REALM: Um Framework Computacional para Investigar os Impactos de Pesquisas Através de Métricas Alternativas

Luís Fernando Monsores Passos Maia[1], **Jonice Oliveira**[1]

[1]Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil

luisfmpm@ufrj.br, jonice@dcc.ufrj.br

Abstract. *In some emergency scenarios or undefined domains, more collaboration among specialists is required. For instance, we can mention the Zika virus, whose epidemic potential became evident in 2014. In Brazil, it had a high occurrence rate, affecting thousands of people and causing overcrowding of public and private emergency services. The social media has been used as the primary alternative to exchange information and create scientific knowledge. The researchers and physicians use social media to communicate their discoveries - abdicating of official and scientific publications - because they needed a faster and proactive way to exchange knowledge. This scenario demands mechanisms to identify experts and to recognize how citizens interpret the efficiency of professionals and their efforts to find solutions. For this purpose, we created a computational framework to identify the social reputation of a researcher or a study, based on alternative impact metrics (altmetrics). To evaluate this framework, we did a Proof of Concept in the Zika context.*

Resumo. *Em alguns cenários de emergência ou em que não há solução conhecida, é necessário mais colaboração entre os especialistas. A exemplo disso tivemos a epidemia de Zika virus, que entre os anos de 2014 e 2016 ganhou destaque internacional, tanto pela sua proporção epidêmica quanto pelas suas consequências. No Brasil, o surto rapidamente evoluiu para uma situação de emergência de saúde pública, exigindo a cooperação dos especialistas e celeridade de resposta. A demanda por resultados rápidos fez com que médicos e pesquisadores publicassem suas descobertas nas mídias sociais, abdicando das publicações científicas oficiais. Cenários como este demandam mecanismos para identificar especialistas e como a população interpreta as soluções por eles criadas. Para este fim, desenvolvemos um framework computacional para medir a reputação social de cientistas e suas pesquisas, com base em métricas de impacto alternativas (altmetrics). Para avaliar o framework, realizamos uma prova de conceito com especialistas no domínio do Zika vírus.*

1. Introdução

Em alguns cenários de emergência ou em que não há solução criada, torna-se necessário uma maior colaboração entre os especialistas. Como exemplo, pode-se citar a epidemia de Zika Virus (ZIKV), cujo potencial epidêmico tornou-se evidente em 2014. No Brasil, a partir de 2015, o surto apresentou uma alta taxa de ocorrências, afetando milhares de pessoas e causando superlotação dos serviços de emergência públicos e privados, embora não tenha sido medido por um sistema de notificação oficial [Zanluca et al. 2015].

Casos de microcefalia e alterações neurológicas em recém-nascidos ocorreram em Pernambuco e em outros estados do Nordeste e, a posteriori, no Sudeste do país, levando o governo brasileiro a declarar estado de Emergência em Saúde Pública de Importância Nacional em novembro de 2015, e posteriormente pela Organização Mundial de Saúde, em 01/02/2016, uma Emergência de Saúde Pública de Importância Internacional (*Public Health Emergency of International Concern*, PHEIC) [Oliveira and Vasconcelos 2016].

Pela urgência de respostas, a ciência se apressou nas investigações. Para garantir que a comunidade científica internacional tivesse condições de tornar públicos os resultados e discussões sobre o tema, foram adotados procedimentos mais rápidos para a aprovação e publicação de artigos sobre o assunto, os chamados *fasttracks*. A demanda pela divulgação de resultados rápidos fez também com que muitos pesquisadores começassem a mostrar seus resultados nas mídias sociais, não esperando o tempo das publicações convencionais [Harmon 2016; McNeil Jr 2016]. Cenários como esse oferecem uma excelente oportunidade para verificar a reputação de especialistas, sua produção científica e como a população interpreta as soluções por eles criadas.

Uma forma eficaz de analisar a produção científica é através de métricas de Análise de Redes Sociais (ARS) e do mapeamento de redes de colaboração científica - também conhecidas como Redes Sociais Científicas (RSC) -, já que atualmente, a colaboração constitui uma característica intrínseca da ciência moderna. Deste modo, a coautoria se apresenta como um importante indicador de colaboração científica na compreensão de diversos fatores relacionados à colaboração entre especialistas [Maia et al. 2018].

Além disso, novas abordagens para avaliar o impacto científico vêm ganhando espaço na medida em que os cientistas mudam seus comportamentos de pesquisa e divulgação para a web [Priem and Hemminger 2010]. Em função disso, métricas alternativas de impacto científico baseadas em mídias sociais estão sendo desenvolvidas e testadas [Priem 2013]. Este novo tipo de medição, também conhecido como Altmetria, consiste em métricas alternativas (*altmetrics*) que permitem mapear a correlação entre os pesquisadores e a sociedade, que vem se estreitando cada vez mais através da troca de experiências, avaliações e conteúdos em mídias sociais, *wikis*, blogs e microblogs, sites de notícias, fóruns de discussão, Redes Sociais On-line (RSO), etc [Bornmann 2014].

Para Priem et al. (2012) a Altmetria pode ser utilizada como uma ferramenta para auxiliar os pesquisadores, não apenas em seus campos de atuação, mas para maximizar a influência e os impactos de suas pesquisas, de modo que seja possível medir sua relevância e contextualizá-la em um universo cada vez mais concorrido de trabalhos científicos.

Deste modo, este trabalho tem como proposta a criação de um *framework* para medir os impactos da ciência na atualidade, tendo em vista entender a representatividade e reconhecimento dos pesquisadores perante a sociedade. O *framework* computacional, denominado REALM (*Researcher Evaluation ALternative Metrics*), visa identificar a reputação social de pesquisadores e suas pesquisas, baseando-se em métricas de impacto alternativas, também conhecidas como *altmetrics* [Priem and Hemminger 2010]. O *framework* foi aplicado no cenário da Zika, onde as questões de pesquisa levantadas foram: “Quem são os pesquisadores mais influentes academicamente?"e "Quem são os pesquisadores com maior inserção na população, possuindo um alto impacto social?”. Os resultados obtidos foram avaliados por especialistas no domínio do ZIKV.

2. O *framework* computacional REALM

O *framework* REALM consiste em uma infraestrutura de *software* que permite a coleta e fusão de dados de publicações em bases de dados indexadas e em mídias sociais com o propósito de extrair e correlacionar diferentes grupos de métricas (produtividade, impacto acadêmico e impacto social). A partir disso é possível identificar, com maior precisão que outros métodos tradicionais (exemplo: número de citações ou h-index), quem são os pesquisadores e/ou grupos de destaque em tópicos de interesse e como esses pesquisadores estão colaborando para engendrar soluções e novas tecnologias. O REALM divide-se em quatro módulos: (a) Coleta e tratamento de dados de publicações acadêmicas; (b) Coleta e tratamento de dados de mídias sociais; (c) Análise do impacto acadêmico; (d) Análise do impacto social.

2.1. Módulo de coleta e tratamento de dados de publicações acadêmicas

Este módulo é responsável pela recuperação de dados de publicações em bases de dados indexadas (exemplo: PubMed[1] e Web of Science[2]) para construção de RSC de coautoria com base em áreas/tópicos de interesse específicos (exemplo: Zika, Dengue e Chikungunya). O módulo opera extraindo das publicações dados como título, nome dos autores, afiliações, data de publicação, identificador do artigo, entre outros. A partir disso ocorrem a separação e tratamento desse conjunto de dados para uma tabela contendo a formatação de grafo, onde os nós representam os pesquisadores e as arestas representam suas publicações em comum. As principais operações realizadas por este módulo são: (i) associação de dois nós (autores), com base no título de uma publicação, caracterizando uma aresta. (ii) Remoção de arestas sem nós associados. (iii) Junção dos nós de autores associados no item (i) em uma única coluna, resultando em um *array* de autores separados por vírgula. (iv) Remoção de itens duplicados. (v) Criação das *labels* dos autores associadas ao *array* do item (iii), resultando na coluna da coautoria. (vi) Verificação do número de vezes que a combinação do item (v) se repete para posterior atribuição de pesos às arestas. (vii) Atribuição de identificadores a cada nó e aresta, possibilitando a leitura e armazenamento dos dados da RSC no banco de dados para posterior visualização do grafo de coautoria e extração de métricas de impacto acadêmico.

2.2. Módulo de coleta e tratamento de dados de mídias sociais

Este módulo é responsável pela coleta, pré-procesamento e triplificação de dados de publicações em mídias sociais como jornais on-line (sites de notícias), blogs científicos, fóruns de discussão e RSO como Facebook, Twitter, Google+, entre outras. Ele corresponde a uma implementação do processo *Extract, Transform, Load* (ETL) descrito em [Maia and Yagui 2017], onde os autores analisaram a repercussão do ZIKV em mídias sociais por ocasião das Olimpíadas Rio 2016. A coleta desses dados ocorre a partir de uma implementação da API do serviço Webhose.io[3], que permite o monitoramento de mídias sociais em tempo real e a coleta de publicações de modo automático 24h/dia. A configuração de '*queries*' específicas no código possibilita reduzir o escopo da coleta para tópicos de interesse específicos, extraindo apenas publicações relacionadas a determinados temas (exemplo: Zika, Dengue e Chikungunya). Após a coleta os dados (não estruturados)

[1]https://www.ncbi.nlm.nih.gov/pubmed/
[2]https://webofknowledge.com/
[3]https://webhose.io/

dessas publicações são convertidos para o formato semi-estruturado (JSON/XML) e enviados para um componente do módulo onde ocorre seu pré-procesamento. Neste ponto são extraídos campos/termos da publicação, tais como URI, título, texto, autor, país, domínio, data, idioma, compartilhamentos em RSO, entre outros, de modo que, a partir desses dados torna-se possível a extração dos índices altmétricos e, consequentemente, das métricas de impacto social. A seguir ocorre a triplificação, que consiste na descrição dos dados através de triplas RDF, seguindo a sintaxe sujeito, predicado e objeto, conforme um modelo de dados RDF[4] adaptado de [Maia and Yagui 2017]. O formato de triplas RDF torna-se uma escolha interessante devido à facilidade em se agregar novas informações ao modelo de dados e, conforme a necessidade, realizar consultas mais elaboradas ou mesmo integrar dados de outros domínios ao modelo proposto. Ao final deste processo ETL é realizada a carga das triplas no banco de dados para posterior extração de índices altmétricos.

2.3. Módulo de análise do impacto acadêmico

Este módulo é responsável pela extração das métricas de produtividade e impacto acadêmico, com base na RSC construída a partir do módulo descrito em 2.1. Assim como o módulo da seção 2.1, este módulo corresponde a uma implementação do método de análise e ranqueamento de pesquisadores descrito em [Maia et al. 2018]. O algoritmo de análise e ranqueamento do REALM permite a análise de RSC em três níveis:

(i) Global - mapeia a rede como um todo, o que permite comparar o comportamento de publicação e colaboração entre pesquisadores de diferentes áreas. Utiliza como parâmetros: número de pesquisadores da rede, número de publicações da rede, número de componentes da rede (sub-redes de nós conectados), somatório de publicações considerando cada pesquisador individualmente, somatório de pesquisadores considerando cada publicação individualmente, média de publicações por pesquisador, média de pesquisadores por publicação e média de colaboração (Grau Médio).

(ii) Local - mapeia as sub-redes existentes, o que permite identificar *clusters* de pesquisadores importantes. Utiliza como parâmetros: número de nós/elementos (NE), Grau Médio, diâmetro e densidade da sub-rede.

(iii) Individual – mapeia pesquisadores influentes a partir do número de publicações (NP) e de sua centralidade na rede, baseada no *Degree* (Grau), *Betweenness* (Intermediação), *Closeness* (Proximidade), e Pagerank. As métricas de centralidade foram escolhidas pela definição de 'prestígio' detalhada em Wasserman e Faust (1994). O prestígio de grau está associado à quantidade de vínculos diretos de um elemento na rede. Quanto mais vínculos o pesquisador tiver na RSC, maior seu prestígio de grau. O prestígio de proximidade considera como mais "centrais" aqueles que possuem uma distância média menor em relação a todos os outros da rede. Pesquisadores que colaboram com elementos mais centrais na RSC possuem proximidade maior. O prestígio de intermediação atribui maior prestígio aos elementos que são pontes, conectando diferentes grupos de pesquisa. Além disso, atribuímos maior status aos elementos mais referenciados na RSC, utilizando a métrica PageRank [Brin and Page 1998]. A visualização da RSC e o cálculo das métricas de centralidade foram adaptados da biblioteca *open source* Cystoscape.js[5].

[4]https://luisfmpm.github.io/realm/datamodel

[5]http://js.cytoscape.org/

Algorithm 1: Algoritmo de análise e ranqueamento do REALM.

```
   Data: array multidimensional sub-rede, NP, NE
   Result: array multidimensional sub-rede_ranqueada
 1 sub-rede_ranqueada ← [];
 2 foreach sub-rede as i do
 3 |   sub-rede[i]['deg'] ← degree(sub-rede[i]['no']) ; sub-rede[i]['bet'] ← betweenness(sub-rede[i]['no']) ;
 4 |   sub-rede[i]['clo'] ← closeness(sub-rede[i]['no']) ; sub-rede[i]['pag'] ← pagerank(sub-rede[i]['no']) ;
 5 end
 6 foreach sub-rede as i do
 7 |   sub-rede[i]['deg_pos'] ← posicao(sub-rede[i]['deg']) ; sub-rede[i]['bet_pos'] ← posicao(sub-rede[i]['bet']) ;
 8 |   sub-rede[i]['clo_pos'] ← posicao(sub-rede[i]['clo']) ; sub-rede[i]['pag_pos'] ← posicao(sub-rede[i]['pag']) ;
 9 end
10 foreach sub-rede as i do
11 |   sub-rede[i]['score'] ← (sub-rede[i]['deg_pos'] + sub-rede[i]['bet_pos'] + sub-rede[i]['clo_pos']) ;
12 end
13 sub-rede_ranqueada ← array_orderby(sub-rede, 'score', SORT_ASC, 'np', SORT_DESC) ; /* Ordena a
       sub-rede de forma crescente pelo score e usa como critério de desempate o maior NP    */
14 foreach sub-rede_ranqueada as i ⇒ var do
15 |   if (sub-rede_ranqueada[var]['np'] < NP) then
16 |   |   unset(sub-rede_ranqueada[i]); /* Remove o Pesquisador do ranking                    */
17 |   end
18 end
19 sub-rede_ranqueada ← array_slice(sub-rede_ranqueada, 0, NE);/* Limita o ranking pelo NE informado */
```

O algoritmo de análise e ranqueamento do REALM[6] (Algoritmo 1) também ordena os pesquisadores em suas respectivas sub-redes (caso haja grafos desconectados) utilizando como parâmetros o NP e as métricas de centralidade, sendo esses parâmetros configuráveis. Exemplo: ordenando somente os 100 primeiros colocados (NE=100) nas quatro métricas de centralidade e que possuam cinco ou mais publicações (NP >=5).

2.4. Módulo de análise do impacto social

Este módulo é responsável pela extração das métricas de impacto social, com base em consultas SPARQL[7] realizadas na base de triplas a partir da interface do sistema. Ele corresponde a uma implementação do método de análise da repercussão social de um pesquisador descrito em [Maia and Oliveira 2017]. A implementação consiste em três grupos de consultas SPARQL que são executadas no banco de triplas Apache Jena Fuseki[8] por meio de requisições HTTP intermediadas pela biblioteca *open source* EasyRDF[9]. Essas consultas são necessárias para a extração de índices altmétricos que permitem medir: (i) o alcance das pesquisas em veículos de comunicação primários (sites de notícias) e secundários (exemplo: blogs e fóruns científicos) (Consulta 1); (ii) sua penetração na população, através da disseminação em RSO como Facebook e Google+ (Consulta 2); e (iii) sua visibilidade a nível global, identificando o país de origem da publicação (Consulta 3). Para isso as consultas utilizam como parâmetros a quantidade de menções a um pesquisador em publicações, a quantidade de menções em publicações compartilhadas em RSO e a quantidade de menções por país, conforme mostrado na Tabela 1.

Conforme as consultas vão sendo realizadas as métricas de impacto social são extraídas e gravadas no banco de dados. A partir disso torna-se possível também indicar em que categoria de reputação um pesquisador se encontra, correlacionando o impacto acadêmico versus impacto social, sendo quatro categorias possíveis:

[6]O pseudocódigo do Algoritmo 1 refere-se ao trecho onde ocorre a análise individual da RSC.

[7]https://www.w3.org/TR/rdf-sparql-query/

[8]https://jena.apache.org/documentation/serving_data/

[9]http://www.easyrdf.org/docs

Tabela 1. Consultas (simplificadas) para extração dos indicadores altmétricos

Prefixos das consultas SPARQL com base no modelo de dados RDF

```
PREFIX ebucore: <https://www.ebu.ch/metadata/ontologies/ebucore/index.html#>
PREFIX schema: <http://schema.org/>
PREFIX realm: <https://luisfmpm.github.io/realm/datamodel#>
PREFIX dbo: <http://dbpedia.org/ontology/>
```

Consulta 1

```
SELECT DISTINCT ?noticia ?texto ?CountFB
?CountGPlus
WHERE {?noticia a ebucore:NewsItem.
       ?noticia schema:text ?texto.
       ?noticia realm:facebookCount ?CountFB.
       ?noticia realm:gplusCount ?CountGPlus.
       FILTER (CONTAINS(LCASE(str(?texto)),
       LCASE("Nome")))}
```

Consulta 2

```
SELECT DISTINCT (COUNT(?noticia) as ?TotalNot)
(SUM(?CountFB) as ?TotalFB)
(SUM(?CountGPlus) as ?TotalGPlus)
WHERE {?noticia a ebucore:NewsItem.
       ?noticia schema:text ?texto.
       ?noticia realm:facebookCount ?CountFB.
       ?noticia realm:gplusCount ?CountGPlus.
       FILTER (CONTAINS(LCASE(str(?texto)),
       LCASE("Nome")))}
```

Consulta 3

```
SELECT DISTINCT (COUNT(?noticia) as ?n)
?pais
WHERE {?noticia a ebucore:NewsItem.
       ?noticia schema:text ?texto.
       ?noticia dbo:country ?pais.
       FILTER (CONTAINS(LCASE(str
       (?texto)),LCASE("Nome")))}
GROUP BY (?pais)
ORDER BY DESC(?n)
```

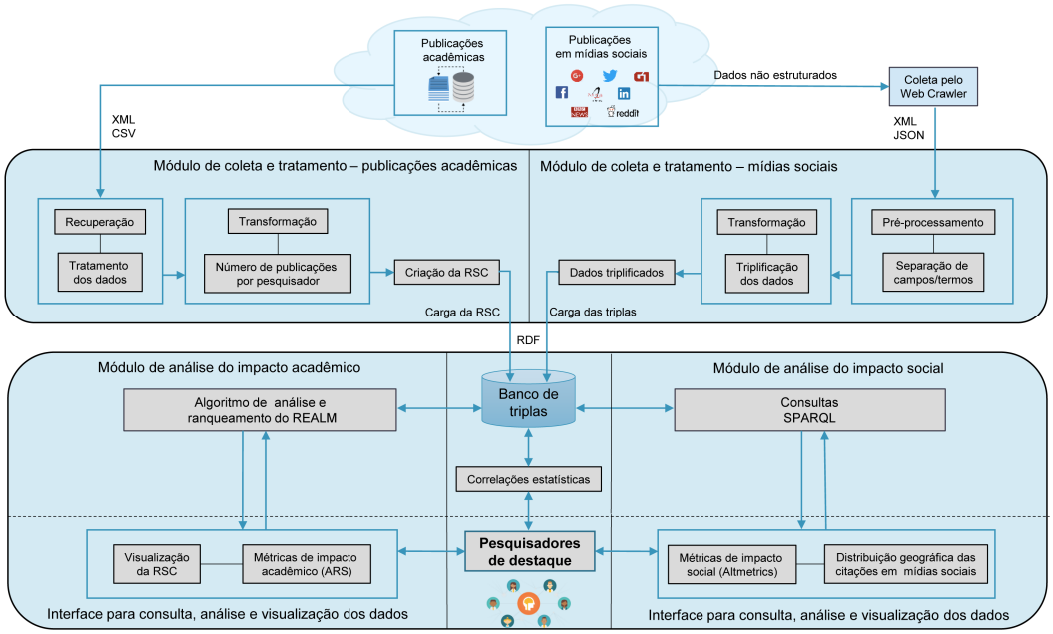

Figura 1. O *framework* computacional REALM

(i) Alto impacto acadêmico e social - Pesquisadores de destaque no cenário. São nomes de grande influência em sua área de atuação, pertencendo a redes de colaboração científica com fortes referências geopolítica/institucional e com forte presença on-line.

(ii) Alto impacto acadêmico - Geralmente são pesquisadores que integram núcleos de pesquisa bem definidos, porém com pouca presença on-line.

(iii) Alto impacto social – São pesquisadores que não possuem uma rede de colaboração bem definida, mas que frequentemente preferem outros meios de compartilhar seus resultados, como *fasttracks* e blogs científicos, o que torna sua divulgação mais prática e célere, sobretudo em mídias sociais.

(iv) Baixo impacto acadêmico e social - Pesquisadores de pouca importância no cenário e pouca ou nenhuma presença on-line.

Além da extração e visualização dos índices altmétricos, a principal contribuição deste módulo é a identificação dos nomes de destaque no cenário. A partir da interface um usuário do sistema pode configurar parâmetros de mapeamento, realizar consultas, acessar métricas gerais da RSC, dos clusters mapeados e dos pesquisadores ranqueados,

visualizar o grafo de colaboração da RSC, visualizar o mapa de citações dos pesquisadores e observar os nomes de destaque no cenário, além de fazer downloads desses dados. A Figura 1 ilustra o *framework* REALM.

3. Estudo de caso Piloto: ZIKV

Como dito na seção 1, a proposta do *framework* é servir como um novo método para entender a representatividade e reconhecimento dos pesquisadores perante a sociedade, a partir de *altmetrics*. Deste modo o *framework* foi testado no cenário do ZIKV, um cenário de emergência que forneceu grande riqueza de dados devido aos surtos ocorridos recentemente. Neste estudo, utilizamos o *framework* para tentar responder as seguintes perguntas: "Quem são os pesquisadores mais influentes academicamente?"e "Quem são os pesquisadores com maior inserção na população, possuindo um alto impacto social?". Findadas nossas análises[10], nosso método foi submetido à prova de conceito, onde especialistas da área verificaram se os pesquisadores mais influentes identificados através do *framework* são, de fato, os nomes de maior reputação no cenário. O estudo divide-se em cinco etapas que serão explicadas a seguir.

Coleta de publicações sobre a Zika na base de dados indexada PubMed e construção da RSC – Nesta etapa utilizamos o mecanismo de consultas[11] do PubMed para recuperar publicações acerca do tema. A partir da *string*[12] "Zika" aplicada nos filtros 'título', '*abstract*' e 'texto' da publicação, recuperamos os dados de 1.932 publicações retroativas a 21/12/2016, nos formatos XML e CSV. A partir desses dados foi possível construir a RSC da temática Zika, conforme as operações descritas em 2.1, para sua posterior leitura, visualização e extração de métricas de impacto acadêmico.

Coleta e triplificação de publicações sobre a Zika em mídias sociais - Nesta etapa o web *crawler* foi configurado para coletar publicações sobre a Zika (*query* "Zika") em sites de notícias, blogs e fóruns de discussão, além de compartilhamentos em RSO como Facebook e Google+, em mais de 100 idiomas e durante o período de 28 de outubro a 28 de dezembro de 2016 (aproximadamente 62 dias de coleta). Neste processo foram coletados, triplificados (conforme descrito em 2.2) e armazenados os dados de 71.898 publicações sobre a Zika para compor uma base de dados temática.

Análise da reputação na Rede Social Científica – Esta etapa visa responder a primeira pergunta: "Quem são os pesquisadores mais influentes academicamente?". Para responder esta pergunta a RSC construída foi analisada em três níveis diferentes.

No primeiro nível foi realizada uma análise global, onde a rede foi verificada como um todo. Neste ponto, foram identificados 6.834 pesquisadores na RSC da doença Zika, onde pesquisadores do tema publicam em média 1,47 artigo, os artigos possuem uma média de 5,39 pesquisadores e a média de colaboração entre eles é de 6,12.

No segundo nível foi realizada uma análise local ou de grupos, onde as sub-redes que se formaram foram verificadas para identificarmos os grupos de pesquisadores mais importantes. Neste nível de análise foram identificados os três *clusters* de pesquisadores

[10]Por razões de escopo neste artigo não mostraremos o mapa da distribuição geográfica das citações.

[11]https://www.ncbi.nlm.nih.gov/pubmed/advanced

[12]((zika[Title/Abstract]) AND zika[Text Word]) AND ("1500/01/01"[Date - Publication] : "2016/12/21"[Date - Publication])

Tabela 2. Categorias de cores baseadas no número de publicações

Categoria	Condição	Categoria	Condição
Vermelha	If NP >= 5 OR NP < 8	Azul	If NP >= 10 OR NP < 15
Roxa	If NP >= 8 OR NP < 10	Verde	If NP >=15

mais importantes, que neste estudo serão referidos como sub-rede 1 (208 nós), sub-rede 2 (133 nós) e sub-rede 3 (96 nós).

No terceiro nível foi realizada uma análise individual, onde foram aplicadas as métricas de centralidade para identificação dos pesquisadores mais influentes dentro das três sub-redes. Para a análise da reputação de um pesquisador no cenário científico, foram utilizados os seguintes parâmetros: número de publicações e centralidade (*Betweenness*, *Closeness*, *Degree* e Pagerank) de cada pesquisador.

Com relação ao número de publicações é importante frisar que este critério deve ser levado em conta nas análises, pois a quantificação da produtividade, a despeito de críticas ao fato, também é um fator essencial para determinarmos se um pesquisador está conduzindo avanços em seu campo de atuação ou em focos específicos (para este estudo de caso, a doença Zika). Deste modo, o mapeamento foi configurado de modo a considerar somente pesquisadores com cinco ou mais publicações, descartando pesquisadores com baixa produção bibliográfica e reduzindo o escopo da próxima análise. Além disso, como forma de facilitar na identificação dos pesquisadores mais produtivos e melhorar a visualização desses números, foram definidas quatro categorias de cores baseadas no NP de cada pesquisador, conforme os critérios definidos na Tabela 2.

Definidos esses critérios, os pesquisadores mais influentes foram ranqueados pelo algoritmo com base nas métricas de centralização de Freeman [Freeman 1978] (*Betweenness*, *Closeness* e *Degree*), a partir do somatório das posições, e individualmente na métrica Pagerank. A última métrica é utilizada como critério de comparação, já que indica se um pesquisador está relacionado com nós que são bastante referenciados na RSC. Estes dois *rankings* encontram-se disponíveis em: https://goo.gl/tAjftd e https://goo.gl/3nLG4c.

Análise da repercussão social de um pesquisador - Esta etapa está relacionada à resolução da segunda pergunta: "Quem são os pesquisadores com maior inserção na população, possuindo um alto impacto social?" Para isso foram realizados dois tipos distintos de consulta, extraindo os índices altmétricos a partir de: (i) veículos de comunicação, para verificar se estão noticiando os avanços e descobertas de um pesquisador em relação ao ZIKV - para isto, utilizamos o total de menções (em sites de notícias, fóruns e blogs) a um pesquisador. (ii) RSO, para verificar a disseminação e alcance das publicações na população – para isto, analisamos a propagação da publicação sobre um pesquisador e suas descobertas a partir de menções no Facebook e no Google+.

Essas duas consultas indicam se: (i) os veículos de comunicação mais importantes (sites de notícias) e os secundários (blogs e fóruns científicos, por exemplo) estão noticiando os avanços e descobertas de um pesquisador em relação à doença. (ii) se essas notícias estão se disseminando pelas RSO e alcançando um público mais amplo.

Todavia, o nome de citação do PubMed é insuficiente para identificar os registros de citações de um pesquisador. Diante disso, foram utilizadas grafias alternativas para auxiliar nas consultas. Por exemplo, para o pesquisador DIAMOND MS, foram identifi-

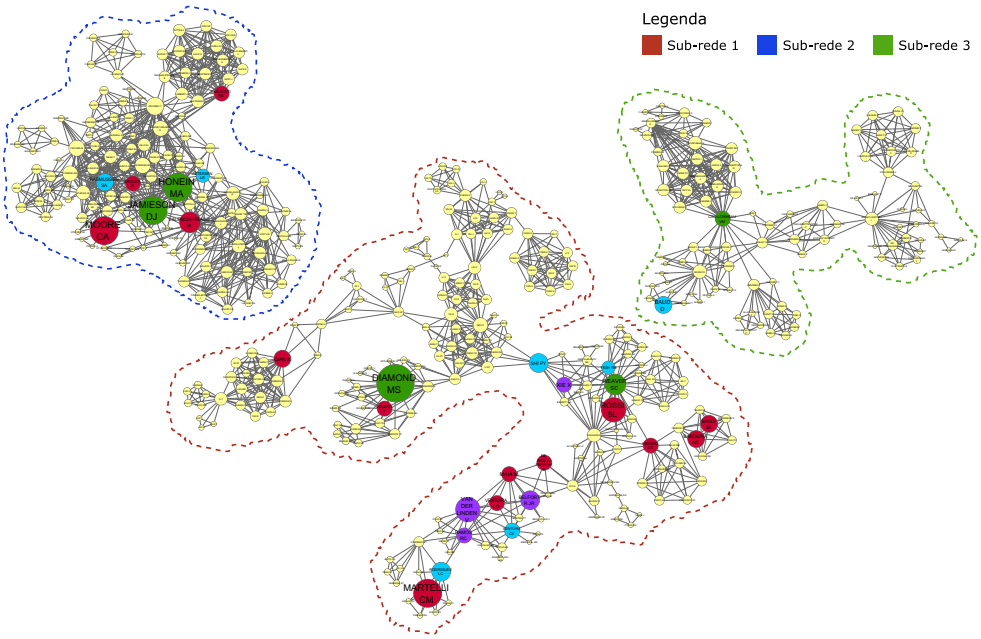

Figura 2. Pesquisadores de destaque no cenário da Zika

cadas três grafias diferentes na base de triplas, 'DIAMOND MS', 'Michael S. Diamond' e 'Michael Diamond', de modo que na primeira consulta as grafias foram testadas individualmente e fazendo uso da cláusula '*distinct*' para contabilizar somente uma referência por publicação. Com isso foram retornados, respectivamente, 6, 24 e 683 resultados para as três grafias identificadas (713 citações). Deste modo, a consulta foi realizada utilizando o operador || (OR) dentro da cláusula '*filter*' e em conjunto com '*distinct*' para retornar, em uma única consulta, os resultados das três grafias para o nome do pesquisador e garantir que não há resultados repetidos. Novamente foram retornados 713 resultados.

A partir disso, realizamos essas duas consultas para os pesquisadores identificados na etapa de análise da reputação na RSC, utilizando a cláusula '*Order by*' para ordená-los de acordo com o volume de menções em publicações (em sites de notícias, blogs e fóruns), e utilizando as menções em RSO (Facebook e Google+) como critério de desempate.

A partir da ordenação dos pesquisadores com base nesses critérios de classificação, foi criado um *ranking* altmétrico unificado das três sub-redes, que encontra-se disponível em https://goo.gl/8QyN6a.

Pesquisadores de destaque no cenário - Com os resultados obtidos também foi possível verificar a correlação entre impacto acadêmico e impacto social. A verificação ocorre a partir da correlação entre as variáveis *Betweenness*, *Closeness*, *Degree*, PageRank e citações em mídias sociais com resultados que variam numa escala entre 0 e 1, para o grupo das métricas de centralização versus as citações individuais na base de dados temática. Quanto mais próximo de 1, maior a correlação entre impacto acadêmico e impacto social. Deste modo, identificamos os 30 pesquisadores de destaque no cenário da Zika (disponíveis em: https://goo.gl/SMzx2b). A Figura 2 ilustra a RSC em Zika, onde estão presentes os 30 nomes de destaque no cenário, conforme as categorias de cores da Tabela 2 e o tamanho do nó variando conforme a importância do pesquisador.

4. Prova de conceito

Para avaliar o *framework*, foi realizada uma prova de conceito. Usando o *framework*, foram criados dois *rankings*, contendo: i) os 20 pesquisadores com maior influência científica e ii) os 20 pesquisadores com maior influência social.

As duas listas, ordenadas de maneira decrescente em relação às respectivas influências, foram apresentadas a três pesquisadores especialistas no domínio, participantes da Rede Zika de Ciências Sociais (chefiada pela Fiocruz) e da ZIKAlliance (consórcio internacional). Para não serem influenciados, foram apresentados a eles apenas as listagens. Nenhuma explicação sobre o *framework* foi dada. Os pesquisadores avaliaram separadamente os dois *rankings* e concordaram integralmente com as ordenações apresentadas.

Um dos pesquisadores (participante do primeiro grupo), que estuda e documenta a evolução da doença no Brasil e no mundo, justificou cada uma das posições dos *rankings*. Tal explicação foi feita baseando-se no seu conhecimento tácito e informações sobre a evolução científica da doença. Este pesquisador mostrou as influências de cada pesquisador (no contexto científico ou social) e demonstrou as diferenças entre suas relevâncias.

5. Trabalhos relacionados

Estudos empíricos no campo da Altmetria podem se basear em diferentes grupos de plataformas que permitem a extração de diferentes grupos de métricas de impacto acadêmico e social. Isso é natural se pensarmos que, fundamentalmente, as métricas alternativas fazem uso de índices que relacionam a quantidade de menções a cientistas e/ou suas pesquisas em diferentes plataformas.

Entre essas plataformas, podemos citar RSO como Facebook e Google+, blogs científicos (ou blogs em geral) e gerenciadores de referências bibliográficas como o Mendeley e CiteULike. Esta forma de quantificar a ciência está amparada por diversos estudos empíricos que visam demonstrar de maneira efetiva o impacto das pesquisas acadêmicas na sociedade em geral. Esta perspectiva é abordada nos trabalhos de: (i) Bornmann (2015), com a utilização de três tipos de plataformas, sendo a primeira a RSO Twitter, a segunda os gerenciadores de referências bibliográficas Mendeley e CiteULike, e a terceira blogs científicos; (ii) Hassan e Gillani (2016), que propuseram um estudo altmétrico baseado em diversas plataformas, como, Google Scholar, Twitter, Mendeley, Facebook, Google+, CiteULike, blogs e *Wiki*; (iii) Kwak e Lee (2014), que utilizaram o Twitter; (iv) Mohammadi et al.(2015) na plataforma Mendeley e (v) Hoffman et al. (2014) que utilizaram o Researchgate.

Neste sentido outro trabalho que merece destaque no campo da Altmetria é o Altmetric.com[13], sendo atualmente a ferramenta mais popular no que se refere a este tipo de medição. O Altmetric.com tem o propósito de rastrear e analisar a atividade on-line no que diz respeito à literatura acadêmica fornecendo feedback de dados de aproximadamente 5 milhões de *papers*. Seus serviços incluem a extensão Altmetric Bookmarklet[14], que quando instalada no navegador oferece métricas a nível de artigo, bastando para isso navegar até a página onde o artigo se encontra e clicar o botão "Altmetric it!" nos favoritos. O Altmetric Bookmarklet, no entanto, funciona apenas em artigos de páginas do PubMed, arXiv ou que possuam DOI.

[13]https://www.altmetric.com/

[14]https://www.altmetric.com/products/free-tools/bookmarklet/

Embora existam diversas ferramentas altmétricas disponíveis, as mais populares, como o Altmetric.com e o Altmetric Bookmarklet exigem o pagamento de taxas para sua utilização ou não satisfazem as necessidades de pesquisadores e instituições acadêmicas. Este é um cenário que incentiva o desenvolvimento de novas ferramentas altmétricas para atender às demandas mais específicas. Neste sentido, o *framework* apresentado se destaca dos demais trabalhos por suprir essa demanda através de uma infraestrutura de aplicações que permitem analisar a reputação de pesquisadores e pesquisas através de diferentes grupos de métricas, de maneira conjunta e com maior precisão do que outros métodos que utilizam somente um tipo de medição. Outra vantagem é a possibilidade de comparar a evolução de áreas a partir de aspectos temporais e macro das RSC construídas.

6. Conclusão

Neste trabalho apresentamos o framework REALM, que permitiu a extração conjunta de diferentes tipos métricas para avaliar a reputação de pesquisadores no cenário da Zika. A extração de três grupos de métricas a partir de uma única ferramenta representa um avanço no que tange a medição dos impactos da ciência, visto que métricas de produtividade, influência acadêmica e impacto social, sozinhas, podem não ser suficientes para evidenciar isso. O alto número de compartilhamentos em alguns casos indica que a pesquisa teve grande repercussão nas RSO (penetração na população). Estes casos referem-se a publicações que reportam progressos em estudos relacionados ao ZIKV e outras descobertas científicas importantes, como novos tratamentos e a proximidade de uma cura. Notamos que essas publicações remetem a estudos desenvolvidos por cientistas de alto prestígio acadêmico e social, sendo integrantes de núcleos de pesquisa bem definidos e de forte referência geopolítica e institucional, conforme corroborado na prova de conceito.

Além do próprio *framework*, este estudo fornece uma contribuição para o cenário da pesquisa em Zika, visto que até o momento não há mapeamentos de RSC acerca da doença a nível micro/individual, ou seja, analisando em profundidade as interações científicas sobre o tema. O *framework* desenvolvido para investigar essas questões se baseia em um conjunto de abordagens sistemáticas que, com a extração de três tipos de medição distintas e sua aplicação combinada, permite avaliar melhor os impactos de pesquisadores, sua produtividade e influência na comunidade acadêmica e na sociedade.

Outra característica do REALM é o uso de conceitos da web semântica para fusão e registro de dados, permitindo consultas mais direcionadas ao conteúdo pretendido. Além disso, diferente dos bancos relacionais, essa abordagem possibilita a inserção de novas categorias conforme a necessidade de ampliar o modelo e reorganizar as informações, permitindo flexibilidade dos dados. Por exemplo, agregando triplas de publicações sobre novas doenças não é necessário alterar a estrutura dos dados já registrados no banco.

Como trabalhos futuros, pretende-se estender o modelo do grafo RDF de modo a analisar impactos de pesquisas em outros cenários como a Dengue e Chikungunya e com maior volume de dados. Também é pretendido investigar o engajamento com o público, ou seja, quantas pessoas e que tipo de audiência lê e publica sobre esses temas.

Referências

Bornmann, L. (2014). Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000prime. *J. Informetr.*, 8(4):935–950.

Bornmann, L. (2015). Alternative Metrics in Scientometrics: A Meta-analysis of Research into Three Altmetrics. *Scientometrics*, 103(3):1123–1144.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. and ISDN systems*, 30(1):107–117.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Soc. Networks*, 1(3):215–239.

Harmon, A. (2016). Handful of Biologists Went Rogue and Published Directly to Internet. *The NYT*.

Hassan, S.-U. and Gillani, U. A. (2016). Altmetrics of "altmetrics"using Google Scholar, Twitter, Mendeley, Facebook, Google-plus, CiteULike, Blogs and Wiki. *arXiv:1603.07992 [cs]*.

Hoffmann, C. P., Lutz, C., and Meckel, M. (2014). Impact Factor 2.0: Applying Social Network Analysis to Scientific Impact Assessment. In *Proceedings of the 47th Int. Conf. on Syst. Sciences*, pages 1576–1585. IEEE.

Kwak, H. and Lee, J. G. (2014). Has Much Potential but Biased: Exploring the Scholarly Landscape in Twitter. In *Proceedings of the 23rd Int. Conf. on World Wide Web*, pages 563–564, New York, NY, USA. ACM.

Maia, L. F. M. P., Lenzi, M., Rabello, E. T., and Oliveira, J. (2018). Colaborações científicas em Zika: Identificação dos principais grupos e pesquisadores através da análise de redes sociais. *Cad. de Saúde Pública*.

Maia, L. F. M. P. and Oliveira, J. (2017). Investigation of research impacts on the Zika virus. An approach focusing on social network analysis and altmetrics. In *Proceedings of the 23rd Brazillian Symp.on Multimedia and the Web*, Gramado.

Maia, L. F. M. P. and Yagui, M. M. M. (2017). Triplificação de dados de notícias sobre a Zika. In *Proceedings of the XIII Brazilian Symp. on Information Systems*, Lavras.

McNeil Jr, D. G. (2016). Zika Data From the Lab, and Right to the Web. *The NYT*.

Mohammadi, E., Thelwall, M., Haustein, S., and Larivière, V. (2015). Who reads research articles? An altmetrics analysis of Mendeley user categories. *J Assn Inf Sci Tec*, 66(9):1832–1846.

Oliveira, C. S. and Vasconcelos, P. F. C. (2016). Microcephaly and Zika virus. *J. Pediatr.*, 92(2):103–105.

Priem, J. (2013). Scholarship: Beyond the paper. *Nature*, 495(7442):437–440.

Priem, J., Groth, P., and Taraborelli, D. (2012). The Altmetrics Collection. *PLoS One*.

Priem, J. and Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, 15(7).

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

Zanluca, C., Melo, V. C. A., Mosimann, A. L. P., et al. (2015). First report of autochthonous transmission of Zika virus in Brazil. *Mem. Inst. Oswaldo Cruz*, 110(4):569–572.