Detecção de comunidades baseada em informações contextuais em redes homogêneas com atributos

M. V. Dias[*a], P. A. Braza, E. B. Silva[b] e R. R. Goldschmidt[a]
[a] *Instituto Militar de Engenharia, Rio de Janeiro, Brasil.*
[b] *Centro Federal de Educação Tecnológica,*
Av. Maracanã, 229 - Maracanã, Rio de Janeiro - RJ, 20271-110
Rio de Janeiro, Brasil

RESUMO: *A detecção de comunidades é uma importante tarefa de análise de redes complexas que vem sendo estudada pela academia e pela indústria nos últimos anos. Trata-se de um problema de otimização que tenta identificar grupos (comunidades) de elementos altamente interligados em redes de grande porte. A maioria dos algoritmos projetados até o momento para resolver este problema concentra-se exclusivamente na topologia da rede de entrada, ignorando qualquer informação existente sobre o contexto da aplicação. Neste artigo, propõe-se uma abordagem de detecção de comunidades que leva em consideração tanto informações topológicas quanto informações contextuais da rede. Esta abordagem introduz o agrupamento de dados como uma etapa prévia ao processo de detecção de comunidades, a fim de identificar comunidades estruturalmente densas, coesas e possivelmente sobrepostas. O artigo apresenta resultados quantitativos e qualitativos obtidos em três redes de natureza distinta, ilustrando o potencial de aplicação da abordagem proposta nos cenários militar, civil e científico. Os experimentos realizados mostram que a combinação de informações contextuais e topológicas das redes pode levar a partições interessantes formando comunidades de conteúdo coeso e útil.*

PALAVRAS-CHAVE: *Detecção de Comunidades, Agrupamento de Dados, Análise de Redes Complexas, Redes Homogêneas, Grafos com Atributos.*

ABSTRACT: ABSTRACT: *Community detection is an important network analysis task that has been studied by academy and industry.It is an optimization problem that tries to identify groups (communities) of highly interconnected nodes in a network. Most algorithms designed so far to solve this problem concentrate exclusively on the topological aspects of the input network, ignoring any existing information about the context of application. This article proposes a community detection approach that takes both topological and contextual information into consideration.This approach introduces data clustering as a pre-processing step for the community detection process in order to identify structurally dense, cohesive and possibly overlapping communities. The article also presents quantitative and qualitative results obtained in three networks of different nature illustrating the potential of application of the proposed approach in military, civil and scientific scenarios. Experiments show that the combination of contextual and topological information may lead to interesting partitions with cohesive and useful content communities.*

KEYWORDS: *Community Detection, Data Clustering, Complex Network Analysis, Homogeneous Network, Attributed Graphs.*

1. Introdução

Nos últimos anos, academia e indústria têm dedicado grande atenção à análise de redes complexas [1]. Uma rede complexa é um multigrafo[1] altamente interconectado, possivelmente contendo atributos[2], onde um vértice (nó) representa um item da rede (por exemplo, pessoa, página da Web, produto, filme, foto, artigo, etc.) e uma aresta representa algum tipo de associação entre os itens correspondentes (por exemplo, amizade ou comunicação entre duas pessoas) [2]. O problema de detectar grupos de nós densamente interconectados é uma importante tarefa de análise de redes complexas conhecida como detecção de comunidades [3].

Inicialmente usada para identificar grupos de pessoas em redes sociais, a detecção de comunidades tem sido aplicada a uma ampla gama de áreas, desde então [4]. Por exemplo: (a) na Ciência da Web, para detectar clusters (grupos) de web sites interconectados [5], [6]; (b) em sistemas de recomendação, para aplicações de recuperação de informação e comércio eletrônico [7], [8], [9]; (c) em Biblioteconomia e Ciência da Informação, para a identificação e remoção de dados duplicados [10], [11]; (d) em Bioinformática, para identificar interações entre proteínas [12], [13]; (e) em aplicações de segurança como a descoberta de grupos virtuais de terroristas e criminosos ocultos em redes sociais [14].

Em essência, a detecção de comunidades é um problema de otimização que tenta organizar os elementos em grupos (comunidades), de forma a maximizar o número de arestas em um mesmo grupo e minimizar a quantidade de arestas em grupos distintos [4]. No entanto, em muitas aplicações reais, a estrutura topológica do grafo e os dados contextuais podem ser importantes para encontrar comunidades estruturalmente densas e coesas [15] [16].

Para ilustrar essa hipótese, considere o exemplo de uma rede fictícia de coautoria de artigos científicos descrita na figura 1. Cada vértice representa um autor e cada aresta representa um artigo escrito em parceria pelos autores por ela conectados. Toda aresta tem um atributo com informações contextuais, ou seja, o conjunto de palavras-chave usadas para rotular o artigo. Do ponto de vista exclusivamente topológico, a detecção de comunidades possivelmente encontraria duas comunidades: uma com os vértices A, B, C e D; e outra com os vértices E, F e G. Ambas as comunidades seriam estruturalmente densas. Porém a primeira conteria autores de artigos envolvendo palavras-chave distintas, o que poderia representar interesses distintos. Por outro lado, se o processo de detecção da comunidade pudesse também levar em conta uma perspectiva contextual, poderia gerar três comunidades sobrepostas[3], tais como {A, B, C}, {B, C, D} e {E, F, G}, que, além de estruturalmente densas, apresentariam uma coesão de conteúdo. Segundo esta divisão, mais refinada que a primeira, cada comunidade agruparia autores

[1] Um multigrafo é um grafo onde dois nós podem ser conectados por múltiplas arestas.
[2] Um grafo com atributos se caracteriza por apresentar atributos em nós e/ou arestas. Em geral, atributos contêm informações sobre o contexto da aplicação. Por exemplo, palavras-chave e data de publicação de um artigo, gênero e renda de uma pessoa, preço de um produto, coordenadas geográficas, etc.
[3] Comunidades sobrepostas são aquelas que têm pelo menos um nó comum. Comunidades não sobrepostas são chamadas de comunidades disjuntas [5].

que escreveram artigos com palavras-chave comuns caracterizando, assim, uma maior coerência entre os assuntos de interesse dos elementos de cada grupo.

Diante do exposto, este artigo tem como objetivo propor uma abordagem de detecção de comunidades que leva em consideração tanto informações topológicas quanto informações contextuais. Chamada de ComDet, a abordagem proposta introduz o agrupamento de dados[4] como uma etapa prévia ao processo de detecção das comunidades a fim de identificar 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, avaliações qualitativas e quantitativas com três redes reais mostram que a combinação de dados contextuais e topológicos pode levar a comunidades densas com conteúdos semelhantes dentro de cada uma. As redes escolhidas ilustram o potencial de aplicação da abordagem proposta nos cenários militar, civil e científico.

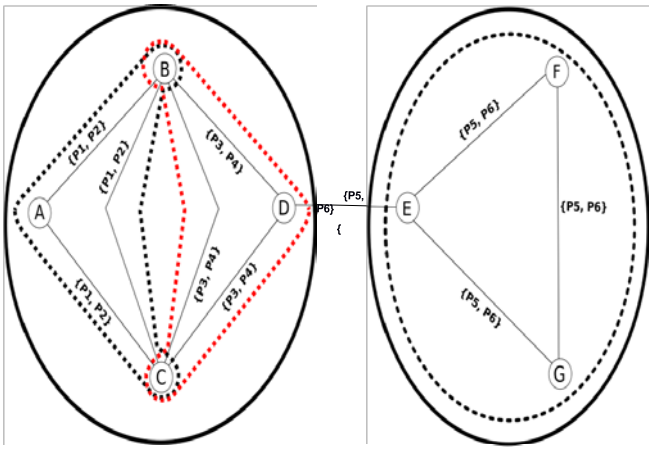

Fig. 1 – Exemplo de comunidades detectadas em dois cenários. As linhas contínuas em negrito representam comunidades identificadas sob um ponto de vista exclusivamente topológico. As linhas tracejadas indicam comunidades detectadas diante da combinação dos pontos de vista topológico e contextual

Este texto contém outras cinco seções. A seção 2 resume os conceitos básicos em detecção de comunidades. Na Seção 3 são apresentados trabalhos relacionados à detecção de comunidades próximos à abordagem proposta neste artigo. O detalhamento da abordagem proposta encontra-se na seção 4. A Seção 5 expõe e comenta os resultados obtidos nos experimentos. Conclusões e trabalhos futuros encontram-se indicados na Seção 4.

2. Fundamentação teórica

Esta seção discute os conceitos relevantes para a compreensão deste trabalho, a saber, a detecção da comunidade e o agrupamento de dados. Também estão descritas algumas métricas de avaliação usadas nos experimentos.

2.1 Detecção de comunidades

Uma comunidade é definida como um conjunto de subgrupos coesos de elementos de um conjunto de dados [4]. A detecção de comunidades em redes complexas é a tarefa de identificar grupos de vértices, na qual os vértices pertencentes a um grupo específico interagem uns com os outros com mais frequência do que com vértices que estão fora do grupo [17]. A maioria dos algoritmos clássicos na área de detecção de comunidades não usa informações do domínio de aplicação [18], [3], [19], [20], [21], que são importantes na identificação efetiva de grupos. Entre os principais algoritmos clássicos para detecção de comunidades estão o Girvan-Newman e o LouvainC.

O algoritmo de Girvan-Newman é um método hierárquico para detecção de comunidades em redes homogêneas[4] [22]. Esse método remove arestas gradativamente do grafo original G. A ordem de remoção dessas arestas segue o critério de centralidade por intermediação, β. Para qualquer aresta $e \in E$, considere que $\beta(e)$ retorna o número de caminhos mais curtos entre o par de nós v e u $\in V$ que contêm e. Ao passo que G é subdividido, a estrutura de suas comunidades são expostas. As etapas deste algoritmo são resumidas pelo seguinte algoritmo:

1) $E \leftarrow G.E$
2) Para cada e ∈ E, calcule β(e);
3) $e_{max} \leftarrow \text{argmax}_{e \in E}\ \beta(e)$;
4) $E \leftarrow E - \{e_{max}\}$
5) Para cada e ∈ E afetado pela remoção do e_{max}, recalcule β (e);
6) Repita os passos 3 e 4 até E = ∅.

A execução do algoritmo acima produz gradualmente um *dendrograma*, uma árvore na qual a raiz representa todos os nós em G.V, e cada folha é um nó de G. As comunidades finais podem ser extraídas escolhendo um limiar de corte em alguma profundidade do *dendrograma*.

O algoritmo LouvainC [21] é um algoritmo de detecção de comunidade que lida com redes heterogêneas[6]. Este algoritmo usa o conceito chamado de modularidade composta a fim de avaliar a partição de uma rede heterogênea em comunidades. Dada uma rede heterogênea G, o LouvainC assume que $G = G^{[1]} \cup G^{[2]} \cup \ldots \cup G^{[s]}$, onde cada $G^{[y]}$ é uma sub-rede que consiste em arestas do mesmo tipo. Na primeira etapa, o algoritmo detecta comunidades em cada sub-rede separadamente. Assim, cada nó pode ser atribuído a diferentes comunidades de diferentes sub-redes. No segundo passo, o LouvainC combina as partições das sub-redes e obtém algumas constantes. Cada constante é associada a um conjunto de vértices que devem ficar juntos se esses forem atribuídos a uma mesma comunidade em cada partição. No terceiro passo, o algoritmo usa as constantes derivadas anteriormente para construir uma nova rede onde cada novo nó representa um grupo de nós que devem ser agrupados em conjunto. Dado dois nós u e v da nova rede, cada nova aresta representa o conjunto de atestas do gráfico original que conecta os nós representados por u e v. Então, o LouvainC otimiza a modularidade composta, que é descrita pela equação 1 ($Q^{[y]}$ representa a modularidade em $G^{[y]}$, $m^{[y]}$ é o número de arestas em $G^{[y]}$, m é o número total de arestas e L é a partição). O LouvainC não leva em consideração dados contextuais durante o processo.

$$Q(L) = \sum_{y=1}^{S} \frac{m^{[y]}}{m} Q^{[y]}(L) \qquad (1)$$

[4] O agrupamento de dados é uma tarefa de mineração de dados que agrupa registros de dados de acordo com sua similaridade de conteúdo [3], [15].
[5]Redes que contêm apenas um tipo de nó e um tipo de aresta.
[6] Redes que contêm nós de vários tipos de nós e/ou conexões.

2.2 Algoritmos de agrupamento

O agrupamento de dados pode ser definido como um problema de otimização, no qual o objetivo é maximizar a similaridade intra-grupo e minimizar a similaridade intergrupos [22], [23], de acordo com um critério predeterminado. Abaixo encontra-se sumarizado o Affinity Propagation, o algoritmo de agrupamento de dados usado nos experimentos deste artigo.

O Affinity Propagation (AP) [24] é um algoritmo de agrupamento que se baseia na ideia de passar mensagens entre os pontos de dados. Considere um conjunto de elementos de um conjunto de dados $\{x_1, x_2, ..., x_n\}$. Esse algoritmo toma como entrada uma matriz cujo elemento a(i, j) corresponde à similaridade entre os pontos x_i e x_j. A seguir, duas matrizes A e R são iterativamente atualizadas até a convergência. Na matriz de responsabilidades R, o elemento r(i, k) indica o quanto o elemento x_k é um representante adequado para o elemento xi na matriz de disponibilidade A, o elemento s(i, k) indica o quão apropriado é o elemento x_k pode ser escolhido como representante do elemento x_i. A atualização dos elementos dessas matrizes é feita por meio de equações 2, 3 e 4, apresentadas a seguir.

$$r(i,k) \leftarrow \mathrm{s(i,k)} - \max_{k' \neq k}\{a\left(i,k'\right), s(i,k')\} \tag{2}$$

$$a(i,k) \leftarrow \min\left\{0, r(k,k) + \sum_{i' \notin \{i,k\}} \max_{k \neq k}\{0, r(i',k)\}\right\} \tag{3}$$

$$a(k,k) \leftarrow \sum_{i' \neq k} \max\{0, r(i',k)\} \tag{4}$$

Uma vez definidas as matrizes A e R, o algoritmo encontra os denominados exemplares, pontos para os quais o somatório da disponibilidade com a responsabilidade é um valor positivo. Esses exemplares definem os grupos identificados pelo AP: cada grupo é composto pelo seu exemplar correspondente e pelos pontos representados por este exemplar.

O AP difere dos algoritmos clássicos de agrupamento como o k-means e o k-medoids, uma vez que não requer que a quantidade de grupos a ser formada seja fornecida como entrada para a tarefa de agrupamento.

2.3 Métricas de avaliação

Neste artigo, foram utilizadas três métricas de avaliação nos experimentos, a saber, modularidade, densidade e informação mútua normalizada (NMI). Abaixo segue uma visão geral destas métricas.

Modularidade (M) mede a qualidade de uma determinada partição de uma rede [25], [26]. Especificamente, é definida como a diferença entre dois valores. O primeiro corresponde à fração de arestas da rede que conecta vértices na mesma comunidade. O segundo é o valor esperado da fração de arestas em uma rede que conecta vértices dentro da mesma comunidade. Se o número de arestas dentro da comunidade não é melhor do que aleatório, então M = 0. Por outro lado, os valores de M aproximando-se de 1 indicam forte estrutura comunitária.

Densidade (ou densidade interna) pode ser definida como o número de arestas, n_a, em uma rede complexa R dividida pelo o total do número de arestas possíveis, n_T. O número total de arestas possíveis em uma rede R é obtido por $n_T = (n \times (n-1))/2$, onde n é o número de vértices em R [27]. n_T está dividido por dois para evitar contar uma aresta duas vezes. Com isso, densidade de uma rede R pode ser definida pela equação 5.

$$\mathrm{Densidade(R)} = \frac{n_a}{n_T} \tag{5}$$

A modularidade expressa como os nós das comunidades no conjunto de dados estão conectados. Por outro lado, a densidade indica a concentração das arestas nas comunidades detectadas. Os altos valores dessas métricas indicam que as comunidades detectadas concentram alto número de arestas em grupos de nós altamente interligados, ou seja, uma boa partição.

Seguindo a metodologia apresentada em [28], utiliza-se a medida de informação mútua (MI) para avaliar como dois agrupamentos C e C' são semelhantes ou diferentes. Esta medida encontra-se definida na equação 6, onde: $p(C_i)$ (resp. $p(C_i, C_j)$) corresponde à probabilidade marginal associada à comunidade C_i (resp. probabilidade conjunta associada às comunidades C_i e C_j).

$$MI(C,C') = \sum_{C_i \in C, C_j \in C'}^{n} p\left(C_i, C_j\right) \times log_2 \frac{p(C_i, C_j)}{p(C_i) \times p(C_j)} \tag{6}$$

A medida de informação mútua normalizada (NMI), versão normalizada da MI, varia entre 0 e 1, e quanto maior o seu valor, mais os dois agrupamentos C e C' são semelhantes.

3. Trabalhos relacionados

Vários algoritmos foram desenvolvidos para resolver o problema de detecção da comunidades [25], [29], [30], [31], [32], [33], [34], [35], [21], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [16], [46], [47]. A maioria deles se concentra exclusivamente nos aspectos topológicos da rede e tenta maximizar o número de conexões em cada comunidade e minimizar o número de conexões entre diferentes comunidades. No entanto, algumas dessas iniciativas se concentraram no uso de agrupamento de dados para detectar comunidades em redes complexas e estão intimamente relacionadas com a abordagem proposta. Por exemplo:

- [16] divide o grafo em k grupos, de modo que cada grupo contenha um subgrafo densamente conectado. Os autores consideram que os vértices pertencentes a um dado grupo possuem valores de atributos, isto é, há informações de domínio associadas a cada vértice. Diferente do presente trabalho, a proposta feita em [16] não é capaz de encontrar comunidades sobrepostas. Trata-se de uma limitação importante, uma vez que diferentes vértices podem pertencer a diferentes comunidades. Por exemplo, uma mesma pessoa pode pertencer a mais de uma comunidade.
- Em [38], os autores procuram detectar comunidades adicionando um vértice à comunidade se seus atributos apresentarem uma boa similaridade com os atributos dos membros já existentes na comunidade, levando em conta também a sua conectividade. De forma análoga a [16], esse trabalho também não detecta comunidades sobrepostas. Além disso, o usuário precisa informar as comunidades anteriormente existentes no conjunto de dados.

- [47] propõe 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). RM-CRAG não trata de atributos nas arestas, diferentemente da abordagem proposta pelo presente artigo. Esta é uma limitação importante, pois, de forma similar ao exemplo da figura 1, muitas redes contêm atributos sobre a aplicação vinculados às arestas.

4. Abordagem proposta

Esta seção descreve conceitualmente os passos da abordagem proposta (ComDet) e resume as principais características do protótipo implementado.

4.1 Descrição conceitual

Toda rede fornecida como entrada para a ComDet deve ser um multigrafo homogêneo com atributos, representado por sua instância e seu esquema. Enquanto o esquema da rede contém meta-dados que indicam os tipos de vértice e de aresta e os atributos contidos no grafo, a instância de rede contém os dados em si. As Figuras 2(a) e 2(b) ilustram, respectivamente, o esquema e a instância de uma rede hipotética de coautoria de publicações. Neste trabalho, foram adotados os diagramas de classe e objeto da UML[7] para representar o esquema e a instância das redes, respectivamente.

Conceitualmente, a abordagem proposta possui seis etapas principais que estão representadas graficamente na figura 3 e descritas a seguir.

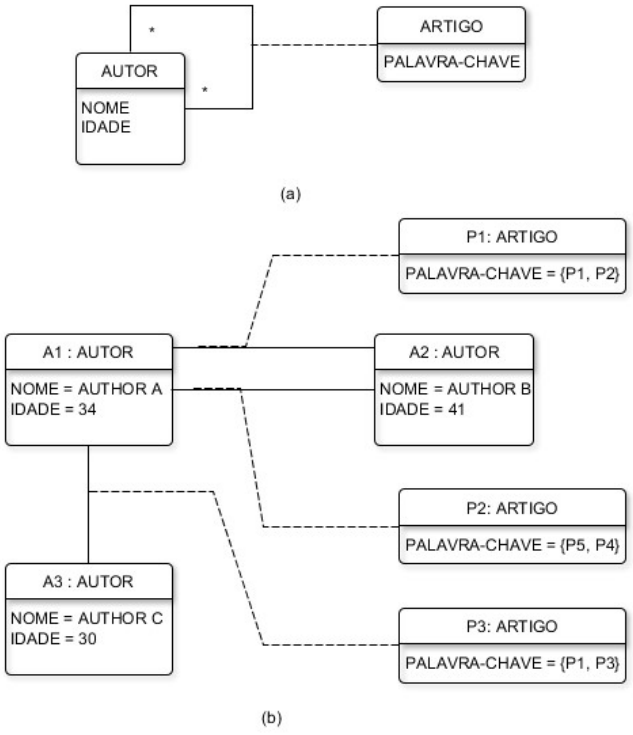

Fig. 2 – Representações de Rede – Exemplo de uma rede de coautoria de publicações: (a) Esquema - Diagrama de Classe; (b) Instância - Diagrama de Objetos.

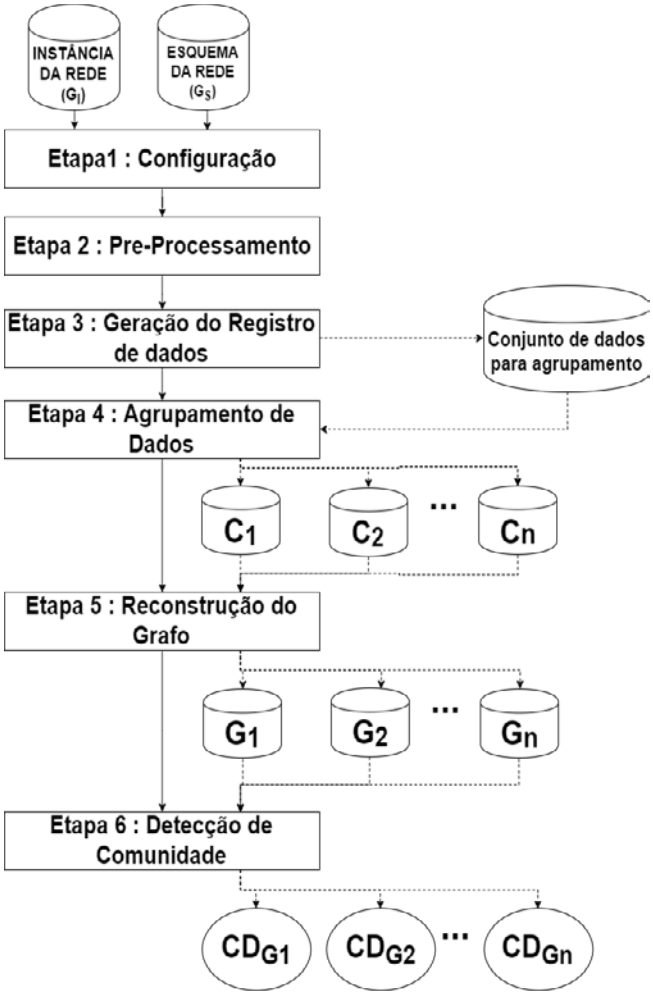

Fig. 3 – Etapas do ComDet: Visão Geral do Processo.

- Configuração - Esta etapa permite ao usuário escolher quais atributos devem ser considerados pelo processo de agrupamento. Cabe ressaltar que tal escolha tem influência direta no conteúdo das comunidades que são identificadas durante o processo. Depende, portanto, dos interesses do domínio da aplicação.
- Pré-Processamento - Esta etapa compreende diversas atividades de preparação de dados feitas antes do processo de agrupamento e que podem incluir, por exemplo, a normalização e a codificação de dados da rede.
- Geração do Registro de Dados - Esta etapa converte a representação de grafo em uma estrutura relacional. Cada par (vértice v, aresta a) é convertido em um registro de dados que contém os atributos tanto de v quanto de *a* que tenham sido escolhidos pelo usuário na etapa de Configuração. A Tabela 1 ilustra o efeito desta conversão aplicada ao exemplo da Figura 2(b). Considere que os atributos escolhidos na primeira etapa tenham sido idade e palavras-chave.
- Agrupamento de Dados - Esta etapa é responsável por separar os registros de dados em grupos de registros similares c_i. O resultado desta etapa é uma coleção de grupos $\{c_i\}$. Embora qualquer algoritmo de agrupamento possa, em princípio, ser aplicado aqui, foi utilizado o algoritmo Affinity Propagation (seção 2.1).
- Reconstrução do Grafo - Para cada grupo c_i identificado na etapa anterior, esta etapa restaura as conexões entre todos os pares de nós de c_i que existem no grafo original. Assim, o grafo resultante G_i é um subgrafo da rede original que

[7] Unified Modeling Language.

compreende vértices e arestas com atributos que compartilham conteúdos semelhantes.

- Detecção de Comunidade - Esta etapa é aplicada a cada grafo reconstruído G_i a fim de identificar subgrafos G_{ij} fortemente conectados (comunidades detectadas em G_i). Na figura 3, CD_{Gi} denota o conjunto de G_{ij} de G_j. Cabe ressaltar que CD_G corresponde à partição identificada pela ComDet, onde $CD_G = CD_{G1} \cup ... \cup CD_{Gn}$. De forma análoga ao descrito em relação à etapa de agrupamento de dados, qualquer algoritmo de detecção de comunidades pode ser aplicado neste ponto. A escolha de qual algoritmo deverá ser aplicado cabe ao analista de dados responsável.

Tabela 1: Resultado da etapa de geração de registros de dados aplicada sobre o exemplo da figura 2, considerando que os atributos escolhidos tenham sido idade e palavras-chave

Record Id	Idade	Palavra-Chave
1	34	{P1, P2}
2	41	{P1, P2}
3	34	{P3, P4}
4	41	{P3, P4}
5	34	{P1, P3}
6	30	{P1, P3}

4.2 Protótipo

O protótipo da ComDet foi codificado em Python[8]. Foram utilizadas as APIs do Scikit-Learn [48] e NetworkX [49] para desenvolver os módulos de aprendizado de máquina e de acesso às estruturas de grafos, respectivamente. Os grafos instância e esquema foram codificados em um formato XML[9] específico chamado DyNetML [50].

Os algoritmos de agrupamento de dados e de detecção de comunidades adotados foram o Affinity Propagation [24] e Girvan-Newman [3], respectivamente. A escolha desses algoritmos foi devido ao fato de que eles não exigem que o usuário informe a quantidade de grupos / comunidades como entrada.

5. Experimentos e resultados

Inicialmente, esta seção descreve o método utilizado para avaliar o desempenho da ComDet. Descreve também as redes utilizadas nos experimentos e os resultados obtidos em cada caso.

5.1 Processo de validação

O processo de avaliação comparou os resultados da ComDet com os resultados obtidos por outros dois algoritmos: Girvan-Newman [3] e LouvainC [21]. Como exposto na Seção 2, o primeiro é um algoritmo de detecção de comunidade que se baseia unicamente na topologia da rede. O segundo é um algoritmo de detecção de comunidades que utiliza como base informações disponíveis sobre o contexto da aplicação. A Figura 4 descreve o processo de avaliação adotado.

Na primeira etapa, cada rede foi submetida aos três algoritmos e cada um deles detectou comunidades. Girvan-Newman e LouvainC geraram um conjunto de comunidades cada (chamado GN_G e LC_G, respectivamente). A ComDet também gerou um conjunto de comunidades, chamado CD_G , onde $CD_G = CD_{G1} \cup ... \cup CD_{Gn}$.

Essas comunidades foram processadas na segunda etapa para uma análise quantitativa. Esta etapa calculou três métricas de avaliação: modularidade, densidade e informação mútua normalizada (NMI). As duas primeiras foram aplicadas separadamente à cada conjunto de comunidades gerado pelos três algoritmos. A modularidade expressa como os vértices das comunidades do conjunto estavam conectados e a densidade indica a concentração de arestas nas comunidades detectadas. Valores elevados dessas métricas indicam que as comunidades detectadas concentram um número elevado de arestas em grupos de vértices altamente interligados, isto é, em termos topológicos, uma boa partição. Por outro lado, a NMI expressa similaridade entre dois conjuntos de comunidades. Tal métrica foi útil para mensurar o quanto as comunidades detectadas pela ComDet foram distintas em relação às comunidades identificadas pelos outros dois algoritmos. Os valores de NMI entre dois conjuntos de comunidades variam de 0 (sem similaridade) a 1 (correlação perfeita). Cabe destacar que a NMI entre os conjuntos de comunidades geradas pelo Girvan-Newman e pelo LouvainC não foi calculada, uma vez que os experimentos não tinham como objetivo comparar os dois algoritmos entre si.

A terceira etapa compreendeu uma análise qualitativa das comunidades detectadas. Foi realizada por um usuário que pesquisou padrões que representavam algum tipo de coerência entre os dados agrupados em cada comunidade.

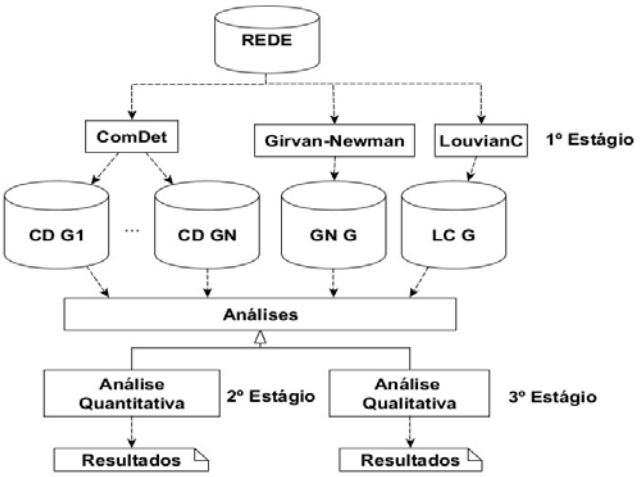

Fig. 4 – Visão Geral do Processo de Avaliação e Comparação dos Resultados.

5.2 Redes

A fim de ilustrar a dualidade de aplicação da abordagem proposta, foram selecionadas três redes para realizar os experimentos, sendo uma de natureza militar, uma pertencente ao contexto científico e a terceira originada em um cenário civil. As três encontram-se disponíveis na web para *download*:

- Militarized Interstate Dispute (MID) - Este conjunto de dados contém informações sobre conflitos entre vários países que ocorreram de 1816 a 2010 [51]. O esquema simplificado do MID é apresentado na Figura 5.
- ArXiv - É um repositório de artigos científicos de várias

[8] eu código está disponível para download em https://goo.gl/Rfuwrf.
[9] https://www.w3.org/XML/

áreas de conhecimento e que podem ser acessados online[1]. Tal repositório inclui informações sobre coautoria e palavras-chave de artigos (ver esquema da rede ArXiv na Figura 6). Nos experimentos foram considerados autores e artigos publicados de 1994 a 1997 em cinco seções de Física: astrofísica, matéria condensada, relatividade geral e cosmologia quântica, física de alta energia fenomenologia e física de alta energia - teoria.

- Enron Email Dataset (Enron) - Este conjunto de dados contém e-mails gerados pelos funcionários da Enron Corporation e adquiridos durante a investigação após o colapso da empresa [52]. Os dados contextuais disponíveis neste conjunto de dados estão no formato textual. A Figura 7 apresenta o esquema da Enron.
- A tabela 2 fornece algumas informações estatísticas sobre esses conjuntos de dados.

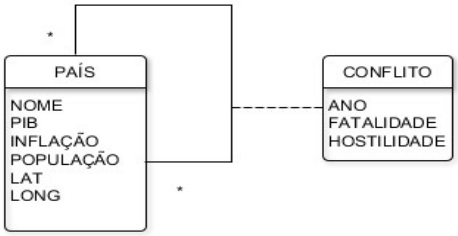

Fig. 5 – Esquema do conjunto de dados MID.

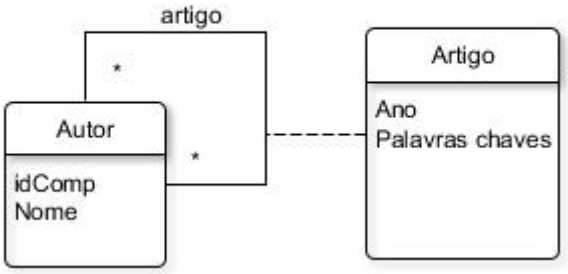

Fig. 6 – Esquema do conjunto de dados ArXiv.

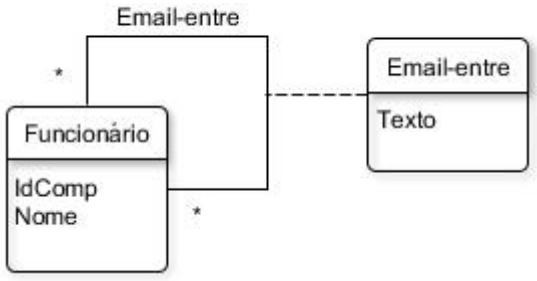

Fig. 7 – Esquema do conjunto de dados Enron.

Tab 2: Estatísticas sobre os conjuntos de dados utilizados nos experimentos

Conjunto de Dados	Nós	Atestas	Coeficiente de agrupamento
MID	2.779	16.544	0,644
ArXiv	902	8.422	0,521
Enron	158	7.139	0,493

5.3 Experimento 1: rede MID

Neste experimento, a ComDet foi configurada para considerar três atributos contextuais: latitude, longitude e o número de mortes (fatalidade). Na etapa de pré-processamento, esses atributos foram normalizados.

A Figura 8(a) apresenta os resultados quantitativos obtidos pela ComDet (CD_G), pelo Girvan-Newman (GN_G) e pelo LouvainC (LC_G) no experimento com a rede MID. O conjunto de comunidades identificado pela ComDet apresentou uma modularidade superior às dos conjuntos identificados pelos outros algoritmos. As comunidades identificadas 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 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.

1. 5.4 Experimento 2: rede arXiv

Neste experimento, foram usadas as palavras-chave dos artigos como informações contextuais. Com tais dados, foram construídas: primeiro, uma matriz TF-IDF associando artigos com palavras-chave e, em seguida, uma matriz de similaridade (entre os artigos). Essa última foi a entrada para a etapa de agrupamento de dados.

Conjunto	Densidade	Modularidade	NMI GN_G	NMI LC_G
CD_G	0,42	0,26	0,70	0,45
GN_G	0,19	0,11	1,00	–
LC_G	0,45	0,15	–	1,00

Conjunto	Densidade	Modularidade	NMI GN_G	NMI LC_G
CD_G	0,80	0,85	0,93	0,93
GN_G	0,82	0,90	1,00	–
LC_G	0,83	0,91	–	1,00

Conjunto	Densidade	Modularidade	NMI GN_G	NMI LC_G
CD_G	0,37	0,59	0,53	0,63
GN_G	0,47	0,32	1,00	–
LC_G	0,45	0,64	–	1,00

Fig. 8 – Resumo dos resultados quantitativos dos experimentos nas redes: (a) MID; (b) ArXiv; e (c) Enron.

1 http://export.arxiv.org/

Sob o ponto de vista quantitativo, os resultados produzidos pela ComDet ficaram bem próximos aos produzidos pelos outros dois algoritmos, conforme mostra a Figura 8(b). No entanto, cabe chamar a atenção para um dos conjuntos produzidos pela ComDet (CD_{G9}). 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 coautoria. 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.

5.5 Experimento 2: rede ENRON

Neste experimento, a ComDet foi configurada para considerar as informações textuais disponíveis no conteúdo dos e-mails. O pré-processamento de tais conteúdos exigiu *tokenização*, remoção de palavras de parada (*stopwords*) e identificação de radicais [53]. Subsequentemente, matrizes de TF-IDF e de similaridade foram construídas e utilizadas pela ComDet.

Em termos quantitativos, a ComDet detectou um conjunto de comunidades, com densidade e modularidade 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 das Figuras 9, 10 e 11. Cada uma delas foi extraída de uma comunidade identificada pela ComDet. 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 9 são comuns e frequentes nas discussões de negociações e finanças. A figura 10 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 11 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.

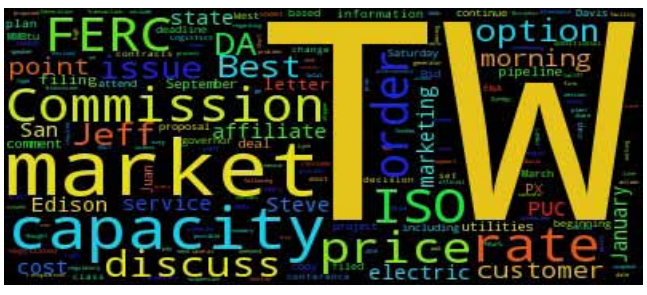

Fig. 9 – Nuvem de palavras extraída de uma comunidade detectada pela ComDet na Rede Enron: negociações e discussões financeiras.

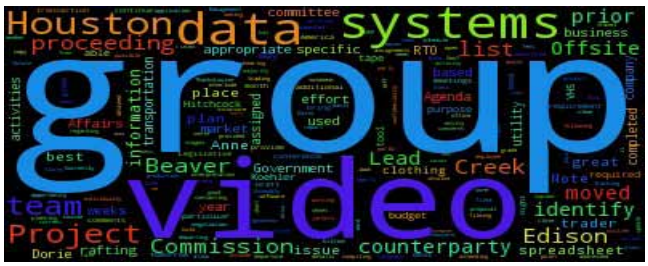

Fig. 10 – Nuvem de palavras extraída de uma comunidade detectada pela ComDet na Rede Enron: problemas de TI.

Fig. 11 – Nuvem de palavras extraída de uma comunidade detectada pela ComDet na Rede Enron: logística, gerenciamento e prazos.

6 Conclusão

A detecção de comunidades é uma importante tarefa de análise de redes complexas que tem sido investigada nos últimos anos tanto pela academia quanto pela indústria. É um problema de otimização que tenta identificar grupos (comunidades) de nós altamente interligados nas redes. Vários algoritmos têm sido propostos para detectar comunidades. A maioria deles está restrita ao uso de informações topológicas. No entanto, apesar da reconhecida importância dos dados contextuais para o problema de detecção de comunidades, poucas iniciativas investigaram o uso desse tipo de informação para identificar grupos coesos de nós conectados que compartilhem conteúdos comuns.

Neste artigo, foi proposta uma abordagem de detecção de comunidades que leva em consideração tanto informações topológicas quanto contextuais. Esta abordagem introduz o agrupamento de dados como uma etapa de pré-processamento para o processo de detecção da comunidade a fim de identificar comunidades estruturalmente densas, coesas e possivelmente sobrepostas. Tal abordagem permite ao analista humano selecionar as informações de contexto consideradas relevantes sob a perspectiva da aplicação e que devem ser utilizadas no processo. Resultados quantitativos e qualitativos obtidos nos experimentos com três redes de diferentes cenários (militar, civil e científico) mostram que, de fato, a combinação de informações contextuais e topológicas pode levar a partições interessantes formadas por comunidades de conteúdo coeso, diferentemente das abordagens tradicionais cujo processo de detecção de comunidades é puramente topológico. Pelos experimentos realizados, foi possível também perceber a relação entre os atributos contextuais selecionados e o conteúdo das comunidades identificadas pela ComDet.

Como trabalho futuro, planeja-se estender a abordagem proposta para lidar com redes heterogêneas. Redes heterogêneas estão cada vez mais frequentes no dia a dia e se caracterizam por conter diferentes tipos de elementos e diferentes tipos de conexões entre eles. Também pretende-se aplicar a abordagem proposta no problema de predição de links. Tal problema consiste em procurar prever se elementos não conectados em uma rede se ligarão no futuro. Acredita-se que prever links a partir de comunidades coesas possa tornar o processo preditivo mais preciso. Além do exposto, novos experimentos aplicando a abordagem proposta em outras redes complexas também deverão ser executados.

Referências Bibliográficas

[1] Wang, P., Xu, B., Wu, Y., and Zhou, X. (2015). Link prediction in social networks: the state-of-the-art. *SCIENCE CHINA Information Sciences*, 58(1):1–38.

[2] Pecli, A., Giovanini, B., Pacheco, C. C., Moreira, C., Ferreira, F., Tosta, F., Tesolin, J., Dias, M. V., Filho, S., Cavalcanti, M. C., and Goldschmidt, R. (2015). Dimensionality Reduction for Supervised Learning in Link Prediction Problems. In Proceedings of the 17th International Conference on Enterprise Information Systems, pp. 295– 302.

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

[4] Fortunato, S. and Darko, H. (2016). Community detection in networks:

[5] A user guide. *CoRR*, volume abs1608.00163.

[6] Adafre, S. F. and de Rijke, M. (2005). Discovering missing links in Wikipedia. In Proceedings of the 3rd international workshop on Link discovery, pp. 90–97. ACM.

[7] Zhu, J., Hong, J., and Hughes, J. G. (2002). Using markov models for web site link prediction. In *HYPERTEXT 2002, Proceedings of the 13th ACM Conference on Hypertext and Hypermedia, June 11-15, 2002, University of Maryland, College Park, MD, USA*, pp. 169–170.

[8] Li, X. and Chen, H. (2009). Recommendation as link prediction: a graph kernel-based machine learning approach. In *Proceedings of the 2009 Joint International Conference on Digital Libraries, JCDL 2009, Austin, TX, USA, June 15-19, 2009*, pp. 213–216.

[9] Liu, Y. and Kou, Z. (2007). Predicting who rated what in large-scale datasets. *SIGKDD Explorations*, 9(2), pp. 62–65.

[10] Huang, Z., Li, X., and Chen, H. (2005). Link prediction approach to collaborative filtering. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2005, Denver, CO, USA, June 7-11, 2005, Proceedings*, pp. 141–142.

[11] Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate Record Detection: A Survey. *IEEE Trans. Knowl. Data Eng.*, volume 19(1), pp. 1–16.

[12] Malin, B., Airoldi, E., and Carley, K. M. (2005). A Network Analysis Model for Disambiguation of Names in Lists. *Computational & Mathematical Organization Theory*, 11(2), pp. 119–139.

[13] Airoldi, E. M., Blei, D. M., Fienberg, S. E., Xing, E. P., and Jaakkola, T. (2006). Mixed membership stochastic block models for relational data, with applications to protein-protein interactions. In *Proceedings of International Biometric Society-ENAR Annual Meetings*.

[14] Freschi, V. (2009). A Graph-Based Semi-supervised Algorithm for Protein Function Prediction from Interaction Maps. In *Learning and Intelligent Optimization, Third International Conference, LION 3, Trento, Italy, January 14-18, 2009, Selected Papers*, pp. 249–258.

[15] Krebs, V. E. (2002). Mapping Networks of Terrorist Cells. *Connections*, 24(3), pp. 43–52.

[16] Guedes, G. P., Ogasawara, E., Bezerra, E., and Xexeo, G. (2016). Discovering top-k non-redundant clusterings in attributed graphs. *Neurocomputing*, 210:45 – 54. SI:Behavior Analysis In {SN}.

[17] Zhou, Y., Cheng, H., and Yu, J. X. (2009). Graph clustering based on structural/attribute similarities. *PVLDB*, 2(1):718–729. 15 dez. de 2015.

[18] Tang, L. and Liu, H. (2010). Community Detection and Mining in Social Media. *Synthesis Lectures on Data Mining and Knowledge Discovery*. 18 jun. de 2015.

[19] Li, X. and Chen, H. (2009). Recommendation as link prediction: a graph kernel-based machine learning approach. In *Proceedings of the 2009 Joint International Conference on Digital Libraries, JCDL 2009, Austin, TX, USA, June 15-19, 2009*, pp. 213–216.

[20] Goldberg, M. K., Kelley, S., Magdon-Ismail, M., Mertsalov, K., and Wallace, A. (2010). Finding Overlapping Communities in Social Networks. In Elmagarmid, A. K. and Agrawal, D., editors, *SocialCom/PASSAT*, pp. 104–113. IEEE Computer Society.

[21] Naruchitparames, J., Gunes, M. H., and Louis, S. J. (2011). Friend recommendations in social networks using genetic algorithms and network topology. In *IEEE Congress on Evolutionary Computation*, pp. 2207– 2214. IEEE.

[22] 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. 15 jan. de 2015.

[23] Anderberg and Michael, R. (1973). Cluster Analysis for Applications. *Academic Press*, serie Probability and Mathematical Statistics.

[24] Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002). Clustering Validity Checking Methods: Part II. *SIGMOD Rec.*, 31(3):19–27. 01 set. de 2015.

[25] Brusco, M. J. and Köhn, H. (2008). Clustering by Passing Messages Between Data Points. *Science*, volume 319, pp. 726–726. 31 nov. de 2015.

[26] Naruchitparames, J., Gunes, M. H., and Louis, S. J. (2011). Friend recommendations in social networks using genetic algorithms and network topology. In *IEEE Congress on Evolutionary Computation*, pp. 2207– 2214. IEEE.

[27] Newman, M. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

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

[29] 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), pp. 222–242. 14 jun. de 2016.

[30] Airoldi, E. M., Blei, D. M., Fienberg, S. E., Xing, E. P., and Jaakkola, T. (2006). Mixed membership stochastic block models for relational data, with applications to protein-protein interactions. In *Proceedings of International Biometric Society-ENAR Annual Meetings*.

[31] Yang, S. and Luo, S. (2009) Community detection based on adaptive kernel affinity propagation. *2009 2nd IEEE International Conference on Computer Science and Information Technology*, USA, Aug, pp. 1–4. IEEE.

[32] Ding, F., Luo, Z., Shi, J., and Fang, X. (2010) Overlapping Community Detection by Kernel-Based Fuzzy Affinity Propagation *2010 2nd International Workshop on Intelligent Systems and Applications*, USA, May, pp. 1–4. IWISA.

[33] Leskovec, J., Lang, K. J., and Mahoney, M. W. (2010) Empirical comparison of algorithms for network community detection. cite arxiv:1004.3539.

[34] Chen, Y.-L., Chuang, C.-H., and Chiu, Y.-T. (2014) Community detection based on social interactions in a social network. *JASIST*, 65, 539–550.

[35] 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.

[36] Yizhou, Sun, Norick, Brandon, Han, Jiawei, Yan, Xifeng, Yu, S., P., and Yu, X. (2012) Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In Yang, Q., Agarwal, D., and Pei, J. (eds.), *KDD*, USA, 7, pp. 1348–1356. ACM. 19 maio de 2016.

[37] Melamed, D. (2014) Community structures in bipartite networks: A dual-projection approach. *PLoS ONE*, 9, e97823. 10 mar. de 2015.

[38] Meng, Q., Tafavogh, S., and Kennedy, P. (2014) Community detection on heterogeneous networks by multiple semantic-path clustering. *Computational Aspects of Social Networks (CASoN), 2014 6th International Conference on*, USA, July, pp. 7–12. CASoN.

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

[40] Günnemann, S., Farber, I., Boden, B., and Seidl, T. (2010). Subspace Clustering Meets Dense Subgraph Mining: A Synthesis of Two Paradigms. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 845–850.

[41] Dang, T. A. and Viennet, E. (2012) Community Detection based on Structural and Attribute Similarities. *ICDS*, pp. 7–14. 10 mar. de 2015.

[42] Akoglu, L., Tong, H., Meeder, B., and Faloutsos, C. (2012) PICS: Parameter-free Identification of Cohesive Subgroups in Large Attributed Graphs. *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, USA.

[43] Ruan, Y., Fuhry, D., and Parthasarathy, S. (2012) Efficient Community Detection in Large Networks using Content and Links. *ArXiv e-prints*, 1, 1.

[44] Ruan, Y., Fuhry, D., and Parthasarathy, S. (2013) Efficient community detection in large networks using content and links. *Proceedings of the 22nd international conference on World Wide*

Web, USA, 3, pp. 1089– 1098. ACM ACM.

[45] Leskovec, J., Lang, K. J., and Mahoney, M. W. (2010) Empirical comparison of algorithms for network community detection. cite arxiv:1004.3539.

[46] Zhang, Y., Levina, E., and Zhu, J. (2015) Community Detection in Networks with Node Features. *ArXiv e-prints*, 1, 1.

[47] LAI, D. and LU, H. (2008) IDENTIFICATION OF COMMUNITY STRUCTURE IN COMPLEX NETWORKS USING AFFINITY PROPAGATION CLUSTERING METHOD. *Modern Physics Letters B*, 22, 1547–1566.

[48] Sachan, M., Contractor, D., Faruquie, T. A., and Subramaniam, L. V. (2012) Using content and interactions for discovering communities in social networks. In Mille, A., Gandon, F. L., Misselis, J., Rabinovich, M., and Staab, S. (eds.), *WWW*, USA, 8, pp. 331–340. ACM.

[49] Guedes, G. P., Ogasawara, E., Bezerra, E., and Xexeoa, G. (2015). Discovering top-k non-redundant clusterings in attributed graphs. *Programa de Engenharia de Sistemas e Computação/COPPE, Universidade Federal do Rio de Janeiro*, 1:1.

[50] 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.

[51] 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*, pp. 11–15.

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

[53] 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), pp. 222–242. 14 jun. de 2016.

[54] 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 International Conference on Knowledge Discovery and Data Mining*, number 1 in KDD '07, pp. 687–696, New York, NY, USA. ACM. 15 jun. de 2016.

[55] Porter, M. F. (2001). Snowball: A language for stemming algorithms. Published online. Accessed 11.03.2015, 15.00h.