais, apresenta-se o TransH. Ao invés de utilizar apenas um vetor para cada tipo de relação, TransH [Wang et al. 2014] emprega dois vetores. Em particular, um vetor $\mathbf{v}_r \in \mathbb{R}^d$ de norma de um hiperplano e um vetor $\mathbf{w}_r \in \mathbb{R}^d$ de projeção. A ideia é que para triplas verdadeiras (s,r,o) a projeção de $\mathbf{v}_s$ e $\mathbf{v}_o$ estejam aproximadamente conectadas por $\mathbf{v}_r$. Com essa mudança o método é capaz de modelar de maneira mais adequada tipos de relação que não são funcionais e nem injetivos. Dito isso, a função de escore de TransH é definida como:

$$\phi^{\text{TransH}}_{(s,r,o)} := -\|(\mathbf{v}_s - \mathbf{w}_r^\top \mathbf{v}_s \mathbf{w}_r) + \mathbf{v}_r - (\mathbf{v}_o - \mathbf{w}_r^\top \mathbf{v}_o \mathbf{w}_r)\|_2^2 \tag{4}$$

1.6.3.2. Modelos de correspondência semântica

Diversos modelos de correspondência semântica têm sido propostos nos últimos anos; por exemplo, RESCAL, ANALOGY, SimplE, ConvE e R-GCN.

Tabela 1.2. Parâmetros de modelos translacionais.

Método	Embedding de entidade	Embedding de relação	Função de Escore
TransE	$\mathbf{v}_s, \mathbf{v}_o \in \mathbb{R}^d$	$\mathbf{v}_r \in \mathbb{R}^d$	$-\|\mathbf{v}_s + \mathbf{v}_r - \mathbf{v}_o\|_{1\vee 2}$
TransH	$\mathbf{v}_s, \mathbf{v}_o \in \mathbb{R}^d$	$\mathbf{v}_r, \mathbf{w}_r \in \mathbb{R}^d$	$-\|(\mathbf{v}_s - \mathbf{w}_r^\top \mathbf{v}_s \mathbf{w}_r) + \mathbf{v}_r - (\mathbf{v}_o - \mathbf{w}_r^\top \mathbf{v}_o \mathbf{w}_r)\|_2^2$
TransR	$\mathbf{v}_s, \mathbf{v}_o \in \mathbb{R}^d$	$\mathbf{v}_r \in \mathbb{R}^k, \mathbf{W}_r \in \mathbb{R}^{k\times d}$	$-\|\mathbf{W}_r\mathbf{v}_s + \mathbf{v}_r - \mathbf{W}_r\mathbf{v}_o\|_2^2$
TransA	$\mathbf{v}_s, \mathbf{v}_o \in \mathbb{R}^d$	$\mathbf{v}_r \in \mathbb{R}^d, \mathbf{W}_r \in \mathbb{R}^{d\times d}$	$-\lvert\mathbf{v}_s + \mathbf{v}_r - \mathbf{v}_o\rvert^\top \mathbf{W}_r \lvert\mathbf{v}_s + \mathbf{v}_r - \mathbf{v}_o\rvert$

RESCAL [Nickel et al. 2011] modela a plausibilidade de uma tripla por meio das interações par a par entre as características latentes das entidades nela retratadas. Especificamente, ele modela o escore de uma tripla (s,r,o), isto é, sua plausibilidade de ser verdadeira, como:

$$\phi^{\text{RESCAL}}_{(s,r,o)} := \mathbf{v}_s^\top \mathbf{W}_r \mathbf{v}_o = \sum_{i=1}^{d}\sum_{j=1}^{d} \mathbf{W}_{k_{ij}} \mathbf{v}_{s_i} \mathbf{v}_{o_j} \tag{5}$$

onde $d \in \mathbb{N}$ é a dimensão do espaço de *embedding* de entidades e $\mathbf{v}_s \in \mathbb{R}^d$, $\mathbf{v}_o \in \mathbb{R}^d$ e $\mathbf{W}_r \in \mathbb{R}^{d\times d}$ são respectivamente as representações vetoriais para s, o e r. Perceba que cada escalar $\mathbf{W}_{r_{ij}}$ especifica o quanto as características não observadas i e j, respectivas às representações de s e o, interagem na relação r.

ANALOGY [Liu et al. 2017] lança mão da ideia de que propriedades analógicas entre entidades e relações ajudam na predição de fatos. Por exemplo, suponha a analogia "*homem é para rei aquilo que mulher é para rainha*". A ideia é que o conjunto $\{(homem, r_1, rei), (mulher, r_1, rainha), (homem, r_2, mulher), (rei, r_2, rainha)\}$ forme uma estrutura analógica, onde r_1 e r_2 denotam tipos de relação. Essa estrutura indica que a relação entre *homem* e *rei* ajuda a predizer os relacionamentos não observados entre *mulher* e *rainha*. Nesse sentido, ANALOGY emprega a mesma função de escore de RESCAL:

$$\phi^{\text{ANALOGY}}_{(s,r,o)} := \mathbf{v}_s^\top \mathbf{W}_r \mathbf{v}_o \tag{6}$$

entretanto, para capturar estruturas analógicas, o modelo impõe que as matrizes de relação sejam normais ($\mathbf{W}_r \mathbf{W}_r^\top = \mathbf{W}_r^\top \mathbf{W}_r$) e comutem entre si ($\mathbf{W}_{r_1}\mathbf{W}_{r_2} = \mathbf{W}_{r_2}\mathbf{W}_{r_1}$).

SimplE [Kazemi and Poole 2018] é um modelo baseado na decomposição de posto tensorial. Nele são considerados dois vetores $\mathbf{v}_e^{(+)} \in \mathbb{R}^d$ e $\mathbf{v}_e^{(-)} \in \mathbb{R}^d$ para representar cada entidade $e \in E$. Os vetores $\mathbf{v}_e^{(+)}$ e $\mathbf{v}_e^{(-)}$ são respectivamente as representações de e como sujeito e objeto das relações. De mesmo modo, dois vetores são considerados para cada relação $r \in R_E$, $\mathbf{v}_r^{(+)}$ e $\mathbf{v}_r^{(-)}$, onde $\mathbf{v}_r^{(-)}$ visa representar a relação inversa de r. Dito isto, o escore dado por SimplE é definido como:

$$\phi^{\text{SimplE}}_{(s,r,o)} := \frac{1}{2}\sum_{i=1}^{d} \left(\mathbf{v}_{s_i}^{(+)} \mathbf{v}_{r_i}^{(+)} \mathbf{v}_{o_i}^{(-)} + \mathbf{v}_{o_i}^{(+)} \mathbf{v}_{r_i}^{(-)} \mathbf{v}_{s_i}^{(-)}\right) \tag{7}$$

Além disso, para capturar conhecimento ontológico existente — em especial, relações simétricas, antissimétricas e inversas — são feitas restrições aos vetores de *embedding* das relações; por exemplo, no caso de r ser simétrica, impõe-se que $\mathbf{v}_r^{(+)} \approx \mathbf{v}_r^{(-)}$ e que os vetores de *embedding* sejam não negativos.

Os modelos de *embedding* apresentados acima empregam diretamente as representações vetoriais no cômputo do escore de predição. Uma das desvantagens desse tipo de abordagem é que a única maneira de aumentar a expressividade de uma representação — i.e., a quantidade de características latentes — é adotar um espaço de *embeddings* com maior dimensão. Todavia, isso não escala para grafos de larga escala, uma vez que o número de parâmetros do *embedding* é da ordem do grafo. O aumento da quantidade de características de forma independente do espaço de *embedding* requer o uso de múltiplas camadas de características. Entretanto, esse tipo de abordagem exige cuidados adicionais para que o modelo gerado não superajuste (*overfitting*) aos dados de treinamento e consequentemente não generalize [Nickel et al. 2016].

ConvE [Dettmers et al. 2018] é um modelo convolucional que ataca os desafios apresentados. Em particular, ele emprega camadas de convolução bidimensional e totalmente conectadas na modelagem de interações entre as representações vetoriais de relação e entidades. O modelo utiliza uma camada de convolução para capturar a interação entre as representações da entidade s e relação r e camadas não lineares para aumentar a expressividade das interações entre s, r e o. Em suma, ConvE possui uma arquitetura definida por três camadas: convolução, projeção e produto interno, sendo sua função de escore definida como:

$$\phi^{\text{ConvE}}_{(s,r,o)} := \underbrace{f_2\left(\underbrace{\text{vec}\left(f_1\left(\underbrace{\text{conv}_{\omega}\left(\text{concat}(\bar{\mathbf{v}}_s, \bar{\mathbf{v}}_r)\right)}_{\text{Convolução}}\right)\right)\mathbf{W}}_{\text{Projeção}}\right)\mathbf{v}_o}_{\text{Produto Interno}} \tag{8}$$

onde (i) f_1, f_2 são funções de ativação não linear (e.g., $f_1 = ReLU$[16] e $f_2 =$ sigmoide); (ii) concat concatena duas matrizes uma embaixo da outra; (iii) conv_{ω} é a camada de convolução parametrizada pelos filtros ω; (iv) $vec(\cdot)$ é uma operação de achatamento, a qual ordena um tensor ou matriz na forma de um vetor; (v) $\mathbf{W}$ é uma matriz de parâmetros utilizada na projeção ao espaço de *embedding*; e finalmente (vi) $\bar{\mathbf{v}} \in \mathbb{R}^{m \times n}$ é a representação matricial adotada para $\mathbf{v} \in \mathbb{R}^d$ $(m \times n = d)$[17]. Na Figura 1.3 é mostrada graficamente a arquitetura do modelo ConvE.

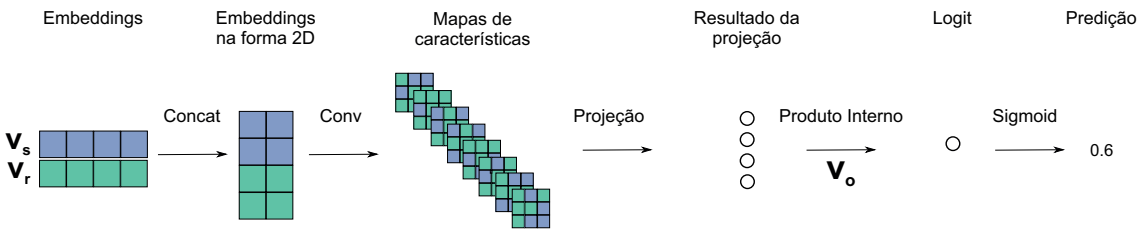

Figura 1.3. Ilustração do modelo ConvE. Fonte: Adaptado de [Dettmers et al. 2018, p.1814].

[16] $ReLU(x) = max(0, x)$

[17] Convoluções bidimensionais esperam que o dado de entrada seja bidimensional.

R-GCNs (*Relational Graph Convolution Neural Network*) [Schlichtkrull et al. 2018] são modelos de aprendizado que estendem redes de convolução de grafos para o cenário multirrelacional. Na complementação de grafos de conhecimento, esse tipo de modelo é utilizado como codificador em um modelo auto-codificador $\phi^{\text{auto-encoder}}_{(o,r,s)} = h_r(g)$. Especificamente, o codificador $g : E \to \mathbb{R}^d$ embute as entidades do grafo no espaço de *embedding* e o decodificador $h_r : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (parametrizado pelo tipo de relação $r \in R_E$) dá um escore de plausibilidade para uma tripla. Note que o decodificador pode ser qualquer um dos modelos apresentados anteriormente. Além disso, o processo de aprendizado é feito *end-to-end*, isto é, o ajuste do modelo auto-codificador é feito de forma conjunta (codificador mais decodificador).

De forma concreta, R-GCNs são redes neurais multicamada, que quando utilizadas como codificadores, visam aprender representações vetoriais para entidades. Em particular, R-GCNs implementam duas ideias básicas (i) as representações vetoriais de uma entidade devem ser o produto de múltiplas camadas; e (ii) a representação vetorial de uma entidade deve estar relacionada com as representações vetoriais das suas entidades vizinhas. Cada camada oculta $l \in 1, 2, .., L$ da rede é da forma:

$$\mathbf{v}_e^{(l)} = f\left(\underbrace{\mathbf{W}_0^{(l-1)}\mathbf{v}_e^{(l-1)}}_{\text{Própria entidade.}} + \underbrace{\sum_{r\in R_E}\sum_{e'\in N_e^r}\frac{1}{c_{e,r}}\mathbf{W}_r^{(l-1)}\mathbf{v}_{e'}^{(l-1)}}_{\text{Entidades vizinhas.}}\right) \tag{9}$$

onde (i) $\mathbf{v}_e^{(l)} \in \mathbb{R}^{d_l}$ é a representação de $e \in E$ na camada l; (ii) $\mathbf{W}_r^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ é uma matriz de parâmetros para a relação r; (iii) $\mathbf{W}_0^l \in \mathbb{R}^{d_l \times d_{l-1}}$ transforma a representação de e da camada anterior para o espaço da camada atual; (iv) $N_e^r = \{o \mid (e,r,o) \in T_E\}$ é a vizinhança de e[18]; (v) $c_{e,r}$ é uma constante de normalização; (vi) f é uma função de ativação não linear; e (vii) $d^l \in \mathbb{N}$ é o tamanho da dimensão das representações ocultas das entidades na camada l. Note que as representações $\{\mathbf{v}_e^{(L)} \in \mathbb{R}^d \mid e \in E\}$ são o resultado da codificação. Na Figura 1.4 o cômputo das representações é exemplificado.

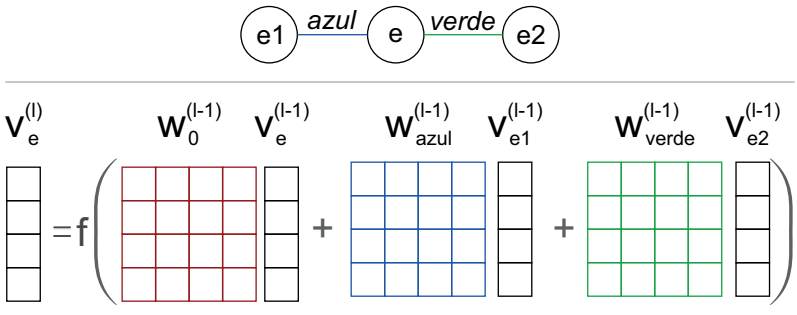

Figura 1.4. Representação de uma camada do modelo R-GCN. A ilustração mostra o cômputo da representação da entidade e na camada l. A entidade e relaciona-se com as entidades $e1$ e $e2$ por meio da relação "azul" e "verde", respectivamente.

[18]Para o cômputo da representação de uma entidade e, além das arestas saintes, também são consideradas as entrantes já que no modelo, R_E contém as inversas de relações (e.g., *pai-de* e sua inversa *pai-de*$^{-1}$).

1.6.3.3. Literais e relacionamentos atributivos

Na Seção 1.2 mostrou-se que grafos de conhecimento usualmente contêm um conjunto de triplas $T_{E \to V}$ atributivas, que ligam entidades a valores literais (de atributo). Esses valores podem ser de diferentes tipos, incluindo dados textuais (e.g., nomes e comentários), numéricos (e.g., altura e ano) ou mesmo imagens. Note que os considerar na modelagem de grafos de conhecimento pode ajudar a produzir melhores representações vetoriais para entidades e, portanto, modelos de *embedding* mais adequados [Gesese and Russa Biswas 2019]. Em particular, eles podem ser úteis no aprendizado de representações para entidades que possuem poucos ou nenhum relacionamento observado no grafo.

A maior parte dos modelos de *embedding* (incluindo os discutidos anteriormente) não leva em conta de forma explícita esse tipo de informação. Isso dificulta, por exemplo, a realização da tarefa de predição de atributos. Nesse contexto, uma abordagem imediata seria modelar literais da mesma forma que entidades. Entretanto, apesar de sua simplicidade, essa abordagem sofre como alguns problemas, em especial, ela pode aumentar drasticamente a quantidade de parâmetros a serem aprendidos, assim como apenas consegue lidar com atributos categóricos [Wang et al. 2017]. Em face ao exposto, modelos de *embedding* capazes de lidar com valores literais têm sido propostos [Gesese and Russa Biswas 2019]; dentre eles, destaca-se MKBE.

Multimodal Knowledge Base Embeddings (MKBE) [Pezeshkpour et al. 2018] é um modelo de complementação de grafos de conhecimento que emprega diferentes codificadores neurais no aprendizado de *embedding* para tipos diversos de dados (textual, numérico e imagens). Assim como, o modelo R-GCN, apresentado anteriormente, MKBE adota uma abordagem de auto-codificador que é descrita a seguir.

O processo de codificação emprega para cada tipo de elemento — imagem, número, texto, entidade, relação — uma rede neural distinta. A representação vetorial de uma imagem é obtida por meio de uma rede neural de convolução; especificamente, uma VGGNet pré-treinada na base de dados ImageNet[19]. No que lhe diz respeito, o vetor de *embedding* de um literal numérico é obtido por meio de uma rede neural *feed-forward*. Por sua vez, emprega-se dois tipos de redes neurais profundas na codificação de literais textuais. Em particular, a representação de textos de menor comprimento (e.g., nomes) é obtida por uma arquitetura recorrente bidirecional GRU (*Gated Recurrent Unit*) a nível de caractere, enquanto *embeddings* de textos de maior comprimento (e.g., descrições textuais), por meio de uma rede de convolução. Por fim, emprega-se duas camadas densas de rede neural no aprendizado de *embeddings* de entidades e tipos de relação.

O decodificador de MKBE realiza o processo de inferência. Notadamente, ele lança mão dos *embeddings* produzidos pelo codificador para produzir o escore referente a uma tripla. O decodificador pode ser, por exemplo, o modelo ConvE, apresentado anteriormente. Perceba que o processo de aprendizado é realizado *end-to-end*, isto é, os parâmetros do codificador e decodificador são ajustados de forma conjunta.

Na Figura 1.5 é apresentada a arquitetura de MKBE. Os vetores $\mathbf{s} \in \{0,1\}^{|E|}$ e

[19] http://www.image-net.org/

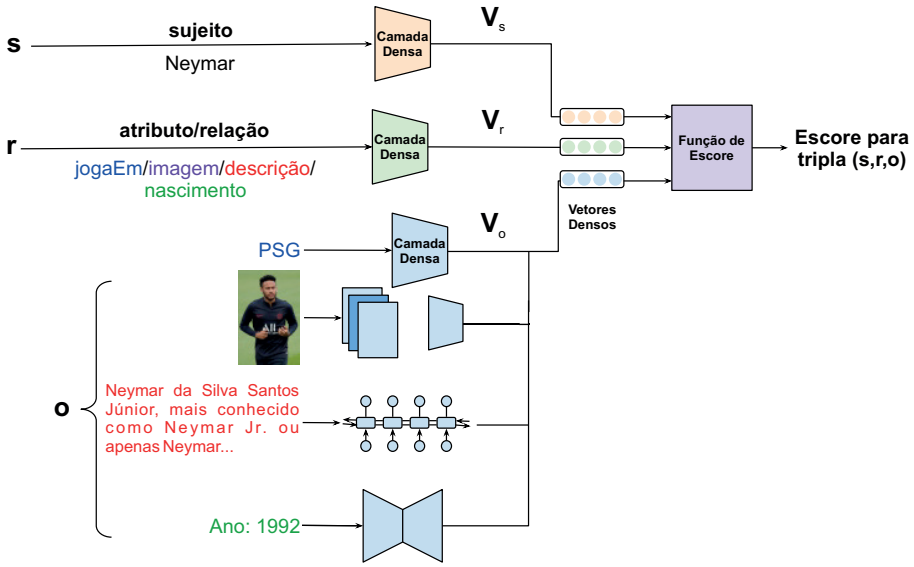

Figura 1.5. Ilustração da arquitetura do modelo MKBE. Informação referente ao jogador de futebol Neymar é utilizada como exemplo. Fonte: Adaptado de [Pezeshkpour et al. 2018, p.3211].

$\mathbf{r} \in \{0,1\}^{|R_E \cup A|}$ são a codificação *one-hot* da entidade s e da relação r respectivamente.[20] Por sua vez, o significado do objeto $\mathbf{o}$ depende daquilo a que ele se refere: entidade (*one-hot encoding*), imagem (tensor real tridimensional), texto (*embeddings* textuais) e literal numérico (números reais padronizados). Perceba ainda os *embeddings* de entidades e atributos $\mathbf{v}_s, \mathbf{v}_o \in \mathbb{R}^d$. Os *embedding* de relação $\mathbf{v}_r$ depende da função de escore empregada; por exemplo, se for adotado ConvE, então $\mathbf{v}_r \in \mathbb{R}^d$.

1.6.3.4. Emprego de ontologias

Conforme exposto na Seção 1.2, grafos de conhecimento geralmente incluem um componente terminológico; em particular, na forma de informação ontológica. Entretanto, ainda é incipiente o uso dessa classe de informação no desenvolvimento de modelos de complementação baseados em *embeddings*. Nesse sentido, a fim de que se modele de forma mais adequada grafos de conhecimento, recentemente alguns trabalhos recorrem a ontologias. Dentre esses, destaca-se aqueles que as empregam no desenvolvimento de novos modelos e na restrição do intervalo de valores das representações vetoriais a serem aprendidas.

JOIE [Hao et al. 2019] é um modelo de complementação que codifica de forma conjunta tanto o componente assertivo quanto terminológico do grafo de conhecimento. Em particular, JOIE é formado por dois componentes. O primeiro deles, chamado de *cross-view association model*, tem como intuito associar o *embedding* de uma entidade ao seu respectivo *embedding* de conceito; por exemplo, a representação vetorial da entidade Albert Einstein à representação da classe pessoa. Para isso, duas técnicas são empregadas, *cross-view grouping* e *cross-view transformation*. Em resumo, a primeira delas visa agrupar entidades referentes ao mesmo conceito, enquanto a segunda, mapear o es-

[20] O vetor *one-hot* $\mathbf{v}$ associado à n-ésima entidade (relação/atributo) é tal que $\mathbf{v}_n = 1$ e $\mathbf{v}_{j \neq n} = 0$.

paço representacional das entidades ao de conceitos. O segundo componente, chamado de *intra-view embedding model*, visa caracterizar as triplas pertencentes aos componentes terminológico e assertivo em dois espaços de *embeddings* distintos. Para isso, lança-se mão de uma função de escore para as triplas T_E e outra para T_C. Essas funções de escore podem ser os modelos apresentados anteriormente, por exemplo, TransE.

A fim de melhor modelar informação ontológica, e consequentemente melhorar o desempenho de modelos, alguns trabalhos propõem abordagens que limitam o espaço de *embedding* associado ao grafo de conhecimento. Nesse contexto, [Ding et al. 2018] propõem alterações no modelo Complex [Trouillon et al. 2017]. Em particular, eles impõem que os *embeddings* das entidades sejam não negativos e seus valores contidos em $[0,1]^d$. Além disso, eles restringem os valores dos *embeddings* das relações a fim de melhor capturar subsunções aproximadas (e.g., a relação nascido-em usualmente implica na relação nacionalidade). De modo similar, [Fatemi et al. 2019b] adotam uma estratégia para garantir que o modelo SimplE seja capaz de capturar subsunções (e.g., $(X, r_1, Y \rightarrow (X, r_2, Y))$) entre tipos de relação. Em particular, eles impõem que os *embeddings* de entidades sejam não negativos e que o *embedding* de um tipo de relação seja sempre menor ou igual aos *embeddings* das relações que ele subsume.

1.6.3.5. Avaliação e treinamento de modelos

Em sua maioria, o desempenho preditivo de modelos de *embedding* é avaliado por meio dos protocolos de classificação de triplas (*triple classfication*) e ranqueamento de entidades (*entity ranking*), sendo o último mais frequentemente utilizado [Wang et al. 2019c]. Em ambos os protocolos, segundo a prática usual em aprendizado de máquina, a coleção de triplas T do grafo de conhecimento é dividida em três conjuntos disjuntos, nomeadamente, treinamento $T^{(+)}_{train} \subset T$, validação $T^{(+)}_{val} \subset T$ e teste $T^{(+)}_{test} \subset T$. Além disso, como em tarefas de complementação, assume-se que o grafo de conhecimento não abarque proposições falsas, os conjuntos acima contêm apenas triplas tidas como verdadeiras.

O objetivo do protocolo de *classificação de triplas* é testar a habilidade de um modelo ϕ em discriminar triplas verdadeiras das falsas [Wang et al. 2019c]. Esse protocolo está associado, por exemplo, com a tarefa de inferência de triplas. Nesse cenário, a fim de avaliar o modelo, triplas pseudonegativas são geradas. Essa geração pode ser realizada ao substituir de maneira aleatória o sujeito ou objeto de cada tripla de teste por outro elemento do grafo que aparece como sujeito ou objeto respectivamente. Além disso, a tripla (s,r,o) é classificada como verdadeira se o escore $\phi_{(s,r,o)}$ exceder um limiar λ_r dependente do tipo de relação r, o qual é ajustado durante o processo de treinamento e validação do modelo. O desempenho do modelo é medido a partir dos rótulos das triplas de teste por meio de métricas de classificação, incluindo acurácia, precisão e revocação.

No que lhe diz respeito, o objetivo do protocolo de *ranqueamento de entidades* é avaliar o desempenho de um modelo ϕ na inferência de determinadas consultas [Wang et al. 2019c]. Esse protocolo está associado, por exemplo, com a tarefa de predição de ligações. Em particular, para cada tripla de teste $t = (s,r,o)$ duas consultas são produzidas $q_s = (?,r,o)$ e $q_o = (s,r,?)$. Substitui-se "?"em q_s e q_o por cada elemento de interesse x (e.g, entidade) do grafo e ranqueia-se em ordem decrescente, com base nos escores $\phi_{(s,r,x)}$

e $\phi_{(x,r,o)}$, as triplas (s,r,x) e (x,r,o), respectivamente. Com base na posição de cada tripla de teste (s,r,o) nesse *ranking*, o desempenho do modelo é avaliado. Essa avaliação usualmente emprega métricas de recuperação de informação (*information retrieval*), por exemplo, *hits@k* e *mean recipocral ranking*. Além disso, para evitar resultados enganosos, na avaliação usualmente são desconsideradas as triplas (s,r,x) e (x,r,o) presentes no conjunto de treinamento e validação.

Tabela 1.3. Funções de custo utilizadas no treinamento de modelos de *embedding*.

	Erro quadrático	*Hinge*	*Logística*
Pontuais	$\frac{1}{2}\sum_{t\in T_{train}} (\phi_t - y_t)^2$	$\sum_{t\in T_{train}} [\lambda + (-1)^{y_t}\phi_t]_+$	$\sum_{t\in T_{train}} [1+\exp((-1)^{y_t}\phi_t)]_+$
	Hinge		*Logística*
Emparelhadas	$\sum_{t\in T^{(+)}_{train}}\sum_{t'\in T^{(-)}_{train}} [\lambda + \phi_{t'} - \phi_t]_+$		$\sum_{t\in T^{(+)}_{train}}\sum_{t'\in T^{(-)}_{train}} \log(1+\exp(\phi_{t'} - \phi_t))$

Legenda: $[x]_+ = \max(x,0)$, $\lambda \in \mathbb{R}_{\geq 0}$ e $T_{train} = T^{(-)}_{train} \cup T^{(+)}_{train}$.

O aprendizado de modelos de *embedding* envolve a escolha de uma função de custo, a qual geralmente é minimizada por meio do método de Gradiente Descendente Estocástico (ou uma de suas variações). Essa funções consideram os escores dados por um modelo, o valor verdade das triplas, além de restrições (e.g., regularização) associadas aos parâmetros do modelo. Como o conjunto de restrições depende de cada modelo, elas não são apresentadas.

Funções de custo podem ser divididas em pontuais (*pointwise*) ou emparelhadas (*pairwise*) [Mohamed et al. 2019]. Funções de custo pontuais abordam uma tripla por vez. Por exemplo, a função erro quadrático, disposta na Tabela 1.3, mede a diferença quadrada entre o escore ϕ_t e o rótulo $y_t \in \{0,1\}$ de uma tripla de treinamento. Note que y_t é igual a um se a tripla for positiva (pertencer ao grafo de conhecimento) e zero se ela for negativa ou pseudo-negativa. Por sua vez, funções de custo emparelhadas tomam um par contendo uma tripla positiva e uma (pseudo)-negativa. Por exemplo, a função *hinge*, disposta na Tabela 1.3, considera de forma conjunta os escores ϕ_t e $\phi_{t'}$ de uma tripla positiva $t \in T^{(+)}_{train}$ e (pseudo) negativa $t' \in T^{(-)}_{train}$, respectivamente. Por fim, é válido notar que usualmente não são considerados todos os pares de triplas $(t,t') \in T^{(+)}_{train} \times T^{(-)}_{train}$ no cômputo da função de custo, mas sim uma amostra de triplas (pseudo) negativas para cada tripla positiva.

1.7. Considerações Finais

O interesse pela construção, inferência e aplicações de grafos de conhecimento têm florescido nos últimos anos. Esse interesse se deve a diversos fatores, dentre eles, a naturalidade com que conhecimento e informação são dispostos na forma de rede, a abundância de dados heterogêneos, multimodais e multirrelacionais, assim como o surgimento de técnicas que propiciam a construção de bases e grafos de conhecimento de forma cada vez automatizada. Diante disso, neste capítulo foi apresentado de forma introdutória o emprego de técnicas de aprendizado de máquina em tarefas relacionadas a grafos de conhecimento (ver Tabela 1.4). Especialmente, foi apresentado um conjunto de modelos destinados à ta-

refa de complementação, baseados no aprendizado de representações vetoriais para grafos de conhecimento. Dito isto, cita-se oportunidades e desafios de pesquisa em aberto.

Tabela 1.4. Sistemas e modelos de aprendizado de máquina apresentados neste capítulo.

Construção Automatizada	*Sistemas:* • DeepDive [Zhang et al. 2016]. • Fonduer [Wu et al. 2018].
Complementação (Inferência de Fatos)	*Modelos gráficos probabilísticos:* • Redes Lógicas de Markov [Richardson and Domingos 2006] *Modelos de características de grafo:* • *Path Ranking Algorithm* [Lao et al. 2011]. *Modelos de características latentes:* • Distância translacional: TransE [Bordes et al. 2013] e TransH [Wang et al. 2014]. • Correspondência semântica: ANALOGY [Liu et al. 2017], ConvE [Dettmers et al. 2018], JOIE [Hao et al. 2019], MKBE [Pezeshkpour et al. 2018], RESCAL [Nickel et al. 2011], R-GCN [Schlichtkrull et al. 2018] e SimplE [Kazemi and Poole 2018].

Pode-se elencar algumas perspectivas de pesquisa relacionadas ao desenvolvimento de modelos baseados em *embedding* para grafos de conhecimento. Primeiramente, é preciso que se avalie em que nível esse tipo de modelo é capaz de vencer a falta de estruturas simbólicas mais formais, como regras e restrições [Trouillon et al. 2019]; mais ainda, como eles se comparam a métodos que fazem uso dessas estruturas, por exemplo, *Probabilistic Soft Logic* [Bach et al. 2017]. Nesse sentido, há indícios de que o desempenho preditivo desses modelos sofra com problemas de generalização quando o grafo modelado é demasiadamente esparso e/ou ruidoso [Pujara et al. 2017]. Isso indica que o emprego conjunto de diferentes abordagens de inferência é potencialmente mais adequado. Relacionado a isso, como embutir ou considerar conhecimento formal no processo de aprendizado de modelos de *embedding* é de interesse; isso pois alguns modelos de características latentes são incapazes de induzir certas regras lógicas (e.g., subsunções) a partir das asserções presentes no grafo [Gutiérrez-Basulto and Schockaert 2018] e serem, portanto, logicamente consistentes.

Também é relevante o desenvolvimento de modelos latentes que considerem um espectro maior de informação, por exemplo, dinâmica temporal, estruturas diversas do grafo e relações de maior aridade. Apesar de usualmente serem consideradas de forma atemporal, as asserções em grafos de conhecimento costumam ser sensíveis ao tempo. Nesse contexto, considerar a dinâmica evolutiva de entidades e relações pode propiciar tanto o desenvolvimento de melhores modelos quanto novas tarefas e aplicações, por exemplo, a predição temporal de ligações [Trivedi et al. 2017]. Além disso, é de interesse produzir modelos que infiram estruturas mais complexas do grafo de conhecimento, por exemplo, caminhos entre entidades, os quais podem ser vistos como relacionamentos de mais alta ordem. Por sua vez, é relevante que sejam desenvolvidas abordagens latentes capazes de lidar com relações de maior aridade uma vez que parte importante das relações

em bases de conhecimento não são binárias [Fatemi et al. 2019a].

Concernente ao aprendizado, é importante o desenvolvimento de metodologias que gerem *embeddings* para novas entidades sem que se precise aprender novamente as representações vetoriais das entidades já presentes no grafo [Wang et al. 2019b]. Isso é importante uma vez que esse retreino é potencialmente impraticável em aplicações reais onde novas entidades surgem diariamente.

Por fim, são relevantes o desenvolvimento de metodologias para construção de bases de conhecimento a partir de informação multimodal. Há uma grande quantidade de informação em imagem, sensorial e em áudio que raramente é integrada a dados textuais em um repositório de conhecimento comum no qual consultas possam ser realizadas [Dong and Rekatsinas 2018]. Nesse contexto, os métodos de aprendizado profundo possivelmente provejam as ferramentas necessárias para integração multimodal de dados.

Agradecimentos

Os autores agradecem ao CNPq, à FAPERJ e ao CENPES/Petrobras pelo financiamento.

Referências

[Bach et al. 2017] Bach, S. H. et al. (2017). Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research (JMLR)*, 18:1–67.

[Baker et al. 1998] Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley FrameNet project. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics* -. Association for Computational Linguistics.

[Bengio et al. 2013] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

[Bonatti et al. 2019] Bonatti, P. A. et al. (2019). Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371).

[Bordes et al. 2013] Bordes, A. et al. (2013). Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 2787–2795, USA. Curran Associates Inc.

[Bouraoui et al. 2018] Bouraoui, Z., Jameel, S., and Schockaert, S. (2018). Relation induction in word embeddings revisited. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1627–1637, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[Brodie and Mylopoulos 1986] Brodie, M. L. and Mylopoulos, J. (1986). Knowledge bases vs databases. In *On Knowledge Base Management Systems*, pages 83–86. Springer.

[Dettmers et al. 2018] Dettmers, T. et al. (2018). Convolutional 2d knowledge graph embeddings. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Ding et al. 2018] Ding, B. et al. (2018). Improving knowledge graph embedding using simple constraints. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 110–121.

[Dong et al. 2014] Dong, X. L. et al. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *SIGKDD - International Conference on Knowledge Discovery and Data Mining*, pages 601–610.

[Dong and Rekatsinas 2018] Dong, X. L. and Rekatsinas, T. (2018). Data integration and machine learning: A natural synergy. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, pages 1645–1650, New York, NY, USA. ACM.

[Ehrlinger and Wöß 2016] Ehrlinger, L. and Wöß, W. (2016). Towards a definition of knowledge graphs. In *SEMANTiCS (Posters, Demos, SuCCESS)*.

[Fatemi et al. 2019a] Fatemi, B. et al. (2019a). Knowledge hypergraphs: Extending knowledge graphs beyond binary relations. *CoRR*, abs/1906.00137.

[Fatemi et al. 2019b] Fatemi, B., Ravanbakhsh, S., and Poole, D. (2019b). Improved knowledge graph embedding using background taxonomic information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3526–3533.

[Gesese and Russa Biswas 2019] Gesese, G. A. and Russa Biswas, H. S. (2019). A comprehensive survey of knowledge graph embeddings with literals: Techniques and applications. In *Workshop on Deep Learning for Knowledge Graphs*.

[Gutiérrez-Basulto and Schockaert 2018] Gutiérrez-Basulto, V. and Schockaert, S. (2018). From knowledge graph embedding to ontology embedding? an analysis of the compatibility between vector space representations and rules. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR*, pages 379–388.

[Hamilton et al. 2017] Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Representation learning on graphs: Methods and applications. *IEEE Data Eng. Bull.*, 40:52–74.

[Hao et al. 2019] Hao, J. et al. (2019). Universal representation learning of knowledge bases by jointly embedding instances and ontological concepts. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD 19*. ACM Press.

[Jia et al. 2016] Jia, Y. et al. (2016). Locally adaptive translation for knowledge graph embedding. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 992–998. AAAI Press.

[Kazemi and Poole 2018] Kazemi, S. M. and Poole, D. (2018). Simple embedding for link prediction in knowledge graphs. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 4289–4300, USA. Curran Associates Inc.

[Kejriwal 2019] Kejriwal, M. (2019). Advanced topic: Knowledge graph completion. In *Domain-Specific Knowledge Graph Construction*, pages 59–74. Springer International Publishing.

[Lao et al. 2011] Lao, N., Mitchell, T., and Cohen, W. W. (2011). Random walk inference and learning in a large scale knowledge base. In *Proceedings of the Conference on Empirical Methods in Natural Processing*, EMNLP '11, pages 529–539, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Lehmann 1992] Lehmann, F. (1992). Semantic networks. *Computers & Mathematics with Applications*, 23(2-5):1–50.

[Lehmann et al. 2015] Lehmann, J. et al. (2015). Dbpedia – a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

[Lenat 1995] Lenat, D. B. (1995). CYC: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

[Lin et al. 2015] Lin, Y. et al. (2015). Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2181–2187. AAAI Press.

[Liu et al. 2017] Liu, H., Wu, Y., and Yang, Y. (2017). Analogical inference for multi-relational embeddings. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2168–2178, Sydney, Australia.

[Mohamed et al. 2019] Mohamed, S. et al. (2019). A comprehensive survey of knowledge graph embeddings with literals: Techniques and applications. In *Workshop on Deep Learning for Knowledge Graphs*.

[Nickel et al. 2016] Nickel, M. et al. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

[Nickel et al. 2011] Nickel, M., Tresp, V., and Kriegel, H.-P. (2011). A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 809–816, USA. Omnipress.

[Niu et al. 2011] Niu, F. et al. (2011). Tuffy: Scaling up statistical inference in markov logic networks using an RDBMS. *Proc. VLDB Endow.*, 4(6):373–384.

[Noy et al. 2019] Noy, N. et al. (2019). Industry-scale knowledge graphs: Lessons and challenges. *Queue*, 17(2):20:48–20:75.

[Noy and Mcguinness 2001] Noy, N. F. and Mcguinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. Technical report, Standford.

[Pan et al. 2018] Pan, J. Z. et al. (2018). Content based fake news detection using knowledge graphs. In *Lecture Notes in Computer Science*, pages 669–683. Springer International Publishing.

[Paulheim 2017] Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3):489–508.

[Pezeshkpour et al. 2018] Pezeshkpour, P., Chen, L., and Singh, S. (2018). Embedding multimodal relational data for knowledge base completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[Pujara et al. 2017] Pujara, J., Augustine, E., and Getoor, L. (2017). Sparsity and noise: Where knowledge graph embeddings fall short. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1751–1756, Copenhagen, Denmark. Association for Computational Linguistics.

[Raedt et al. 2016] Raedt, L. D., Kersting, K., and Natarajan, S. (2016). *Statistical Relational Artificial Intelligence: Logic, Probability, and Computation*. Morgan & Claypool Publishers.

[Ratner et al. 2018] Ratner, A., Ré, C., and Bailis, P. (2018). Research for practice: Knowledge base construction in the machine-learning era. *Commun. ACM*, 61(11):95–97.

[Ratner et al. 2016] Ratner, A. J. et al. (2016). Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*, pages 3567–3575.

[Rebele et al. 2016] Rebele, T. et al. (2016). YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II*, pages 177–185.

[Richardson and Domingos 2006] Richardson, M. and Domingos, P. (2006). Markov logic networks. *Machine learning*, 62(1-2):107–136.

[Sang et al. 2019] Sang, S. et al. (2019). GrEDeL: A knowledge graph embedding based method for drug discovery from biomedical literatures. *IEEE Access*, 7:8404–8415.

[Schlichtkrull et al. 2018] Schlichtkrull, M. et al. (2018). Modeling relational data with graph convolutional networks. In *The Semantic Web*, pages 593–607, Cham. Springer International Publishing.

[Trivedi et al. 2017] Trivedi, R. et al. (2017). Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3462–3471, International Convention Centre, Sydney, Australia. PMLR.

[Trouillon et al. 2017] Trouillon, T. et al. (2017). Knowledge graph completion via complex tensor factorization. *Journal of Machine Learning Research*, 18(130):1–38.

[Trouillon et al. 2019] Trouillon, T. et al. (2019). On inductive abilities of latent factor models for relational learning. *J. Artif. Int. Res.*, 64(1):21–53.

[van Harmelen et al. 2008] van Harmelen, F., Lifschitz, V., and Porter, B. W., editors (2008). *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*. Elsevier.

[Vrandečić and Krötzsch 2014] Vrandečić, D. and Krötzsch, M. (2014). Wikidata. *Communications of the ACM*, 57(10):78–85.

[Vu et al. 2018] Vu, M. H. et al. (2018). Statistical relational learning with unconventional string models. *Frontiers in Robotics and AI*, 5.

[Wang et al. 2019a] Wang, H. et al. (2019a). Exploring high-order user preference on the knowledge graph for recommender systems. *ACM Transactions on Information Systems*, 37(3):1–26.

[Wang et al. 2019b] Wang, P. et al. (2019b). Logic attention based neighborhood aggregation for inductive knowledge graph embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:7152–7159.

[Wang et al. 2017] Wang, Q. et al. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

[Wang et al. 2019c] Wang, Y. et al. (2019c). On evaluating embedding models for knowledge base completion. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 104–112, Florence, Italy. Association for Computational Linguistics.

[Wang et al. 2014] Wang, Z. et al. (2014). Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 1112–1119. AAAI Press.

[Wu et al. 2018] Wu, S. et al. (2018). Fonduer: Knowledge base construction from richly formatted data. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, pages 1301–1316, New York, NY, USA. ACM.

[Yan et al. 2016] Yan, J. et al. (2016). A retrospective of knowledge graphs. *Frontiers of Computer Science*, 12(1):55–74.

[Yu et al. 2018] Yu, Y. et al. (2018). PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. *Nucleic Acids Research*, 47(D1):D1090–D1101.

[Zhang et al. 2016] Zhang, C. et al. (2016). Extracting databases from dark data with DeepDive. In *Proceedings of the 2016 International Conference on Management of Data - SIGMOD16*. ACM Press.

REALIZATION

EXECUTION

SUPPORT

SILVER SPONSOR

ACADEMIC SUPPORT