Contextual Information Based Community Detection in Attributed Heterogeneous Networks

M. Dias, P. Braz, E. Silva, and R. Goldschmidt

***Abstract*—Community detection is an important network analysis task that has been studied by academy and industry for the last years. Community detection algorithms try to maximize the number of connections in each community and minimize the number of connections between different communities. Some of them consider not only the topological aspects of the networks but also try to explore existing information about the context of the application available in attributes of nodes and/or connections in order to find cohesive content communities. Those algorithms were designed to run exclusively over homogeneous networks and cannot deal with heterogeneous structures. Nevertheless, typical real-world networks are heterogeneous. Thus, this article proposes ComDet, a community detection approach that fills this gap by taking into account topological and contextual information to detect communities in heterogeneous networks. The proposed approach uses data clustering as a pre-processing step for the community detection process in order to identify similar nodes that are directly or indirectly linked and organize them in cohesive and possibly overlapping communities. Experiments in three attributed heterogeneous networks show that ComDet leads to interesting partitions with cohesive content communities.**

***Index Terms*—Community Detection, Data Clustering, Heterogeneous Networks, Attributed Graphs.**

I. Introdução

NOS últimos anos, academia e indústria têm dedicado grande atenção à análise de redes complexas. Uma rede complexa é um multigrafo altamente conectado, possivelmente contendo atributos[1], onde um vértice (nó) representa um item da rede (e.g., pessoa, postagem, artigo, etc.) e uma aresta representa algum tipo de associação entre os itens correspondentes (e.g., amizade ou comunicação entre duas pessoas).

Identificar grupos de nós densamente interconectados é uma tarefa de análise de redes complexas conhecida como detecção de comunidades [2]. Em essência, essa tarefa é um problema de otimização que tenta organizar os elementos em grupos (comunidades), de forma a maximizar o número de arestas entre vértices de um mesmo grupo e minimizar a quantidade de arestas entre vértices de grupos distintos [1]. Inicialmente usada para identificar grupos de pessoas em redes sociais, a detecção de comunidades tem sido aplicada em diversas áreas como marketing, educação e segurança, dentre outras [1].

Guiada pela essência topológica da definição da tarefa, a maioria dos algoritmos de detecção de comunidades considera somente aspectos estruturais das redes. Alguns deles também levam em conta informações sobre o contexto da aplicação disponíveis nos atributos de vértices e arestas do grafo a fim de detectar comunidades em que os elementos nelas alocados apresentem algum tipo de coesão de conteúdo [29]. Inspirados em princípios de homofilia, tais algoritmos tentam identificar comunidades coesas por meio do agrupamento de vértices e conexões que compartilhem informações similares.

Em geral, os algoritmos de detecção de comunidades que exploram informação contextual disponível nas redes complexas foram projetados para serem aplicados em redes homogêneas (i.e., redes que contém somente um tipo de vértice e um tipo de aresta). Tal tendência decorre da definição da tarefa de detecção que pressupõe sua aplicação em grafos homogêneos. Por outro lado, os algoritmos de detecção de comunidades que são capazes de lidar com redes heterogêneas (i.e., aquelas que contêm vários tipos de vértices e/ou arestas) não levam em conta informação contextual. No entanto, a maior parte das redes do mundo real são heterogêneas e contêm informações em sua estrutura. Além de serem semanticamente mais ricas do que as redes homogêneas, as redes heterogêneas fornecem ligações indiretas entre vértices que podem ser exploradas pelas pesquisas e aplicações em detecção de comunidades.

Diante do exposto, este artigo tem como objetivo propor uma abordagem de detecção de comunidades em redes heterogêneas que leva em consideração tanto informações topológicas quanto informações contextuais. Chamada ComDet, a abordagem proposta utiliza agrupamento de dados (do inglês, *data clustering*) como uma etapa prévia ao processo de detecção de comunidades, a fim de identificar vértices similares que estejam direta ou indiretamente conectados para, então, organizá-los em comunidades estruturalmente densas, de conteúdo coeso e possivelmente sobrepostas. O agrupamento de dados garante a coesão de conteúdo entre as informações dos vértices e das arestas de cada comunidade detectada. Nos experimentos realizados, avaliações quantitativas e qualitativas com três redes heterogêneas mostram que a utilização da ComDet leva a comunidades mais densas e com conteúdos mais coesos em cada uma delas do que as obtidas por algoritmos que se baseiam exclusivamente na topologia das redes.

Este texto contém mais cinco seções. A seção II resume conceitos sobre redes heterogêneas. A seção III apresenta os trabalhos relacionados. O detalhamento conceitual da ComDet encontra-se na seção IV. A seção V descreve os experimentos realizados e analisa os resultados obtidos. Conclusões e trabalhos futuros encontram-se indicados na seção VI.

II. Redes Heterogêneas

Em geral, nos trabalhos científicos, redes complexas são representadas por meio de grafos. Aquelas com um (resp.

[1]Um grafo com atributos se caracteriza por apresentar atributos em nós e/ou arestas que contêm informações sobre o contexto da aplicação. Exemplos: palavras-chave e data de publicação de um artigo, gênero e renda de uma pessoa, preço de um produto, coordenadas geográficas, etc.

dois ou mais) tipo(s) de nó são chamadas unimodais (resp. multimodais). De forma análoga, redes com um (resp. dois ou mais) tipo(s) de aresta são chamadas unidimensionais (resp. multidimensionais). Redes unimodais e unidimensionais são denominadas homogêneas. Neste trabalho, estamos particularmente interessados em *redes heterogêneas*, isto é, estruturas multimodais e/ou multidimensionais.

Uma rede heterogênea pode ser representada por dois grafos: esquema (G_S) e instância (G_I). O primeiro especifica a estrutura conceitual e a semântica dos elementos da rede. Indica metadados, como tipos de nó, tipos de aresta e tipos de informação (atributos) que a rede contém. Por outro lado, G_I representa um retrato da rede em um determinado momento, mostrando as instâncias existentes de nós, arestas e os valores de seus atributos. A figura 1 apresenta um exemplo de rede heterogênea por meio de esquema e instância.[2] Tal exemplo é um multigrafo heterogêneo com atributos. Contém diferentes tipos de nós e arestas. Seus nós possuem atributos com informações contextuais. Além disso, alguns nós são interligados por duas ou mais arestas. Esse tipo de grafo (com atributos, heterogêneo e múltiplo) é frequentemente usado para representar instâncias de redes do mundo real.

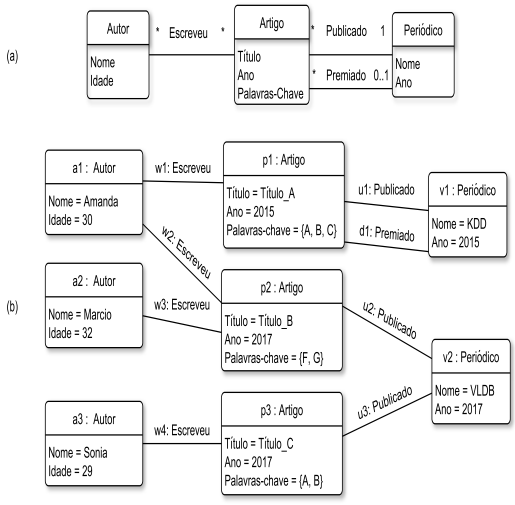

Fig. 1. Exemplo de rede heterogênea. (a) esquema; (b) instância.

Formalmente, o esquema e a instância de um multigrafo heterogêneo podem ser definidos respectivamente por $G_S = (V_S, E_S, X, \beta)$ e $G_I = (V_I, E_I, \alpha, \sigma)$, onde: V_S e V_I são os conjuntos de tipo e de instância de nó, respectivamente; E_S e E_I são os conjuntos de tipo e de instância de aresta; X é o conjunto de atributos; β é uma função que mapeia um tipo de elemento de $V_S \cup E_S$ em uma lista ordenada formada pelos atributos desse tipo; α é uma função que mapeia um elemento de $(V_I \cup E_I)$ no tipo desse elemento $(V_S \cup E_S)$; e σ é uma função que mapeia uma aresta e no par de nós que e conecta. No exemplo da figura 1, o esquema e a instância da rede são formalmente caracterizados pelos itens abaixo. $\beta(Autor) = (Nome, Idade)$, $\alpha(u_2) = Publicado$ e $\sigma(u_3) = (p_3, v_2)$ ilustram o uso das funções de mapeamento.

- V_S={$Autor$, $Artigo$, $Periódico$}
- $E_S = \{Escreveu, Publicado, Premiado\}$
- $X = \{Nome, Idade, Título, Ano, Palavras\text{-}chave\}$
- $V_I = \{a_1, a_2, a_3, p_1, p_2, p_3, v_1, v_2\}$
- $E_I = \{w_1, w_2, w_3, w_4, u_1, u_2, u_3, d_1\}$

Redes heterogêneas são semanticamente mais ricas do que as homogêneas. Diferentes tipos de nó e de aresta podem levar a diferentes análises de dados. Por exemplo, enquanto que em uma rede homogênea, a análise para detecção de comunidades é restrita a ligações diretas entre nós, em uma estrutura heterogênea, essa análise pode explorar conexões indiretas. Para lidar com essa diversidade, o conceito de caminho semântico (ou meta-caminho) foi proposto em [13]. Este conceito é usado para representar um conjunto de caminhos da instância da rede (denominados *instâncias de caminho*) que conectam direta ou indiretamente nós de dois tipos. Formalmente, um caminho semântico é um caminho definido a partir de um esquema de rede. Dados dois tipos de nó arbitrários, A_1 e A_k, um caminho semântico P que os conecta é denotado por uma sequência $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_{k-1}} A_k$, onde $A_i \in V_S$ e $R_j \in E_S$. A_1 (resp. A_k) é chamado de tipo de nó de origem (resp. tipo de nó alvo) de P. Para toda instância de caminho p, onde $p \in P$, o nó origem (resp. alvo) de p pertence ao tipo de nó A_1 (resp. A_k). $S : Autor \xrightarrow{Escreveu}$ Artigo $\xrightarrow{Publicado}$ *Periódico* é um exemplo de caminho semântico onde *Autor* e *Periódico* são, respectivamente, tipos de nós de origem e alvo. $a_3 \xrightarrow{w_4} p_3 \xrightarrow{u_3} v_2$ é um exemplo de instância de caminho de S.

III. Trabalhos Relacionados

Basicamente, os algoritmos de detecção de comunidades podem ser divididos em dois grupos. Orientados pela essência topológica da definição da tarefa de detecção, os algoritmos do primeiro grupo se concentram exclusivamente nos aspectos estruturais da rede e tentam maximizar o número de conexões em cada comunidade, minimizando o número de conexões entre comunidades distintas. Além dos aspectos topológicos das redes, os algoritmos do segundo grupo tentam explorar informações existentes sobre o contexto da aplicação disponíveis em nós e/ou arestas. Com base em princípios de homofilia, esses algoritmos tentam identificar comunidades coesas agrupando nós e arestas com informações semelhantes. A Tabela I resume esta taxonomia. Por meio dela, pode-se perceber que a maioria dos algoritmos de detecção se limita a tratar exclusivamente redes homogêneas. Tal limitação também decorre da definição da tarefa de detecção de comunidades que pressupõe sua aplicação em grafos homogêneos. No entanto, embora muitas redes do mundo real sejam heterogêneas, poucas são as iniciativas de pesquisa voltadas à detecção de comunidades em grafos heterogêneos. Mais restrito ainda é o conjunto de trabalhos que, como a ComDet, são capazes de lidar com redes heterogêneas considerando tanto aspectos topológicos quanto informações sobre o contexto da aplicação. Tais trabalhos encontram-se comentados a seguir.

O trabalho descrito em [29] divide o grafo em k grupos, de modo que cada grupo contenha um subgrafo densamente

[2]Neste trabalho, adotamos diagramas de classes e objetos da UML para representar esquema e instância de redes heterogêneas, respectivamente.

TABLA I
TRABALHOS RELACIONADOS - VISÃO COMPARATIVA

Grupos	Redes Homogêneas	Redes Heterogêneas
Somente Topológico	[19], [20], [22], [23], [24], [11], [12]	[13], [14], [5], [15]
Topológico e Contextual	[16], [4], [17], [18], [24], [25], [26]	[29], [27], [3]

conectado. Diferente da ComDet, a proposta apresentada em [29] não é capaz de encontrar comunidades sobrepostas. Trata-se de uma limitação importante, uma vez que diferentes vértices podem pertencer a diferentes comunidades simultaneamente. Por exemplo, uma mesma pessoa pode pertencer a mais de uma comunidade ao mesmo tempo.

[27] propõe um modelo probabilístico para a detecção de comunidades que mistura a relação entre vértices, o tipo de interação e a informação trocada com membros de outras comunidades. Da mesma forma que o trabalho anterior, esta proposta não detecta comunidades sobrepostas. Além disso, o usuário precisa informar comunidades previamente existentes e o número de tópicos (assuntos) a ser considerado pelo modelo, o que nem sempre é viável na prática.

Por fim, [3] propõe o RM-CRAG, um algoritmo de agrupamento de grafos com atributos. Para um valor k dado pelo usuário, o RM-CRAG gera os *top-k* agrupamentos (possivelmente sobrepostos), nos quais esses agrupamentos são distintos uns dos outros (não redundantes). O RM-CRAG não trata atributos nas arestas, diferentemente da ComDet. Esta é uma limitação importante, pois, de forma similar ao exemplo da figura 1, muitas redes contêm atributos vinculados às arestas que podem ser relevantes no processo de detecção.

IV. ABORDAGEM PROPOSTA

A ComDet, abordagem proposta neste trabalho, combina o agrupamento de dados contextuais com a detecção de comunidades tradicional baseada em topologia. Seu processamento é dividido em seis etapas, conforme ilustrado na Figura 2. A rede heterogênea com atributos fornecida como entrada para a ComDet deve ser representada por um grafo esquema $G_S = (V_S, E_S, X, \beta)$ e um grafo instância $G_I = (V_I, E_I, \alpha, \sigma)$.

Na primeira etapa, o analista de dados deve configurar o experimento escolhendo os seguintes itens: (a) Os algoritmos a serem usados pelas etapas 4 e 5. Potencialmente quaisquer algoritmos de agrupamento de dados e de detecção de comunidades podem ser utilizados nas respectivas etapas; (b) Um tipo de nó de referência t de V_S. Essa escolha determina quais nós de G_I (aqueles cujo tipo é t) devem ser considerados pelo algoritmo de detecção de comunidades. Embora a rede de entrada possa ser heterogênea, como consequência natural dessa escolha, as comunidades detectadas ao final do processo somente envolverão nós de G_I que sejam do tipo t (grafos homogêneos). Isso é possível porque a ComDet usa um caminho semântico fornecido pelo usuário (veja abaixo) e o processo de agrupamento de dados para identificar nós que estão indiretamente ligados em G_I e que, portanto, devem estar na mesma comunidade. A escolha de t depende da

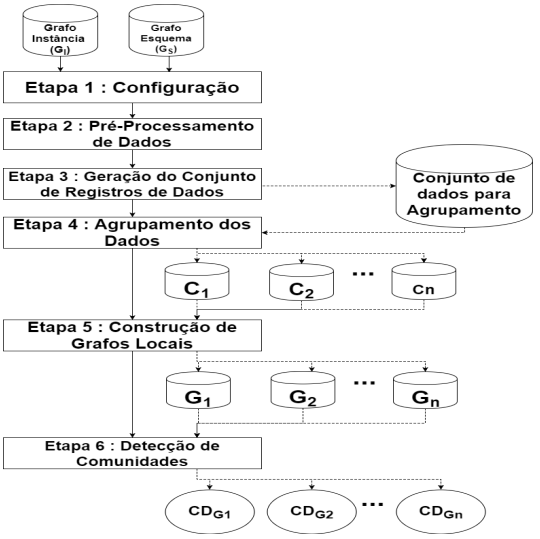

Fig. 2. Etapas da ComDet: Visão Geral do Processo.

interpretação do analista sobre que tipo de vértice pode levar a comunidades que sejam de interesse sob o ponto de vista do domínio da aplicação; (c) Os atributos de X que devem ser considerados pelo processo de agrupamento. Essa escolha é muito importante porque determina a forma como os dados devem ser agrupados. Diferentes opções de atributos podem levar a diferentes agrupamentos e, portanto, a diferentes visualizações de dados. Potencialmente, qualquer elemento não vazio de 2^X pode ser escolhido; (d) O caminho semântico P que deve orientar o processo. Deve incluir, pelo menos, o tipo de nó de referência t como tipo de nó origem de P e todos os tipos associados aos atributos escolhidos. A ComDet apresenta para escolha do analista todos os caminhos semânticos que atendem às restrições mencionadas.

A etapa 2 compreende atividades de preparação de dados tais como normalização, codificação, imputação de dados ausentes, lematização de texto, entre outros. A escolha de quais atividades devem ser realizadas depende dos atributos contextuais previamente selecionados e da opinião do analista.

A etapa 3 é responsável por converter os dados provenientes de valores dos atributos de G_I em uma estrutura relacional. A conversão é baseada no caminho semântico P e nos atributos previamente escolhidos. P determina quais combinações de instâncias de nó e instâncias de aresta devem gerar os registros de dados. Cada instância de caminho de P produz um registro de dados. Para descrever formalmente o algoritmo de conversão, primeiro definimos alguns operadores, todos inspirados na Álgebra Relacional [21]. Considere que os operadores apresentados abaixo são definidos no contexto de uma rede heterogênea dada pelo par (G_S, G_I). Além disso, note que os exemplos fornecidos são baseados na Figura 1. Tendo $c \in V_I \cup E_I$, $x \in \beta(\alpha(c)))$ e f sendo um elemento do esquema ($f \in V_S \cup E_S$), definimos:

- $\omega(c, x)$ retorna o valor do atributo $x \in X$ encontrado no elemento $c \in V_I \cup E_I$. Exemplo: $\omega(a_1, Idade) = 30$.
- $\gamma_I(c)$ mapeia c em uma lista ordenada de valores $\omega(c, x_i)$ chamada tupla, onde $x_i \in \beta(\alpha(c))$, $i = 1, \ldots, |\beta(\alpha(c))|$.

Por exemplo: $\gamma_I(p_2)$ = (p_2,Título_B, 2017, $\{F, G\}$).

- $\gamma_S(c)$ mapeia um elemento de instância c no esquema correspondente. Portanto, $\gamma_S(c) = \alpha(c)(id, x_1, x_2, \ldots, x_{|\beta(\alpha(c))|})$, onde $x_i \in \beta(\alpha(c))$ e Id é um atributo artificialmente criado cujos valores identificam de forma única os componentes associados a esse esquema. Por exemplo: $\gamma_S(a_1) = Autor(Id, Nome, Idade)$.
- $\rho(f)$ mapeia um componente $f \in V_S \cup E_S$ em uma relação R. Para cada elemento c tal que $\alpha(c) = f$, a relação produzida por $\rho(f)$ contém a tupla $\gamma_I(c)$ cujo esquema correspondente é $\gamma_S(c)$. A figura 3(a) mostra um exemplo de relação produzida por $\rho(Artigo)$.
- $\pi_L(R)$: dada uma lista de atributos $L \subset X$ e uma relação R, produz uma projeção relacional de R. Por exemplo, a figura 3(b) mostra a relação produzida por $\pi_{(Id,Ano)}(R)$, onde $R = \rho(Artigo)$.
- $\oplus(v_i, v_j)$: dado um par de elementos instância v_i e v_j de $V_I \cup E_I$, concatena as tuplas correspondentes $\gamma_I(v_i)$ e $\gamma_I(v_j)$, produzindo uma terceira tupla formada por todos os valores das tuplas originais. Por exemplo: $\oplus(a_1, p_1)$ = (a_1, *Amanda*, 30, p_1,Título_A, 2015, $\{A, B, C\}$).
- $(R \bowtie S)_{(f_i,f_j)}$: dado um par de elementos esquema f_i e f_j de $V_S \cup E_S$ e um par de relações R e S cujos esquemas pertencem a 2^X, produz uma relação T onde cada tupla é construída através do processo a seguir. Para cada tupla $r \in R$ e $s \in S$, aplica o operador $\oplus(r, s)$ para concatenar r e s. Então, verifica se r e s são tuplas dos elementos conectados em G_I. Em caso de resposta positiva, a tupla é adicionada a T. Caso contrário, é descartada. A Tabela II contém um exemplo que mostra a relação produzida por $(R \bowtie S)_{(Autor,Artigo)}$, onde R e S são relações resultantes de $\rho(Autor)$ e $\pi_{(Id,Titulo)}(\rho(Artigo))$, respectivamente.

Id	Título	Ano	Palavras-chave
p_1	Título_A	2015	{A, B, C}
p_2	Título_B	2017	{F, G}
p_3	Título_C	2017	{A, B}

(a)

Id	Ano
p_1	2015
p_2	2017
p_3	2017

(b)

Fig. 3. Relações produzidas por: (*a*) $R = \rho(Artigo)$; (*b*) $\pi_{(Id,Ano)}(R)$.

TABLA II
EXEMPLO DE RELAÇÃO PRODUZIDA POR $(R \bowtie S)_{(Autor,Artigo)}$

eAutor.Id	Nome	Idade	Artigo.Id	Título
a_1	Amanda	30	p_1	Título_A
a_1	Amanda	30	p_2	Título_B
a_2	Marcio	32	p_2	Título_B
a_3	Sonia	29	p_3	Título_C

O algoritmo 1 apresenta o pseudo-código desta etapa. São necessárias duas entradas: P e SelAttr, um dicionário que mapeia cada componente de P nos atributos correspondentes selecionados pelo usuário. Consideramos que o i-ésimo tipo em P é acessado por $P[i]$, e que SelAttr[$P[i]$] fornece o conjunto de atributos selecionados para $P[i]$.

Como exemplo, suponha que P e os atributos escolhidos tenham sido $Autor \xrightarrow{Escreveu} Artigo$ e (*Idade*, *Palavras-chave*). A tabela III apresenta a relação produzida pelo Algoritmo 1 quando aplicado à rede da Figura 1.

Algoritmo 1: RecordSetGeneration(P, $SelAttr$)

Input : P e SelAttr, um dicionário para acessar os atributos selecionados em P.
Output: Conjunto de registros de G_I guiados por P.

$R \leftarrow \pi_{\text{SelAttr}[P[1]]}(\rho(P[1]))$;
for $i = 2$ **to** $P.size$ **do**
  $R \leftarrow (R \bowtie \pi_{\text{SelAttr}[P[i]]}(\rho(P[i])))_{(P[i-1],P[i])}$;
return R

TABLA III
EXEMPLO DE RELAÇÃO PRODUZIDA PELO ALGORITMO 1.

Autor.Id	Idade	Artigo.Id	Palavras-chave
a_1	30	p_1	{A, B, C}
a_1	30	p_2	{F, G}
a_2	32	p_2	{F, G}
a_3	29	p_3	{A, B}

A etapa 4, agrupamento de dados, consiste da aplicação de um algoritmo separe os registros de dados em grupos de forma a maximizar a similaridade dos registros alocados em cada grupo e minimizar a similaridade entre registros de grupos distintos.

No exemplo mostrado na Tabela III, somente os atributos *Idade* e *Palavras-chave* foram levados em consideração na etapa 4. A Tabela IV apresenta uma saída possível produzida por esta etapa, onde os registros do exemplo foram divididos em dois *clusters* disjuntos.

É importante destacar que o agrupamento de dados garante que os registros alocados no mesmo grupo compartilhem conteúdo similar. Tal fato é relevante na abordagem proposta, uma vez que cada grupo produzido por esta etapa deve ser processado individualmente pelas etapas subsequentes.

Para cada grupo C_i identificado na etapa anterior, a etapa 5, construção de grafos locais, cria um grafo homogêneo G_i, formado por pares de nós conectados (v_j, v_z) de $V_I \times V_I$ que satisfaçam as seguintes condições: (i) $v_j \neq v_z$; (ii) $\alpha(v_j) = \alpha(v_z) = t$, onde t é o tipo de nó de referência escolhido pelo usuário na etapa de configuração; (iii) Existem dois registros de dados em C_i, r e s, de tal forma que $\omega(r, t.id) = v_j$, $\omega(s, t.id) = v_z$ e $\omega(r, A_k.id) = \omega(s, A_k.id)$, onde A_k é o tipo de nó destino do caminho semântico P definido pelo usuário. Isso significa que v_j e v_k estão indiretamente conectados no grafo original G_I (i. e., existem duas instâncias de caminho em G_I com nó destino comum cujos nós origem são v_j e v_k).

TABLA IV
EXEMPLO DE GRUPOS PRODUZIDOS PELA ETAPA 4 DA *ComDet*.

Grupo #	Autor.Id	Idade	Artigo.Id	Palavras-chave
1	a_1	30	p_1	{A, B, C}
2	a_1	30	p_2	{F, G}
2	a_2	32	p_2	{F, G}
1	a_3	29	p_3	{A, B}

O Algoritmo 2 descreve a etapa 5. Recebe como entrada uma relação R correspondente a um dos grupos formados no passo anterior. Cada linha $r \in R$ armazena informações contextuais associadas a algum vértice $v_r \in G_I$. Este algoritmo usa três funções auxiliares: (a) TargetVertex(r, P) retorna o vértice $v_r \in G_I$ referenciado na linha $r \in R$; (b) SemanticNeighborhood(r, t) retorna o conjunto de vértices $\mathcal{N}_{v_r} \subset G_I$ acessíveis por v_r, todos eles do tipo t; (c) Arestas$(v_r, \mathcal{N}_{v_r})$ retorna um conjunto de arestas (pares não ordenados) ligando v_r a todos os vértices em $\mathcal{N}_{v_r}$.

Algoritmo 2: BuildLocalGraph(t; P; R)

Input : t; P; R (relação com registros de grupo gerado na etapa 4)
Output: $G(V, E)$ grafo local correspondente a R

$G.V \leftarrow \emptyset$;
$G.E \leftarrow \emptyset$;
for $r \in R$ **do**
  $v_r \leftarrow$ TargetVertex(r, P);
  $\mathcal{N}_{v_r} \leftarrow$ SemanticNeighborhood(v_r, R);
  $G.V \leftarrow G.V \cup \mathcal{N}_{v_r} \cup \{v_r\}$;
  $G.E \leftarrow G.E \cup$ Edges$(v_r, \mathcal{N}_{v_r})$;
return G

A figura 4 apresenta os grafos locais construídos por este processo quando aplicados aos grupos identificados na tabela IV. Ambos são homogêneos e contêm arestas que não existem na rede original. Essas arestas foram criadas artificialmente para conectar nós que estão indiretamente ligados no grafo instância de entrada (figura 1). A criação das conexões indiretas foi orientada pelo caminho semântico definido pelo usuário.

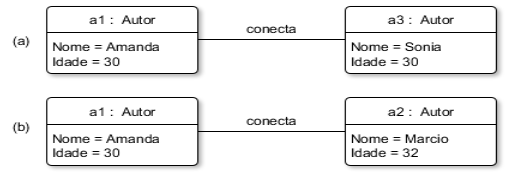

Fig. 4. Grafos induzidos a partir da tabela IV: (a) Grupo #1; (b) Grupo #2.

Enfatizamos dois pontos-chave relacionados ao Algoritmo 2. Primeiro, ele constrói um grafo local de cada grupo de conteúdo coeso identificado na etapa anterior. Assim, garante que os elementos dentro de cada grafo local contenham conteúdo similar/coeso. Em segundo lugar, este algoritmo é capaz não só de restaurar conexões entre nós diretamente ligados no grafo original G_I, mas também é capaz de induzir conexões entre nós indiretamente vinculados em G_I. Este é um dos pontos fortes da ComDet em relação às soluções do estado da arte: permitir a detecção de comunidades não somente entre nós diretamente ligados mas também entre vértices indiretamente conectados. Tal capacidade é uma providência útil em redes heterogêneas onde podem existir vários tipos de nó/aresta e, consequentemente, de conexões indiretas entre vértices.

Finalizando o processamento da ComDet, a etapa 6, detecção de comunidades, é aplicada a cada grafo local G_i induzido pela etapa anterior a fim de detectar subgrafos altamente conectados G_{ij} (comunidades detectadas em G_i). CD_{G_i} indica o conjunto de G_{ij}. Uma vez que os grafos induzidos pela etapa anterior são homogêneos, qualquer algoritmo de detecção de comunidades baseado em topologia pode ser aplicado.

V. Protótipo, Experimentos e Resultados

O protótipo da ComDet[3] foi codificado em Python, utilizando as APIs *Scikit-Learn* [31] e *NetworkX* [32]. Na representação dos grafos foi adotado um formato XML chamado *DyNetML* [33].

Em função de sua popularidade, para o agrupamento de dados e a detecção de comunidade foram utilizados os algoritmos *Affinity Propagation* [12] e Girvan-Newman [2]. O *Affinity Propagation* se baseia na troca de mensagens entre os registros de dados, buscando eleger os representantes de cada grupo. Para tanto, atualiza iterativamente duas matrizes, R e A, onde R indica o quanto cada registro de dados é um representante adequado para os demais registros e A informa o quanto cada registro mostra-se disponível para ser representado pelos demais. A partir de R e A, o *Affinity Propagation* encontra os registros denominados exemplares, i.e. representantes de grupos, definindo, para cada um deles, um grupo. Em seguida, atribui a cada grupo, os registros representados pelo exemplar do grupo. O Girvan-Newman, por sua vez, é um método hierárquico que vai gradativamente removendo as arestas do grafo que apresentam maior centralidade por intermediação, i.e., maior número de caminhos mais curtos entre pares de nós que contenham essas arestas. Ao passo que o grafo é subdividido, a estrutura de suas comunidades são expostas.

Três redes heterogêneas com atributos foram utilizadas nos experimentos (vide esquemas na figura 5). A *Militarized Interstate Dispute* (MID) contém informações sobre conflitos entre vários países que ocorreram de 1816 a 2010 [8]. A principal motivação para sua escolha foi o fato dela possuir natureza militar. A rede ArXiv é um repositório de artigos científicos de várias áreas de conhecimento e que podem ser acessados online[4]. Tal repositório inclui informações sobre coautoria e palavras-chave de artigos. Nos experimentos, consideramos artigos publicados de 1994 a 1997 em cinco seções de Física popularmente usadas em estudos sobre redes complexas. O Enron Email Dataset (Enron) contém e-mails enviados pelos funcionários da *Enron* e adquiridos durante a investigação após o colapso da empresa [9]. As informações contextuais disponíveis estão no formato textual, armazenadas em arestas, tendo sido este o principal motivo para escolha deste conjunto. A tabela V fornece informações estatísticas sobre essas redes.

O processo de avaliação adotado nos experimentos está ilustrado na Figura 6. Teve como objetivo comparar os resultados da ComDet com os obtidos pelos algoritmos Girvan-Newman [2] e LouvainC [10]. Tais algoritmos foram escolhidos pois, diferentemente dos propostos em [29], [27], [3], não demandam configuração pelo usuário, assegurando assim imparcialidade em sua aplicação.

[3]O código-fonte está disponível em https://goo.gl/Rfuwrf.
[4]http://export.arxiv.org/

TABLA V
ESTATÍSTICAS SOBRE AS REDES UTILIZADAS NOS EXPERIMENTOS

Rede	No. Nós	No. Arestas	Coeficiente de agrupamento
MID	2.779	16.544	0,644
ArXiv	902	8.422	0,521
Enron	158	7.139	0,493

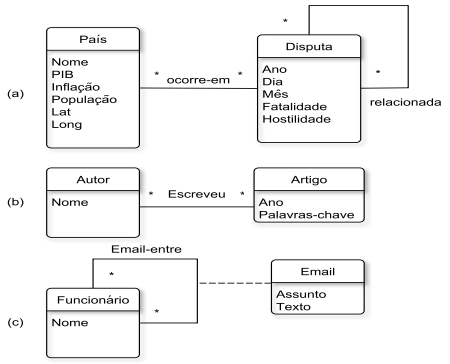

Fig. 5. Esquemas das redes: (a) MID; (b) ArXiv; (c) Enron.

Diferentemente da ComDet, os algoritmos Girvan-Newman e LouvainC exigem grafos homogêneos como entradas. Assim, dada uma rede heterogênea representada por seu esquema e instância, o primeiro estágio do processo de avaliação consiste em converter essa rede em um grafo homogêneo. Com base no esquema da rede, o analista deve selecionar um tipo de nó de referência (t') e um caminho semântico (P') que contenha t'. Para comparar as comunidades detectadas pelos algoritmos Girvan-Newman e LouvainC com as identificadas pela ComDet, as escolhas de t' e P' devem ser as mesmas feitas para a ComDet (ou seja, $t' = t$ e $P' = P$). Assim, dados t' e P', o primeiro estágio do processo de avaliação cria uma aresta artificial entre cada par de nós u e v ($u \neq v$) do tipo t' que estão conectados por alguma instância de caminho de P'. Em seguida, todas as arestas reais são removidas da rede original, assim como todos os vértices cujo tipo não seja t', resultando em um grafo homogêneo onde somente as arestas artificiais e os vértices do tipo t' são preservados.

No segundo estágio do processo de avaliação, a rede de entrada é submetida à ComDet e a versão homogênea da rede induzida pelo estágio anterior é submetida ao Girvan-Newman e ao LouvainC separadamente. Os algoritmos Girvan-Newman e LouvainC geram um conjunto de comunidades cada (chamado GN_G e LC_G) e a ComDet, por sua vez, gera n conjuntos de comunidades (chamados CD_{G_1}, CD_{G_2}, ..., CD_{G_n}).

Os resultados produzidos pelo estágio anterior são avaliados quantitativamente no terceiro estágio do processo. Para tanto, são utilizadas três métricas de avaliação popularmente empregadas em experimentos envolvendo detecção de comunidades: *modularidade* [7], *densidade* [28] e *informação mútua normalizada* (*NMI*) [6]. A modularidade e a densidade são aplicadas separadamente a cada conjunto de comunidades detectadas pelos algoritmos Girvan-Newman e LouvainC e pela abordagem ComDet. Altos valores dessas métricas indicam boas partições sob o ponto de vista topológico, ou seja, comunidades com elevado número de arestas em grupos de nós altamente interligados. A fim de verificar a influência das informações contextuais no resultado do processo de detecção de comunidades, são calculados os valores de NMI entre cada conjunto de comunidades detectadas pela ComDet e os conjuntos de comunidades detectadas pelos algoritmos Girvan-Newman e LouvainC. Baixos valores de NMI indicam que a estratégia da ComDet em combinar dados contextuais com a informação da topologia da rede levou a comunidades estruturalmente diferentes das detectadas pelos algoritmos exclusivamente topológicos Girvan-Newman e LouvainC.

O quarto e último estágio do processo de avaliação compreende a análise qualitativa das comunidades detectadas pela ComDet. Deve ser realizada por um especialista do domínio que busque por padrões que representem algum tipo de coesão entre os dados contextuais disponíveis em cada comunidade e que possam ser úteis sob o ponto de vista da aplicação.

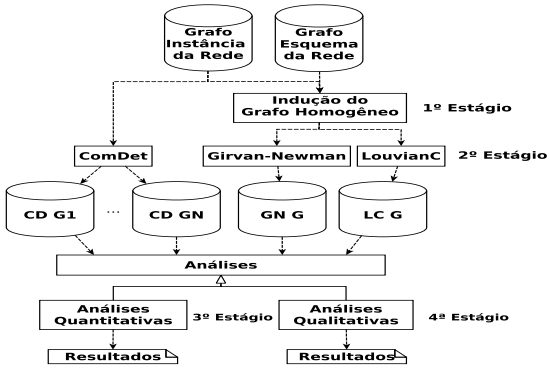

Fig. 6. Visão Geral do Processo de Avaliação e Comparação dos Resultados.

A Tabela VI contém a configuração do experimento realizado em cada rede. Embora outras configurações pudessem ser avaliadas, optamos por priorizar aquelas cujas comunidades formadas pudessem ter um maior sentido prático em termos de aplicação. Por exemplo, é mais natural perceber a utilidade de comunidades formadas por países, autores ou funcionários do que comunidades onde os membros sejam disputas, artigos ou e-mails. A figura 7 apresenta os resultados quantitativos obtidos pela ComDet (CD_G), pelo Girvan-Newman (GN_G) e pelo LouvainC (LC_G). No caso da ComDet, os resultados apresentados indicam as médias ± desvio-padrão calculadas a partir dos valores das respectivas métricas de avaliação apurados nos conjuntos de comunidades CD_{G_i}.

Na etapa de pré-processamento da rede MID, a única operação realizada foi a normalização linear dos atributos contextuais. Em termos quantitativos, a ComDet identificou seis conjuntos de comunidades, apresentando uma média de modularidade superior às dos conjuntos identificados pelos outros algoritmos. Na média, as comunidades detectadas pela ComDet apresentaram uma densidade superior às das identificadas pelo algoritmo Girvan-Newman e bem próxima das obtidas pelo algoritmo LouvainC. Tais fatos são evidências de que a ComDet foi capaz de identificar melhores partições do que os dois algoritmos de base estritamente topológica. Os valores médios de NMI próximos de *0,7* e *0,45* mostram

TABLA VI
CONFIGURAÇÃO DOS EXPERIMENTOS

Rede	Nó Ref. t	Caminho Semântico	Atributos Contextuais	Objetivo - Detectar comunidades de:
MID	País	$País \xrightarrow{Ocorrem-em} Disputa$	Latitude, longitude, nº de fatalidades	Países geograficamente próximos envolvidos em conflitos comuns com números aproximados de mortes.
ArXiv	Autor	$Autor \xrightarrow{Escreveu} Artigo$	Palavras-chave dos artigos	Autores com co-autoria em artigos com conteúdos similares.
Enron	Funcionário	$Funcionário \xrightarrow{Email-entre} Funcionário$	Assunto e texto dos e-mails	Funcionários que trocaram e-mails com conteúdos (interesses) similares.

que as comunidades detectadas pela ComDet foram bem diferentes das detectadas pelo Girvan-Newman e pelo LouvainC. Isso significa que os atributos contextuais considerados pela abordagem proposta influenciaram o processo de detecção e levaram a comunidades estruturalmente distintas das detectadas pelos algoritmos exclusivamente baseados em topologia. Do ponto de vista qualitativo, os algoritmos Girvan-Newman e LouvainC identificaram comunidades disjuntas e que incluíram países de continentes diferentes. Por exemplo, ambos colocaram os EUA em comunidades com a Turquia e a Grécia, e com o Irã e o Iraque. Por outro lado, as comunidades detectadas pela ComDet dividiram os países de acordo com seu continente (ou seja, países geograficamente próximos). Consequentemente, nos resultados obtidos com a abordagem proposta, os EUA só apareceram em comunidades de países do continente americano. Claramente, esses resultados foram influenciados pelos atributos contextuais de latitude e longitude usados pela ComDet no experimento.

Na etapa de pré-processamento da rede ArXiv, foram construídas: primeiro, uma matriz TF-IDF associando artigos com palavras-chave e, então, uma matriz de similaridade entre os documentos (artigos). Esta última foi a entrada para a etapa de agrupamento de dados. Nos resultados, a ComDet identificou onze conjuntos de comunidades. Sob o ponto de vista quantitativo, na média, os resultados produzidos pela ComDet ficaram bem próximos aos produzidos pelos outros dois algoritmos. No entanto, cabe chamar a atenção para um dos conjuntos produzidos pela ComDet (CD_{G_9}). Ele foi o que apresentou menor valor de NMI (*0,92*) e maiores valores de densidade (*0,88*) e modularidade (*0,96*). Ou seja, retratou uma boa partição do ponto de vista topológico, além de uma maior dissimilaridade em relação aos conjuntos identificados pelos algoritmos baseados exclusivamente em topologia. Ao analisar esse mesmo conjunto de comunidades sob uma perspectiva qualitativa, é possível observar que a ComDet foi capaz de identificar comunidades de autores que, além de possuírem interesses (palavras-chave) comuns, também publicaram artigos em co-autoria. Como artigos de áreas comuns normalmente compartilham palavras-chave comuns, o processamento da ComDet foi certamente influenciado pelo uso das palavras-chave como atributos contextuais.

No pré-processamento da rede Enron, o assunto e o corpo (texto) de cada e-mail foram concatenados e submetidos às operações de *tokenização* (*tokenization*), remoção de palavras de parada (*stopword removal*) e *lematização* (*stemming*). Posteriormente, as matrizes TF-IDF e de similaridade entre documentos (e-mails) foram construídas e usadas pela ComDet. Em termos quantitativos, a ComDet detectou quinze conjuntos de comunidades, com densidade e modularidade médias próximas das apresentadas pelos conjuntos de comunidades identificados pelo Girvan-Newman e pelo LouvainC. Entretanto, baixos valores de NMI mostram que as comunidades detectadas pela ComDet foram diferentes das detectadas pelos outros algoritmos. Da mesma forma que nos demais experimentos, os atributos contextuais considerados pela abordagem proposta influenciaram o processo e levaram a comunidades estruturalmente distintas das detectadas pelos algoritmos exclusivamente baseados em topologia. Isso pode ser ilustrado, sob o ponto de vista qualitativo, pelas nuvens de palavras da figura 8. Cada uma delas foi extraída de uma comunidade de um conjunto identificado pela ComDet (CD_{G_3}). Cada comunidade apresentou uma temática (interesse) predominante no grupo, porém distinta das temáticas dos demais grupos. Palavras como *comission*, *market*, *price*, *rate*, entre outras presentes na figura 8(a) são comuns e frequentes nas discussões de negociações e finanças. A figura 8(b) mostra termos usuais entre pessoas do setor de Tecnologia da Informação (TI) como *data*, *system*, *group*, *offsite*, *video*, *information*, *identity*, entre outras. Da mesma forma, a figura 8(c) indica termos frequentes típicos de setores de logística onde as preocupações com datas e prazos são críticas. Por exemplo: *Nov*, *Jan*, *Feb*, *Oct*, *Dec* (menção aos meses do ano), *confirmations*, *management*, *state*, entre outras.

(a)

Conjunto	Densidade	Modularidade	NMI GN_G	NMI LC_G
CD_G	0,42 ± 0,17	0,26 ± 0,17	0,70 ± 0,11	0,45 ± 0,09
GN_G	0,19	0,11	1,00	—
LC_G	0,45	0,15	—	1,00

(b)

Conjunto	Densidade	Modularidade	NMI GN_G	NMI LC_G
CD_G	0,80 ± 0,07	0,85 ± 0,08	0,93 ± 0,01	0,93 ± 0,01
GN_G	0,82	0,90	1,00	—
LC_G	0,83	0,91	—	1,00

(c)

Conjunto	Densidade	Modularidade	NMI GN_G	NMI LC_G
CD_G	0,37 ± 0,08	0,59 ± 0,56	0,53 ± 0,06	0,63 ± 0,07
GN_G	0,47	0,32	1,00	—
LC_G	0,45	0,64	—	1,00

Fig. 7. Resumo dos resultados quantitativos dos experimentos nas redes: (a) MID; (b) ArXiv; e (c) Enron

VI. CONCLUSÃO

Neste artigo, propusemos a ComDet, uma abordagem que leva em consideração tanto informações topológicas e quanto contextuais para detectar comunidades em redes heterogêneas. A ComDet usa o agrupamento de dados como etapa de pré-processamento para o processo de detecção de comunidades, a fim de identificar nós semelhantes altamente conectados e organizá-los em comunidades coesas e possivelmente

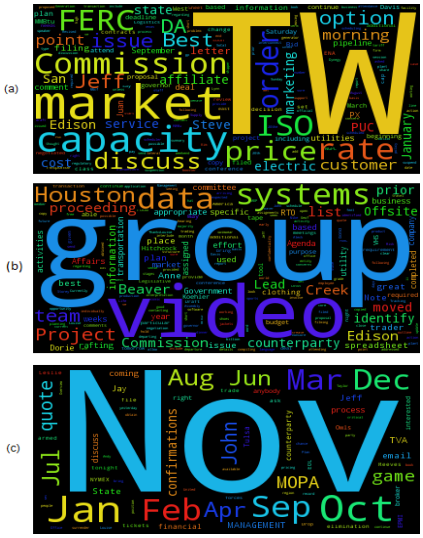

Fig. 8. Nuvens de palavras extraídas de comunidades detectadas pela ComDet na Rede Enron: (a) negociações e discussões financeiras, (b) questões relacionadas à área de TI e (c) logística, gerenciamento e prazos.

sobrepostas. Foram executados experimentos em três redes heterogêneas e os resultados mostraram que, diferente de algoritmos baseados em topologia, a ComDet leva a boas partições com comunidades de conteúdo coeso. Como trabalho futuro está a extensão da ComDet a fim de oferecer recursos para apoiar a escolha de configurações tais como a seleção de atributos contextuais a serem empregados na fase de agrupamento e de caminhos semânticos que possam levar a melhores partições das redes em comunidades. Outro trabalho futuro envolve a aplicação da ComDet para prever se os nós não conectados devem se ligar no futuro (problema de análise de redes complexas conhecido como previsão de links). Experimentos com outras redes heterogêneas com atributos visando uma análise de significância estatística dos resultados também estão planejados.

REFERÊNCIAS

[1] Fortunato, S. and Darko, H. (2016). Community detection in networks: A user guide. *CoRR*, volume abs/1608.00163.

[2] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *PNAS*, 99(12), pages 7821–7826.

[3] Guedes, G. P., et al. (2016). Discovering top-k non-redundant clusterings in attributed graphs. *Neurocomputing*, 210:45 – 54.

[4] Günnemann, et al. (2010). Subspace Clustering Meets Dense Subgraph Mining: A Synthesis of Two Paradigms. In *Data Mining (ICDM), 2010 IEEE 10th Intl. Conf. on*, pages 845–850.

[5] Liu, X., Liu, W., Murata, T., and Wakita, K. (2014). A Framework for Community Detection in Heterogeneous Multi-Relational Networks. *Advances in Complex Systems*, volume 17.

[6] McDaid, A. F., Greene, D., and Hurley, N. J. (2011). Normalized Mutual Information to evaluate overlapping community finding algorithms. *CoRR*, abs/1110.2515.

[7] Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2), pages 026113.

[8] Palmer, G., Dórazio, V., Kenwick, M., and Lane, M. (2015). The MID4 dataset, 2002–2010: Procedures, coding rules and description. *Conflict Management and Peace Science*, 32(2), pages 222–242.

[9] Sun, J., Faloutsos, C., Papadimitriou, S., and Yu, P. S. (2007). GraphScope: Parameter-free Mining of Large Time-evolving Graphs. In *Proceedings of the 13th ACM SIGKDD*, number 1, pages 687–696, NY.

[10] Li, X. and Chen, H. (2009). Recommendation as link prediction: a graph kernel-based machine learning approach. In *Proceedings of the 2009 Joint Intl. Conf. on Digital Libraries, Austin, TX*, pages 213–216.

[11] Chen, Y.-L., et al. (2014) Community detection based on social interactions in a social network. *JASIST*, 65, 539–550.

[12] Guo, K. and Zhang, Q. (2015) Detecting communities in social networks by local affinity propagation with grey relational analysis. *Grey Systems: T&A*, 5, 31–40.

[13] Yizhou, Sun, et al. (2012). Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In *KDD*, USA, 7, pp. 1348–1356.

[14] Melamed, D. (2014) Community structures in bipartite networks: A dual-projection approach. *PLoS ONE*, 9, e97823.

[15] Meng, Q., Tafavogh, S., and Kennedy, P. (2014) Community detection on heterogeneous networks by multiple semantic-path clustering. *Comp. Aspects of Social Networks (CASoN), Intl. Conf. on*, pp. 7–12.

[16] Yang, T., Jin, R., Chi, Y., and Zhu, S. (2009) Combining link and content for community detection: a discriminative approach. *Proc. of the 15th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, USA, 11, pp. 927–936.

[17] Dang, T. A. and Viennet, E. (2012) Community Detection based on Structural and Attribute Similarities. *ICDS*, pages 7–14.

[18] Akoglu, L., et al. (2012) PICS: Parameter-free Identification of Cohesive Subgroups in Large Attributed Graphs. *Proc. of the 2012 SIAM Intl. Conf. on Data Mining*. SIAM, USA.

[19] Porter, M., Onnela, J., and Mucha, P. (2009) Communities in Networks. *Notices of the American Mathematical Society*, 56, 1082–1097.

[20] Aggarwal, C. C. (2011) An Introduction to Social Network Data Analytics. In *Social Network Data Analytics*, pp. 1–15. Springer, GEN.

[21] Elmasri, R. and Navathe, S. (2013) Fundamentals of Database Systems. In *Always learning*, pp. 1–15. Pearson.

[22] Ferrara, E. (2011) A Large-Scale Community Structure Analysis In Facebook. *CoRR*, abs/1106.2503.

[23] Zhao, Z., et al. (2012) Topic oriented community detection through social objects and link analysis in social networks. *Knowl.-Based Syst.*, 26, 164–173.

[24] Ruan, Y., Fuhry, D., and Parthasarathy, S. (2013) Efficient community detection in large networks using content and links. *Proc. of the 22nd intl conf. on World Wide Web*, USA, 3, pp. 1089–1098.

[25] Yang, J., McAuley, J. J., and Leskovec, J. (2013) Community Detection in Networks with Node Attributes. In *ICDM*, USA, 8, pp. 1151–1156.

[26] Newman, M. E. J. and Clauset, A. (2016) Structure and inference in annotated networks. *Nature Communications*, 7, 11863 EP –.

[27] Sachan, M., Contractor, D., Faruquie, T. A., and Subramaniam, L. V. (2012) Using content and interactions for discovering communities in social networks. In *WWW*, USA, 8, pp. 331–340. ACM.

[28] Yang, J. and Leskovec, J. (2013). Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proc. of the Sixth ACM Intl. Conf. on Web Search and Data Mining*, 1, pages 587–596, New York, USA.

[29] Zhou, Y., Cheng, H., and Yu, J. X. (2009). Graph clustering based on structural/attribute similarities. *PVLDB*, 2(1):718–729.

[30] Safira, Nayara. (2013). Identificação Litológica Com Affinity Propagation. *Universidade Federal do Pará, Mestrado em Sistemas e Computação*, 134.

[31] Pedregosa, F., Varoquaux, G., Gramfort, A., andB. Thirion, V. M., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[32] Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, pages 11–15.

[33] M. Tsvetovat, J. Reminga, and K. M. Carley. (2004) DyNetML: Interchange Format for Rich Social Network Data. *SSRN Electronic Journa*, page 25.

M. V. Dias nasceu no Rio de Janeiro, RJ, BRA em 1984. Ele recebeu o bacharelado em Engenharia de Computação pela Universidade Estadual do Rio de Janeiro (UERJ), Rio de Janeiro, em 2013 e M. Sc. no Instituto Militar de Engenharia (IME) em 2016. Trabalha com engenharia de software desde 2001.

P. A. Braz Paulo Alves Braz nasceu no Rio de Janeiro, RJ, BRA em 1981. Ele recebeu o bacharelado em Ciência da Computação da Universidade Federal do Rio de Janeiro (UFRJ), no Rio de Janeiro, em 2009. De 2009 a 2016, foi gerente sênior de dados de perfuração na Halliburton. Atualmente cursa o mestrado em Sistemas e Computação no Instituto Militar de Engenharia (IME).

E. Bezerra Eduardo Bezerra recebeu seu bacharelado em Informática pela Universidade Federal do Rio de Janeiro (1992-1995). Ele tem ambos M.Sc. e D.Sc. em Computação e Engenharia de Sistemas pela COPPE/UFRJ. Trabalha como professor no Centro Federal de Educação Tecnológica do Rio de Janeiro (CEFET/RJ). Seus interesses de pesquisa estão relacionados a aplicações práticas de Aprendizado de Máquina.

R. R. Goldschmidt recebeu seu bacharelado em Matemática pela Universidade Federal Fluminense (1990), o M.Sc. em Sistemas de Computação pelo Instituto Militar de Engenharia (1992) e o D.Sc. em Engenharia Elétrica pela Pontifica Universidade Católica do Rio de Janeiro (2004). Atualmente trabalha como professor associado ao Instituto Militar de Engenharia e seus interesses de pesquisa incluem Ciência de Dados e Inteligência Artificial.