

INSTITUTO MILITAR DE ENGENHARIA

LEANDRO FARIAS MAIA E HEYTOR BRUNO NOBRE

**DETECÇÃO DE COMUNIDADES EM REDE DE
RELACIONAMENTO DE ALUNOS**

Monografia de Projeto Fim de Curso apresentada ao Curso de Graduação em Engenharia de Computação do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Graduação em Engenharia de Computação.

Orientador: Claudia Marcela Justel - D.Sc.

Rio de Janeiro
2012

c2012

INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80-Praia Vermelha
Rio de Janeiro-RJ CEP 22290-270

Este exemplar é de propriedade do Instituto Militar de Engenharia, que poderá incluí-lo em base de dados, armazenar em computador, microfilmear ou adotar qualquer forma de arquivamento.

É permitida a menção, reprodução parcial ou integral e a transmissão entre bibliotecas deste trabalho, sem modificação de seu texto, em qualquer meio que esteja ou venha a ser fixado, para pesquisa acadêmica, comentários e citações, desde que sem finalidade comercial e que seja feita a referência bibliográfica completa.

Os conceitos expressos neste trabalho são de responsabilidade do(s) autor(es) e do(s) orientador(es).

623.85 Maia, Leandro Farias

M217d Detecção de comunidades em rede de relacionamento de alunos/
Leandro Farias Maia, Heytor Bruno Nobre; orientado por Cláudia
Marcela Justel - Rio de Janeiro : Instituto Militar de Engenharia,
2012.

54 p. : il.

Projeto de final de Curso - Instituto Militar de Engenharia -
Rio de Janeiro, 2012.

1. Engenharia de Computação 2. Algoritmos 3. Comunidades-
redes I. Nobre, Heytor Bruno II. Justel, Claudia Marcela III. De-
tecção de comunidades em rede de relacionamento de alunosIV.
Instituto Militar de Engenharia

623.85

INSTITUTO MILITAR DE ENGENHARIA

LEANDRO FARIAS MAIA E HEYTOR BRUNO NOBRE

**DETECÇÃO DE COMUNIDADES EM REDE DE RELACIONAMENTO
DE ALUNOS**

Monografia de Projeto Fim de Curso apresentada ao Curso de Graduação em Engenharia de Computação do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Graduação em Engenharia de Computação.

Orientador: Claudia Marcela Justel - D.Sc.

Aprovada em 13 de Outubro 2011 pela seguinte Banca Examinadora:

Claudia Marcela Justel - D.Sc. COPPE-UFRJ - Orientadora

MAJ Julio Cesar Duarte - D.C. da PUC-Rio

TC Luiz Henrique da Costa Araújo - D.Sc. COPPE-UFRJ

Rio de Janeiro
2012

SUMÁRIO

LISTA DE ILUSTRAÇÕES	6
LISTA DE TABELAS	8
LISTA DE SIGLAS	9
1 INTRODUÇÃO	11
1.1 Objetivo	12
1.2 Motivação	12
1.3 Metodologia	12
1.4 Organização da Monografia	13
2 CONCEITOS DE GRAFOS, REDES SOCIAIS E COMU- NIDADES	14
2.1 Grafos	14
2.2 Redes Sociais e Comunidades	14
3 UM ESTUDO DE TRÊS ALGORITMOS PARA DETECÇÃO DE COMUNIDADES	18
3.1 Identificação da Estrutura da comunidade	18
3.2 O Algoritmo CNM	19
3.2.1 Um exemplo para o algoritmo CNM	20
3.3 O algoritmo de Wakita	23
3.4 O algoritmo de Blondel	24
4 ANÁLISE DA REDE DOS ALUNOS DO SEGUNDO ANO DO IME	28
4.1 Grafo Questionário	28
4.2 Grafo Seção	31
4.3 Comparação das comunidades obtidas por CNM com dados reais do 3º ano de 2012	33
4.3.1 Comparação com o Grafo Questionário reduzido	34
4.4 Construção do novo Grafo Questionário	38

4.4.1	Comparação do novo Grafo Questionário com o Grafo Seção reduzido	40
5	GERADOR DE GRAFOS	42
5.1	Projeto	43
5.1.1	Processo Completo	43
5.1.2	Caso de Uso	45
5.2	Cadastrando Alunos	47
5.3	Preenchendo a Ficha de Escolha de Especialidade	48
5.4	Respondendo o Questionário	49
6	CONCLUSÃO	51
7	REFERÊNCIAS BIBLIOGRÁFICAS	53

LISTA DE ILUSTRAÇÕES

FIG.2.1	Exemplo de grafo com três comunidades (Figura extraída de [4])	15
FIG.2.2	Rede de Zachary (Figura extraída de [6])	16
FIG.2.3	Rede de golfinhos (Figura extraída de [4])	17
FIG.3.1	Pseudo-código para o algoritmo de CNM.	19
FIG.3.2	Bipartição da Rede de Zachary obtida por Girvan e Newman (Figura extraída de [5])	20
FIG.3.3	Bipartição da Rede de Zachary obtida por Clauset, Newman e Moore (Figura extraída de [5])	21
FIG.3.4	Grafo exemplo com 6 comunidades	21
FIG.3.5	Grafo exemplo com 5 comunidades	22
FIG.3.6	Grafo exemplo com 4 comunidades	22
FIG.3.7	Grafo exemplo com 3 comunidades	23
FIG.3.8	Grafo exemplo com 2 comunidades	23
FIG.3.9	Exemplo de aplicação do algoritmo de Blondel (Figura extraída de [2])	25
FIG.3.10	Pseudo-código para o algoritmo de Blondel.	26
FIG.4.1	Grafo Questionário e as 5 comunidades detectadas pelo algoritmo CNM	30
FIG.4.2	Grafo Questionário Reduzido e as 6 comunidades detectadas pelo algoritmo CNM	31
FIG.4.3	Grafo Seção e as 4 comunidades detectadas pelo algoritmo CNM	33
FIG.4.4	Grafo Seção Reduzido e as 4 comunidades detectadas pela algoritmo CNM	34
FIG.4.5	Grafo Seção reduzido e comunidades definidas pela escolha de es- pecialidade em fevereiro de 2012 dos atuais alunos do 3ºano	35
FIG.4.6	Grafo Questionário reduzido e comunidades definidas pela escolha de especialidade em fevereiro de 2012 dos atuais alunos do 3ºano	37
FIG.4.7	Grafo do Novo Questionário montado com a ferramenta desen- volvida, já com as comunidades detectadas	39

FIG.5.1	Exemplo de uma entrada para o software	42
FIG.5.2	Grafo de Exemplo	42
FIG.5.3	Processo Completo	44
FIG.5.4	Processo cadastro de alunos	44
FIG.5.5	Processo preencher formulário(responder questionário/ficha de es- colha)	44
FIG.5.6	Diagrama de casos de uso.	45
FIG.5.7	Página inicial do sistema de Cadastro de alunos.	47
FIG.5.8	Página de confirmação do cadastro de um aluno.	47
FIG.5.9	Página inicial para responder a Ficha de Escolha de Especialidades.	48
FIG.5.10	Página onde o aluno seleciona suas escolhas de especialidades.	48
FIG.5.11	Página de conferência das respostas.	49
FIG.5.12	Página inicial para o sistema de resposta do questionário.	49
FIG.5.13	Página do questionário	50
FIG.5.14	Página de conferência do questionário	50

LISTA DE TABELAS

TAB.4.1	Quantidade de pessoas que estão na interseção das duas Comunidades obtidas pelo Algoritmo CNM nos dois Grafos.	40
---------	---	----

RESUMO

Neste trabalho foram estudados três algoritmos para detecção de comunidades em redes sociais. O primeiro deles, proposto por Clauset, Newman e Moore (conhecido por CNM, as iniciais dos criadores), é baseado na otimização da medida de qualidade denominada modularidade. O Algoritmo de Wakita é o segundo deles e propõe uma melhoria do CNM: uma taxa de consolidação que leva em consideração o tamanho das comunidades e induz a um crescimento equilibrado delas, sendo assim mais adequado para redes maiores. O terceiro algoritmo, conhecido por Blondel, também utiliza a modularidade, e trabalha em duas etapas, dessa forma é possível trabalhar com redes de tamanho maior em tempo satisfatório.

A partir desse estudo, foi realizado um experimento com dados reais obtidos com alunos do segundo ano do curso básico do IME. Foram modelados dois grafos contendo informações fornecidas pelos alunos durante o ano 2011 para analisar a formação de grupos utilizando o algoritmo de CNM e comparamos os grupos obtidos entre si.

Como produto desse experimento, foi criado um site para gerar automaticamente grafos a partir de informações reais inseridas via formulário. O grafo produzido foi criado respeitando as condições da entrada para o algoritmo de CNM.

ABSTRACT

In this paper, three algorithms to detect communities in social networks were studied. The first, proposed by Clauset, Newman and Moore (known by the initials of the creators, CNM), is based on the optimization of the quality measure called modularity. The second one, Wakita algorithm, is based on the algorithm CNM, but proposes an improvement of the same; this algorithm uses a consolidation rate that takes into account the size of communities and leads to a balanced growth between the communities, so it is more appropriate for larger networks. The third algorithm, known as Blondel, also uses the modularity, and works in two steps, so you can work in suitable time with larger networks.

From this study, an experiment was performed with real data obtained with the second year students of the basic course of the IME. Two graphs containing information provided by students during the years 2011 were molded to analyze the formation of groups using the CNM algorithm and were compared with each other about the formation of these groups.

As a product of this experiment, a site was created to automatically generate graphs from actual information obtained in the form. The graph produced was created in compliance with the conditions of the entry for the CNM algorithm.

1 INTRODUÇÃO

As redes sociais geram um grande interesse por partes dos usuários e representam um grande atrativo também para a comunidade científica. Sua análise é encarada como um grande desafio e as descobertas que pesquisas relacionadas podem proporcionar podem ser apenas imaginadas. Um dos problemas mais populares atualmente está relacionado à detecção de comunidades, um problema interessante, mas de difícil abordagem. As redes são normalmente representadas como grafos, onde os indivíduos são os vértices e a relação entre eles, arestas ([1]).

As redes sociais podem ser constituídas a partir das relações interpessoais diárias de um grupo de indivíduos e o estudo dos grupos pode mostrar como estes afetam as decisões e o comportamento de um indivíduo do grupo. A detecção de comunidades por algoritmos é uma das linhas de pesquisa em redes ([5]).

Do ponto de vista da modelagem do problema, uma das grandes dificuldades é identificar a melhor representação da relação entre os indivíduos a ser descrita pelas arestas no grafo, ou seja, qual aspecto dos relacionamentos interpessoais deve ser considerado para que se encontrem resultados significativos na detecção de comunidades e o que essas comunidades detectadas irão representar.

O problema da detecção de comunidades recai na divisão da rede em partições do conjunto de vértices densamente conectados. Existem vários tipos de algoritmos de detecção de comunidades: algoritmos divisivos detectam arestas intercomunitárias e as removem ([5]); algoritmos aglomerativos unem nós/comunidades recursivamente e métodos de otimização são baseados na maximização de uma função objetivo ([3], [2]).

Existem diferentes abordagens para determinar comunidades que dependem da função utilizada para descrever as características dos grupos. Portanto, escolher o algoritmo mais apropriado para cada tipo de problema determinará a corretude e maior exatidão possível dos resultados obtidos.

1.1 OBJETIVO

O objetivo desse trabalho é aplicar diferentes métodos de detecção de comunidades em uma rede de relacionamentos de alunos baseada em dados reais, a fim de estudar seu comportamento e como a formação de diferentes grupos dentro de uma universidade é determinada por essa estrutura. Com uma análise dos resultados obtidos após a aplicação dos algoritmos, espera-se obter uma previsão do comportamento dos membros da rede em situações práticas. Para tornar esse processo de análise mais eficiente, buscar-se-á automatizar o processo de criação do grafo que representa a rede de alunos estudada a partir de um formulário.

1.2 MOTIVAÇÃO

A detecção de comunidades em grafos que representam redes reais é uma área de pesquisa em foco. A popularização da internet nos últimos anos veio acompanhada de uma grande expansão de redes sociais, sites de relacionamento pessoal. A possibilidade de visualização dessas redes como grafos trouxe um grande interesse por pesquisas nessa área e atualmente não só redes sociais e a internet são estudados. Algoritmos eficientes já foram desenvolvidos e a possibilidade de previsão de comportamento através das comunidades em um grafo representante de uma rede é um tema muito interessante.

1.3 METODOLOGIA

Para atingir os objetivos desse trabalho, serão estudados diversos métodos de detecção de comunidades. Os algoritmos correspondentes serão implementados e serão posteriormente aplicados em um grafo de relacionamentos elaborado a partir de dados reais sobre alunos do curso básico do Instituto Militar de Engenharia. Os resultados obtidos serão analisados em comparação com o comportamento real da rede estudada. Dessa forma será determinada a validade do uso do método para prever comportamentos futuros em redes similares. A seguir, são apresentadas as fases do trabalho:

- (1) Estudo dos conceitos básicos em redes sociais e detecção de comunidades, como definições de grafos e comunidades ([1], [4]).
- (2) Estudo de algoritmos existentes na literatura para determinar comunidades, onde podemos destacar o CNM, Wakita e Blondel ([7], [4], [2]).

- (3) Implementação dos algoritmos escolhidos para usar na análise da rede proposta, adicionando como novas funcionalidades de um software desenvolvido em um PFC anterior.
- (4) Construção da rede de relacionamentos baseada em dados reais dos alunos do curso básico do Instituto Militar de Engenharia sobre a escolha da especialidade.
- (5) Aplicação dos algoritmos de detecção de comunidades implementados no passo (3) na rede de relacionamentos construída no passo (4).
- (6) Automatizar o processo de criação do grafo com dados reais de alunos do segundo ano do básico do IME a partir de um formulário.

1.4 ORGANIZAÇÃO DA MONOGRAFIA

O restante da monografia está organizada da seguinte maneira: O Capítulo 2 apresenta uma introdução sobre comunidades e redes sociais, abordando alguns problemas na literatura computacional. O Capítulo 3 apresenta a descrição de três algoritmos estudados sobre detecção de comunidades em uma rede baseado nos artigos [7] e [2]. O Capítulo 4 apresenta a análise dos resultados obtidos pela aplicação de um algoritmo de detecção de comunidades em uma rede de relacionamentos construída a partir de dados de alunos do segundo ano do IME. O Capítulo 5 aborda o projeto do site e suas motivações. Por fim, o Capítulo 6 traz a conclusão dos resultados obtidos no trabalho e o Capítulo 7 apresenta a bibliografia utilizada.

2 CONCEITOS DE GRAFOS, REDES SOCIAIS E COMUNIDADES

Neste capítulo serão introduzidos alguns conceitos básicos sobre grafos, redes sociais e comunidades.

2.1 GRAFOS

Serão abordados apenas algumas definições que serão utilizadas posteriormente, sem entrar em detalhes.

- Um grafo *não orientado* é uma estrutura $G = G(V, E)$ formada por um conjunto finito e não vazio de vértices $V = \{v_1, \dots, v_n\}$ e um conjunto de arestas $E = \{\{v_i, v_j\} | \text{para alguns } 1 \leq i, j \leq n\}$. Nesta monografia, quando nada for dito, entende-se o grafo como não orientado.
- O grau de um vértice, $\deg(v)$ denotada a quantidade de arestas que saem ou entram (grafo não orientado) neste vértice.
- Um corte é o particionamento do conjunto de vértices do grafo $G=(V,E)$ em dois conjuntos disjuntos S e $V - S$. Denota-se o corte pelo par $(S, V - S)$
- Um grafo completo é um grafo onde existe uma aresta entre quaisquer dois vértices.
- Um k -clique em um grafo é um subgrafo que também é um grafo completo com k vértices.

2.2 REDES SOCIAIS E COMUNIDADES

Os primeiros passos da teoria das redes encontram-se principalmente nos trabalhos do matemático *Ėuler* que criou os primeiros resultados da teoria dos grafos. Um grafo é uma representação de um conjunto de nós conectados por arestas que, em conjunto, formam uma **rede**. Existem exemplos, nos diversos domínios, de redes: redes de comunicação, redes sociais, redes biológicas e redes de informação. Uma **rede social** pode ser entendida como uma rede na qual o conjunto de objetos são as pessoas e as conexões entre elas são relações sociais.

A organização de uma rede social compreende a totalidade de relações de um determinado agrupamento social. Neste sentido, pode-se dizer que a organização é composta pela interação social que constitui as relações de determinado grupo. Em cima dessa nova idéia, vários pesquisadores dedicaram-se ao trabalho de compreender quais são as propriedades dos grafos e como se gera o processo de sua construção, ou seja, como seus nós se agrupavam.

Há algum tempo, pode-se perceber um grande crescimento do número de redes sociais obtidas a partir de vários modelos de interação humana e social. Redes sociais como *Facebook*, *LinkedIn*, *Wikipedia*, entre outras, mostram a atual tendência de organização e comunicação entre as pessoas.

Entende-se por **comunidade** um grupo de vértices que possuem propriedades comuns, e/ou desempenham uma função similar dentro de uma rede. A formulação matemática para determinar comunidades em uma rede é denominada particionamento de grafos(graph clustering).

A Figura 2.1 apresenta um grafo onde são detectadas três comunidades, cada uma circundada pela linha tracejada.

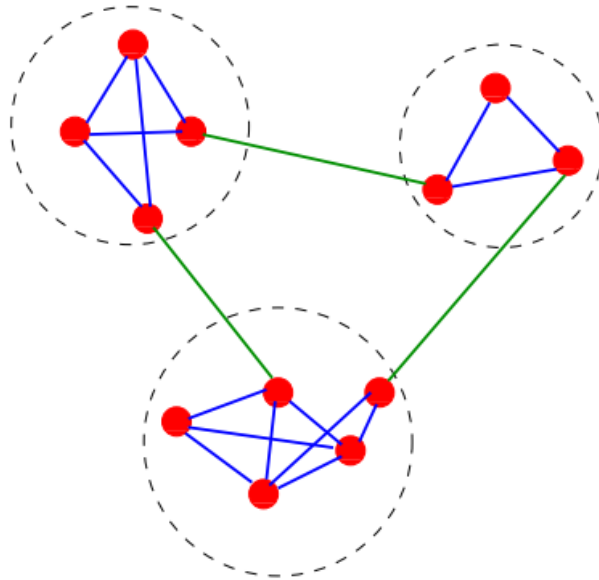


FIG. 2.1: Exemplo de grafo com três comunidades (Figura extraída de [4])

A Figura 2.2 representa uma rede, conhecida como **Rede de Zachary**, formada por membros de um clube de Karatê. Esta rede é composta por 34 vértices, os membros de um clube de karate nos Estados Unidos, que foram observados durante um período de

três anos. As arestas conectando os indivíduos indicam uma interação dos mesmos fora das atividades do clube.

Em algum momento, um conflito entre o presidente do clube e o instrutor levou à cisão do clube em dois grupos separados, um grupo apoiando o instrutor e o outro grupo apoiando o presidente. A questão é se a partir da estrutura de rede original é possível inferir a composição dos dois grupos. De fato, olhando Figura 2.2 pode-se distinguir duas agregações, um em torno de vértices 33 e 34 (34 é o presidente), o outro em torno do vértice 1 (o instrutor). Pode-se também identificar vários vértices situados entre as duas estruturas principais, como o 3, 9, 10; tais vértices são muitas vezes classificados erroneamente por métodos de detecção da comunidade ([4]).

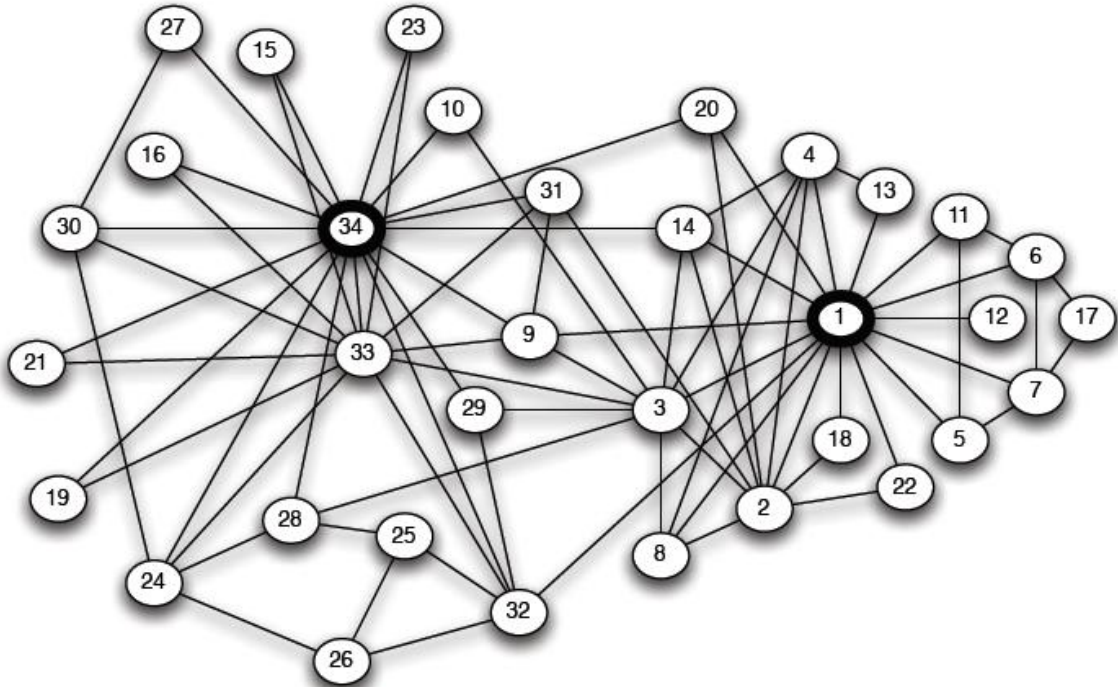


FIG. 2.2: Rede de Zachary (Figura extraída de [6])

A Figura 2.3 mostra a rede de golfinhos que vivem em Doubtful Sound (Nova Zelândia) analisados por Lusseau ([4]). Há 62 golfinhos e as arestas foram estabelecidas entre os animais que foram vistos juntos com mais frequência do que o esperado. Os golfinhos se separaram em dois grupos após um golfinho deixar o local por algum tempo (quadrados e círculos na figura). Tais grupos são bastante coesos, com várias cliques internos, e facilmente identificáveis: há apenas seis arestas unindo vértices de grupos diferentes. Devido a esta classificação natural da rede de golfinhos de Lusseau, como clube de karate

de Zachary, costuma-se frequentemente testar algoritmos para a detecção de comunidades ([4]).

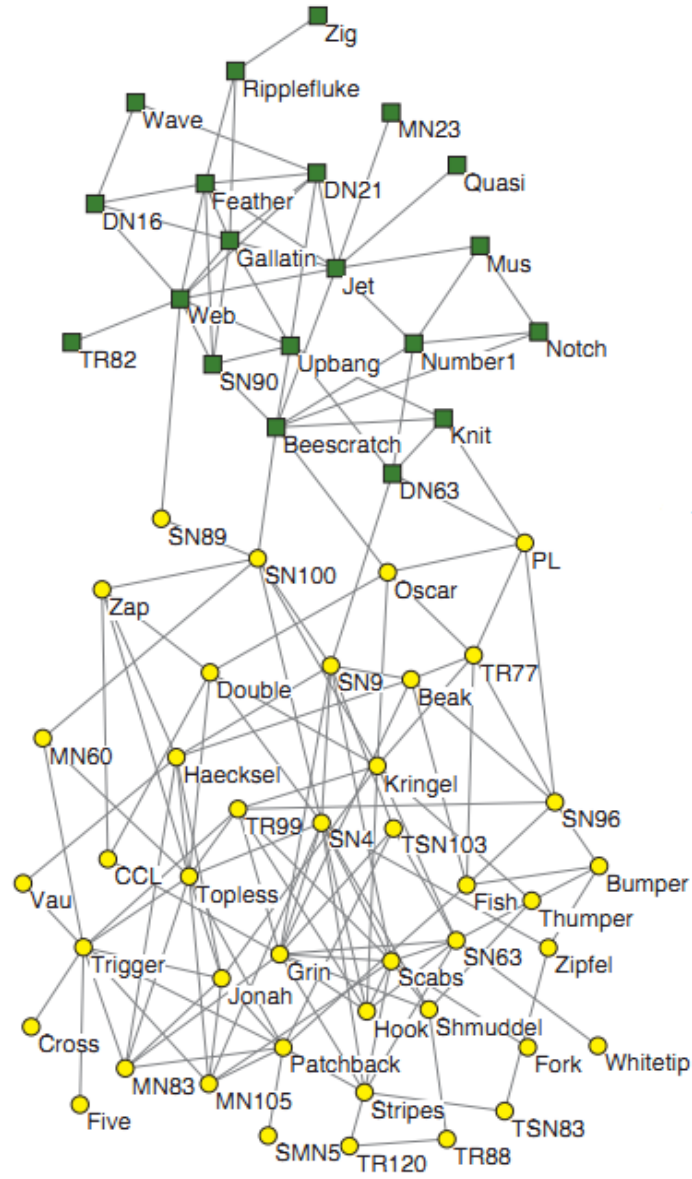


FIG. 2.3: Rede de golfinhos (Figura extraída de [4])

3 UM ESTUDO DE TRÊS ALGORITMOS PARA DETECÇÃO DE COMUNIDADES

As redes sociais apresentam um desafio considerável, tendo em vista seu enorme tamanho e sua rápida expansão. Assim, a detecção de comunidades tornou-se um problema muito interessante para estudar a estrutura da rede tanto topologicamente quanto sintaticamente.

Algoritmos eficientes para detectar comunidades em redes com milhares de vértices são importantes. No artigo [7], são analisados três algoritmos de detecção de comunidades: Clauset-Newman-Moore (CNM), Wakita e Blondel. Estes três algoritmos utilizam a modularidade como métrica de otimização. Serão introduzidas duas métricas importantes no entendimento dos algoritmos que serão abordados: probabilidade e consistência. Será proposta uma solução que melhora a consistência dos resultados sem comprometer a modularidade. A idéia é estabelecer os pares de probabilidades como um link com peso e achar comunidades em uma rede iterativa com pesos. Mostra-se que ambos, a modularidade e a consistência, convergem para a maioria das redes.

3.1 IDENTIFICAÇÃO DA ESTRUTURA DA COMUNIDADE

O conceito de modularidade (Q) representa a qualidade de um particionamento de um grafo em comunidades. Dado um grafo $G = (V, E)$ com $|V| = n$, $|E| = m$ e um particionamento $C = \cup C_i$ onde $C_i \subseteq V$, $C_i \neq \emptyset$, $C_i \cap C_j = \emptyset$, define-se $Q(C) = \sum e_{i,i} - a_i^2$, Sendo $e_{i,i}$ a razão do número de arestas entre vértices pertencentes a comunidade i sobre o número de arestas do grafo (m) e a_i a razão de todas as arestas que atravessam a fronteira da comunidade i sobre todas as arestas do grafo.

O valor da modularidade, $Q(C)$, considerando todas as possíveis partições, varia de -1 a 1. O valor $Q = 0$ implica que o número de arestas dentro de uma comunidade não é melhor do que um número obtido aleatoriamente. O valor $Q = 1$ é o máximo possível, mas na prática nunca é atingido.

O algoritmo de CNM procura um particionamento do grafo em comunidades de maneira tal que seja maximizada a modularidade. A abordagem utilizada para obter

o resultado desejado é um algoritmo guloso que parte de um conjunto de n comunidades unitárias que são unidas, iterativamente, tentando a cada iteração aumentar o valor da modularidade.

3.2 O ALGORITMO CNM

Newman e Girvan definiram a modularidade e propuseram um algoritmo simples, para detectar comunidades, de complexidade $O(n^3)$ para uma rede de n nós, conhecido na literatura com algoritmo de Girvan e Newman ([5]). Newman propôs um algoritmo melhorado baseado em clusterização aglomerativa hierárquica e reduziu a complexidade para $O(n^2)$. Posteriormente, Clauset, Newman e Moore propuseram um algoritmo mais eficiente na detecção de comunidades (denotado como CNM)[3]. A complexidade do algoritmo foi reduzida para $O(m.d.\log(n))$, sendo d a profundidade de uma árvore binária balanceada e m o número de arestas da rede.

A Figura 3.1 apresenta o pseudo-código do algoritmo CNM. Este algoritmo começa considerando cada nó da rede como uma comunidade diferente (linha 4).

```

1:  $V$ : conjunto de vértices
2:  $E$ : conjunto de arestas
3:  $G \leftarrow (V, E)$ 
4:  $C \leftarrow \{\{v_i\} | v_i \in G(V)\}$ 
5:  $H$ : máxima heap de  $\Delta Q_{c_x, c_y}$ 
6: Enquanto  $H$  é não vazio Faça
7:   extraí-se  $\max \Delta Q_{c_x, c_y}$  de  $H$ , onde qualquer  $c_x, c_y \in C$ 
8:   Se  $\max \Delta Q_{c_x, c_y} < 0$  Então
9:     Pare!
10:  Fim do Se
11:   $c_z \leftarrow c_x \cup c_y$ 
12:   $C \leftarrow C - c_x - c_y + c_z$ 
13:   $N_{c_z} \leftarrow \{c_k | v_m \in c_z, v_n \in c_k, e_{mn} \in G(E)\}$ 
14:  Para  $c_k \in N_{c_z}$  Faça
15:     $\Delta Q_{c_k, c_z} \leftarrow Q(G, C - c_k - c_z + (c_k \cup c_z)) - Q(G, C)$ 
16:  fim do loop
17: fim do Enquanto

```

FIG. 3.1: Pseudo-código para o algoritmo de CNM.

Assim, o algoritmo determina os pares de comunidades com o máximo global de ΔQ (linha 7), calculado por $\frac{1}{2m}(A_{ij} - \frac{k_i k_j}{2m})$, onde A_{ij} é o número de arestas entre o par

de comunidades e k_i é o número de arestas saindo da comunidade i , e une as duas comunidades determinadas em uma única comunidade (linhas 11 – 12). O algoritmo repete este procedimento até ΔQ não ser mais positivo (linhas 8 – 10). Durante o processo de **união** de duas comunidades em uma, o algoritmo atualiza o valor de ΔQ que corresponde às comunidades vizinhas da comunidade recém-fundida (linhas 13 – 16).

Se houver empates para o máximo de ΔQ na linha 7, qualquer par de comunidades pode ser selecionado. Na maioria dos casos, a ordem das entradas dos vértices para o algoritmo influencia o resultado. O problema da consistência no resultado final surge aqui e apenas piora, assim como o tamanho da rede aumenta e mais empates surgem.

A Figura 3.2 mostra uma bipartição da Rede de Zachary da Figura 2.2 obtida a partir do algoritmo de Girvan e Newman ([5]). Na mesma, as comunidades são representadas a partir de um dendograma, no qual, o formato do vértice (círculo ou quadrado) representa os vértices que ficaram em cada grupo após a divisão real". Já a Figura 3.3 mostra uma bipartição da mesma rede obtida pelo algoritmo de Clauset, Newman e Moore ([3]).

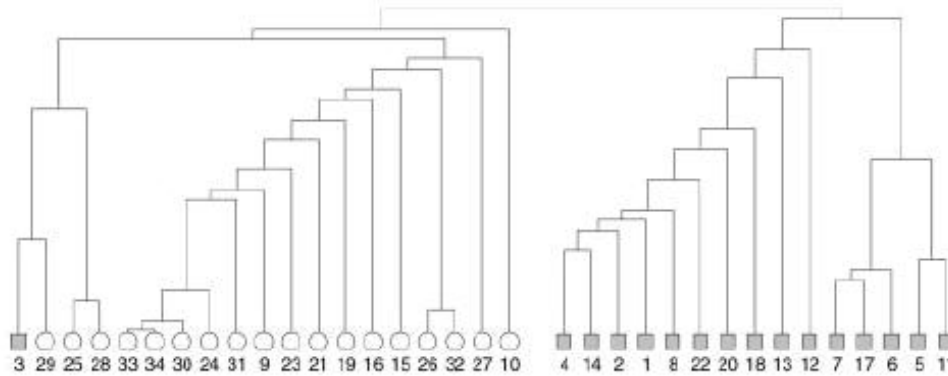


FIG. 3.2: Bipartição da Rede de Zachary obtida por Girvan e Newman (Figura extraída de [5])

A Figura 3.3 mostra a bipartição da mesma Rede de Zachary, porém utilizando agora o algoritmo CNM ([5])

3.2.1 UM EXEMPLO PARA O ALGORITMO CNM

Será feito agora o passo a passo do algoritmo CNM, considerando como entrada do algoritmo o grafo $G=(V,E)$ da Figura 3.4 com $n=6$ vértices e $m=7$ arestas.

A Figura 3.4 mostra cada vértice colorido com uma cor diferente, indicando comunidades diferentes no conjunto C . O algoritmo inicializa $C = \{ \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\} \}$

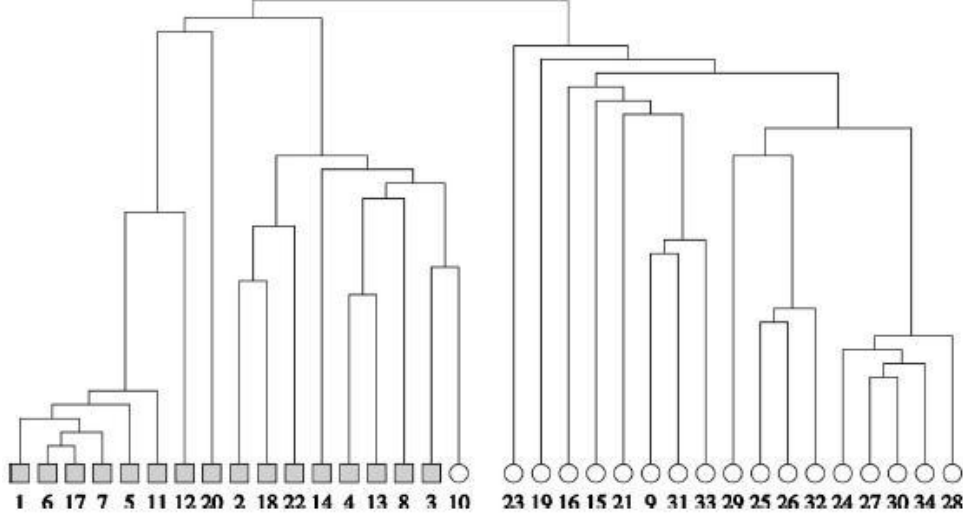


FIG. 3.3: Bipartição da Rede de Zachary obtida por Clauset, Newman e Moore (Figura extraída de [5])

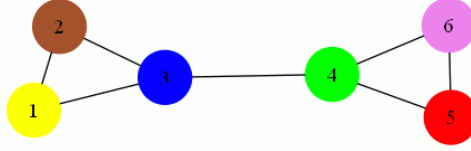


FIG. 3.4: Grafo exemplo com 6 comunidades

e o heap H com os valores:

- $\Delta Q_{1,2} = \Delta Q_{5,6} = 1/14(1 - 1.1/14) = 5/14.7;$
- $\Delta Q_{3,4} = 1/14(1 - 3.3)/14 = 5/14.14;$
- $\Delta Q_{2,3} = \Delta Q_{1,3} = \Delta Q_{4,6} = \Delta Q_{4,5} = 1/14(1 - 2.3/14) = 4/14.7;$

Na primeira iteração:

$$H \neq \emptyset$$

$$max = 5/14.7$$

$$C_1 = C_1 \cup C_2$$

$$C = \{C_1 = \{1, 2\}, C_3 = \{3\}, C_4 = \{4\}, C_5 = \{5\}, C_6 = \{6\}\}$$

No caso do exemplo, a Figura 3.5 mostra as comunidades unitarias $\{1\}$ e $\{2\}$ que são unidas formando a nova comunidade $\{1, 2\}$, cujos vértices são mostrados coloridos amarelo. Uma vez atualizadas as comunidades e o heap, continua a execução do algoritmo com o novo conjunto de comunidades $C = \{\{1, 2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$ e o heap atualizado:

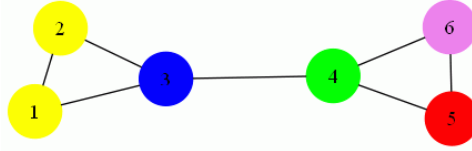


FIG. 3.5: Grafo exemplo com 5 comunidades

- $\Delta Q_{1,3} = 11/14.7$
- $\Delta Q_{3,4} = 5/14.14$
- $\Delta Q_{4,5} = \Delta Q_{4,6} = 8/14.14$
- $\Delta Q_{5,6} = 5/14.7$

Na segunda iteração:

$$H \neq \emptyset$$

$$max = 11/14.7$$

$$C_1 = C_1 \cup C_3$$

$$C = \{C_1 = \{1, 2, 3\}, C_4 = \{4\}, C_5 = \{5\}, C_6 = \{6\}\}$$

A Figura 3.6 mostra essa atualização de comunidades.

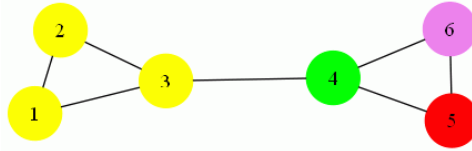


FIG. 3.6: Grafo exemplo com 4 comunidades

O algoritmo une os elementos do par de comunidades $C_x = \{1, 2\}$ e $C_y = \{3\}$, formando $C_z = \{1, 2, 3\}$, atualiza C e o heap:

- $\Delta Q_{1,4} = 11/14.14$
- $\Delta Q_{4,5} = \Delta Q_{4,6} = 4/14.7$
- $\Delta Q_{5,6} = 5/14.7$

Na terceira iteração:

$$H \neq \emptyset$$

$$max = 5/14.7$$

$$C_5 = C_5 \cup C_6$$

$$C = \{C_1 = \{1, 2, 3\}, C_4 = \{4\}, C_5 = \{5, 6\}\}$$

A Figura 3.7 mostra essa atualização de comunidades.

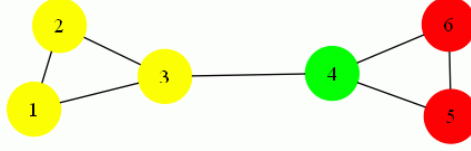


FIG. 3.7: Grafo exemplo com 3 comunidades

O algoritmo une os elementos de $C_x = \{5\}$ e $C_y = \{6\}$, formando $C_z = \{5, 6\}$ e atualiza o conjunto de comunidades C e o heap:

- $\Delta Q_{1,4} = 11/14.14$
- $\Delta Q_{4,5} = 27/14.14$

Na quarta e última iteração:

$$H \neq \emptyset$$

$$max = 27/14.14$$

$$C_4 = C_4 \cup C_5$$

$$C = \{C_1 = \{1, 2, 3\}, C_4 = \{4, 5, 6\}\}$$

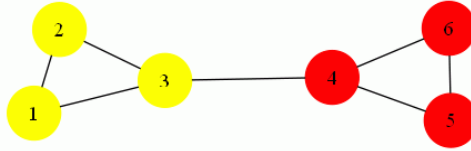


FIG. 3.8: Grafo exemplo com 2 comunidades

Após a formação de $C_z = \{4, 5, 6\}$ a partir de $C_x = \{5, 6\}$ e $C_y = \{4\}$, o conjunto C é atualizado e também o heap. O resultado pode ser visto na Figura 3.8.

Assim o heap indica que não há mais alterações a serem feitas e o conjunto de comunidades fica $C = \{\{1, 2, 3\}, \{4, 5, 6\}\}$.

3.3 O ALGORITMO DE WAKITA

Enquanto o algoritmo CNM melhora o algoritmo guloso utilizando uma estrutura de dados sofisticada, é observado que produz crescimento desequilibrado em algumas

comunidades. Ou seja, o par de comunidades unidas primeiro tende a aumentar até formar uma grande comunidade, sempre que permita aumentar a modularidade.

Quando não é mais possível aumentar o tamanho dessa comunidade, outro par de nós simples é formada e então se repete o mesmo processo até formar grandes comunidades. Dessa forma, poucas comunidades grandes surgem e então os nós restantes unem-se em comunidades menores.

Wakita e Tsurumi usaram uma taxa de consolidação que leva em consideração o tamanho das comunidades e induz a um crescimento equilibrado entre as comunidades. Isto reduz a complexidade, o computador pode trabalhar com mais nós de um forma mais rápida, e permite trabalhar com redes de até 10 milhões de nós. A Figura 3.1 apresenta a linha 7 do pseudo-código do algoritmo de Wakita, que é bastante similar ao algoritmo do CNM, apenas diferindo nesta linha:

$$\text{extract } \max \Delta Q_{c_x, c_y} . \min(|c_x|/|c_y|, |c_y|/|c_x|) \quad (3.1)$$

A expressão 3.1 calcula não apenas ΔQ mas também uma razão de consolidação, e com isso comunidades de tamanhos similares são preferidas para serem unidas. Portanto, como no algoritmo do CNM, existem empates para o máximo de ΔQ em 3.1, e qualquer par de comunidades pode ser selecionado em caso de empates. O problema da consistência persiste neste novo algoritmo. O algoritmo de Wakita se comporta de maneira similar ao CNM, por isso, o exemplo de funcionamento dado na seção anterior pode ser usado também para esse algoritmo.

3.4 O ALGORITMO DE BLONDEL

Blondel et al. [2] considera também a idéia da consolidação e une, em cada iteração, todos os nós e comunidades que maximizam ΔQ . O algoritmo proposto por [2] é um método heurístico baseado em otimização da modularidade de um grafo, unindo a cada iteração as comunidades de modo a maximizar o ΔQ . Sua aplicação é focada em redes extensas devido ao seu rápido tempo de computação, sendo o único algoritmo conhecido que foi aplicado em redes de mais de 10 milhões de nós. A qualidade das comunidades detectadas por Blondel é maior do que outros algoritmos famosos como o CNM, além de ser de simples entendimento e implementação.

O método proposto no artigo encontra partições com modularidade alta em tempo

curto e fornece uma estrutura hierárquica para as comunidades, isso permite visualizar diferentes etapas da detecção de comunidades. O algoritmo é dividido em duas fases repetidas iterativamente e pode ser usado em grafos com ou sem pesos nas arestas.

Primeiramente, cada nó é uma comunidade, ou seja, na partição inicial existem tantas comunidades quanto for a quantidade de nós. Então, para cada nó i tomamos os vizinhos j e encontramos o ganho de modularidade gerado por colocar o nó i na comunidade do nó j . O nó i é então transferido para a comunidade em que o ganho é máximo, mas apenas se o ganho for positivo. Se nenhum dos ganhos foi positivo, i permanece na sua comunidade. Esse processo é aplicado repetidamente e sequencialmente para todos os nós até que não seja possível mais nenhum ganho, então a primeira fase termina.

Observe que um nó pode ser considerado várias vezes. A primeira fase termina quando nenhum movimento de um nó pode melhorar a modularidade. Além disso, a ordem em que os nós são considerados também influi na saída. Resultados preliminares, no entanto, indicam que a influência da ordenação na modularidade não é grande. Contudo, ela pode influenciar no tempo de computação.

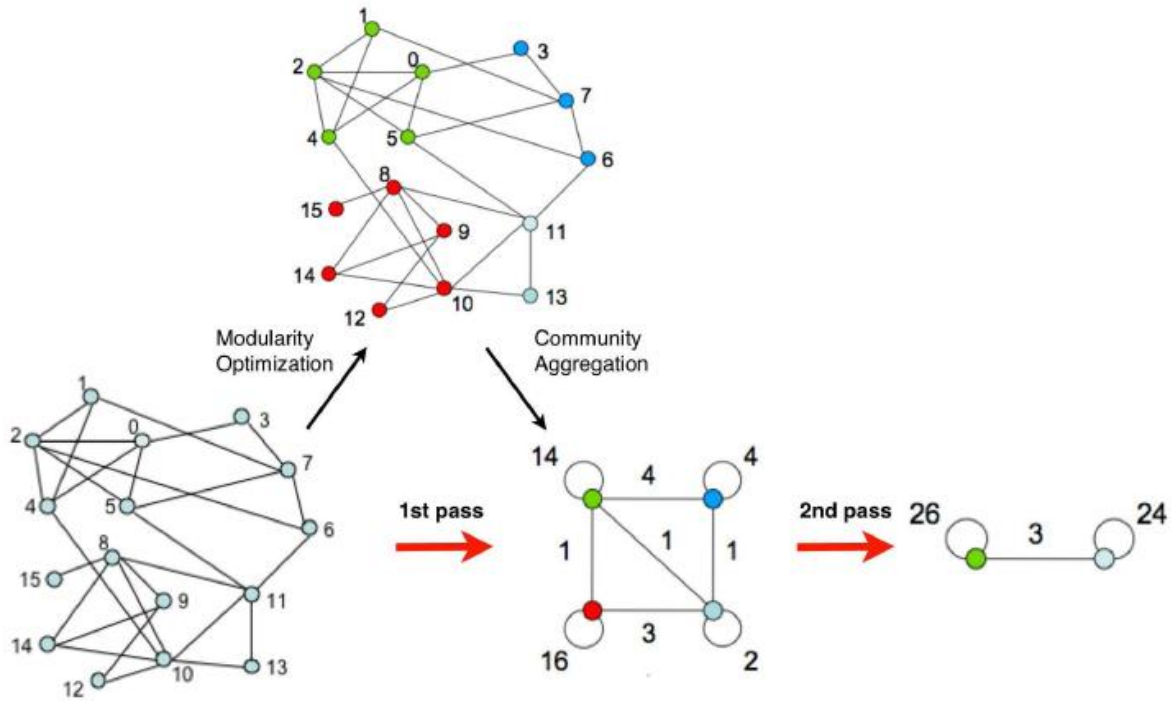


FIG. 3.9: Exemplo de aplicação do algoritmo de Blondel (Figura extraída de [2])

A segunda fase consiste na construção de um novo grafo onde os nós são as comunidades formadas na etapa anterior. Para isso, o peso das arestas entre os novos nós são dados

pela soma das arestas entre nós das duas comunidades. Arestas entre nós na mesma comunidade geram loops. Feito isso, basta aplicar a primeira fase novamente no grafo resultante. Essas duas fases combinadas são chamadas de "passo".

A Figura 3.9 mostra um exemplo de um grafo cujas comunidades são construídas usando o algoritmo de Blondel. Cada passo tem duas fases: uma onde a modularidade é otimizada apenas por meio de mudanças locais de comunidades; outra onde as comunidades encontradas são agregadas a fim de construir uma nova rede de comunidades. Os passos são repetidos iterativamente até que nenhum aumento da modularidade é possível.

```

1:  $V$ : conjunto de vértices
2:  $E$ : conjunto de arestas
3:  $W$ : conjunto de arestas
4:  $G \leftarrow (V, E, W)$ 
5: repita
6:    $C \leftarrow \{\{v_i\} | v_i \in G(V)\}$ 
7:   Calcular a modularidade atual  $Q_{at}$ 
8:    $Q_{new} \leftarrow Q_{at}$ 
9:    $Q_{old} \leftarrow Q_{new}$ 
10:  repita
11:    para cada  $v_i \in V$  faça
12:       $Q_{at} \leftarrow Q_{new}$ 
13:      remova  $v_i$  da sua comunidade atual
14:       $N_{v_i} \leftarrow \{c_k | v_i \in G(V), v_j \in c_k, e_{ij} \in G(E)\}$ 
15:      encontre  $c_x \in N_{v_i}$  que tem  $\max \Delta Q_{v_i, c_x} > 0$ 
16:      coloque  $v_i$  em  $c_x$ 
17:    fim do para
18:    calcule a nova modularidade  $Q_{new}$ 
19:  até que nenhuma comunidade mude ou  $Q_{new} = Q_{at}$ 
20:   $V' \leftarrow \{c_i | c_i \in C\}$ 
21:   $E' \leftarrow \{e_{ij} | \forall e_{ij} \text{ se } v_i \in C_i, v_j \in C_j \text{ e } C_i \neq C_j\}$ 
22:   $W' \leftarrow \{w_{ij} | \sum w_{ij}, \forall e_{ij} \text{ se } v_i \in C_i \text{ e } v_j \in C_j\}$ 
23:   $G \leftarrow (V', E', W')$ 
24: até que  $Q_{new} = Q_{old}$ 

```

FIG. 3.10: Pseudo-código para o algoritmo de Blondel.

O número de comunidades decresce a cada passo, uma consequência disso é que o tempo de computação é gasto, em sua maioria, no primeiro passo. É feita a iteração de passos até que não ocorra mais nenhuma mudança e a modularidade máxima é alcançada. O algoritmo incorpora a noção de hierarquia, já que são construídas comunidades de comunidades. A altura da hierarquia é dada pelo numero de passos e é, geralmente,

pequena.

A Figura 3.10 apresenta o pseudo-código para o algoritmo de Blondel, que consiste de duas fases:

4 ANÁLISE DA REDE DOS ALUNOS DO SEGUNDO ANO DO IME

Neste capítulo, será apresentado o trabalho de construção de grafos de relacionamento baseados em dados reais de alunos do 2º ano de 2011 e a obtenção das comunidades a partir da aplicação dos métodos de CNM, utilizando o software em desenvolvimento.

Posteriormente serão montados dois tipos de grafos: O primeiro grafo, denominado **Grafo Questionário**, será montado com base em um questionário de três perguntas; o segundo grafo, denominado **Grafo Seção**, será montado de acordo com a ficha de escolha de engenharia preenchida pelos alunos no final do segundo ano.

Após a montagem dos grafos, o software será utilizado tanto para a divisão dos grafos em comunidades e quanto para montar a visualização dos mesmos.

As subseções 4.1, 4.2 abordam como foi feita a montagem de cada grafo, assim como as comunidades geradas pela aplicação do algoritmo CNM. A subseção 4.3 compara os resultados obtidos pelo algoritmo CNM para os dois grafos escolhidos, com a divisão em comunidades obtida da escolha real de especialidade e a comparação dos grupos de alunos que forma o terceiro ano de 2012.

4.1 GRAFO QUESTIONÁRIO

O Grafo Questionário foi construído a partir das informações obtidas através de uma pesquisa realizada com o segundo ano no fim do ano de 2011, antes da realização de escolha de especialidades. O formulário visa obter informações para montagem do grafo com o objetivo de detectar comunidades que determinasse uma previsão das escolhas de especialidade de cada aluno. As perguntas realizadas foram:

- a) Quais são os colegas do 2º ano com os quais você mais estuda para as provas e trabalhos das disciplinas do IME?
- b) Dentre as disciplinas do 1º e 2º ano básicos, quais são as 3 que você mais se identifica?
- c) Quais são os amigos dentre os colegas do 2º ano que você divide seu tempo livre (atividades esportivas, culturais e sociais, cursos fora do IME e outras)?

Com as respostas em mãos, foi realizada a montagem do grafo. Inicialmente especificamos os nós como sendo o conjunto de alunos mais o conjunto de disciplinas incluídas no formulário pelos alunos estudadas durante o período do curso básico. A partir de então são estabelecidas as arestas de acordo com as respostas dadas por cada um nas perguntas realizadas. Cada vértice do grafo representa um aluno ou uma disciplina. Então, se o aluno A e a disciplina B foram citadas no questionário do aluno C , existe uma aresta ligando A e C e outra ligando B e C .

Nessa fase, problemas foram detectados devido à ausência de alguns alunos no dia da pesquisa e, conseqüentemente, ausência de suas fichas indicando sua preferência. Além disso, alguns alunos deixaram em branco ou indicaram que não tinham preferência em alguma pergunta e, em outros casos, mesmo com a condição de no máximo 3 respostas por pergunta, mais respostas do que o condicionado inicialmente foram dadas. Esses fatos determinaram a decisão de escolher entre duas situações: eliminar respostas ou considerar todas. Foi adotado o segundo método, deixando alguns alunos com mais arestas do que outros.

A Figura 4.1 mostra o Grafo Questionário e as comunidades detectadas utilizando o algoritmo de Clauset-Newman-Moore. Os vértices foram numerados de 1 até 105 para os alunos e de 106 até 120 para as disciplinas informadas.

Esperava-se que alunos que são da mesma turma ficassem agrupados devido a uma convivência maior no dia-a-dia, visto que isso implica que as respostas dadas, principalmente para a primeira pergunta, sejam nomes de colegas da mesma turma. Pelo mesmo motivo e devido a uma falta de convivência entre os oficiais do CG e os alunos do CFG, espera-se também uma separação do CG em uma comunidade.

O que interessa aqui é como os alunos se relacionam, as comunidades sociais em que cada um se insere, bem como as disciplinas que mais têm a ver com cada comunidade, e como essa preferência de disciplina reflete nas escolhas de especialidade e nas preferências de engenharia de cada um.

Com o objetivo de reduzir a quantidade de nós do Grafo Questionário e a fim de obter um grafo com um significado mais amplo em termos de disciplinas e especialidades, foram feitas as seguintes modificações:

- As disciplinas Introdução a Computação, Cálculo Numérico, Álgebra Linear, Cálculo e Probabilidade e Estatística foram consideradas um mesmo nó, número 106.

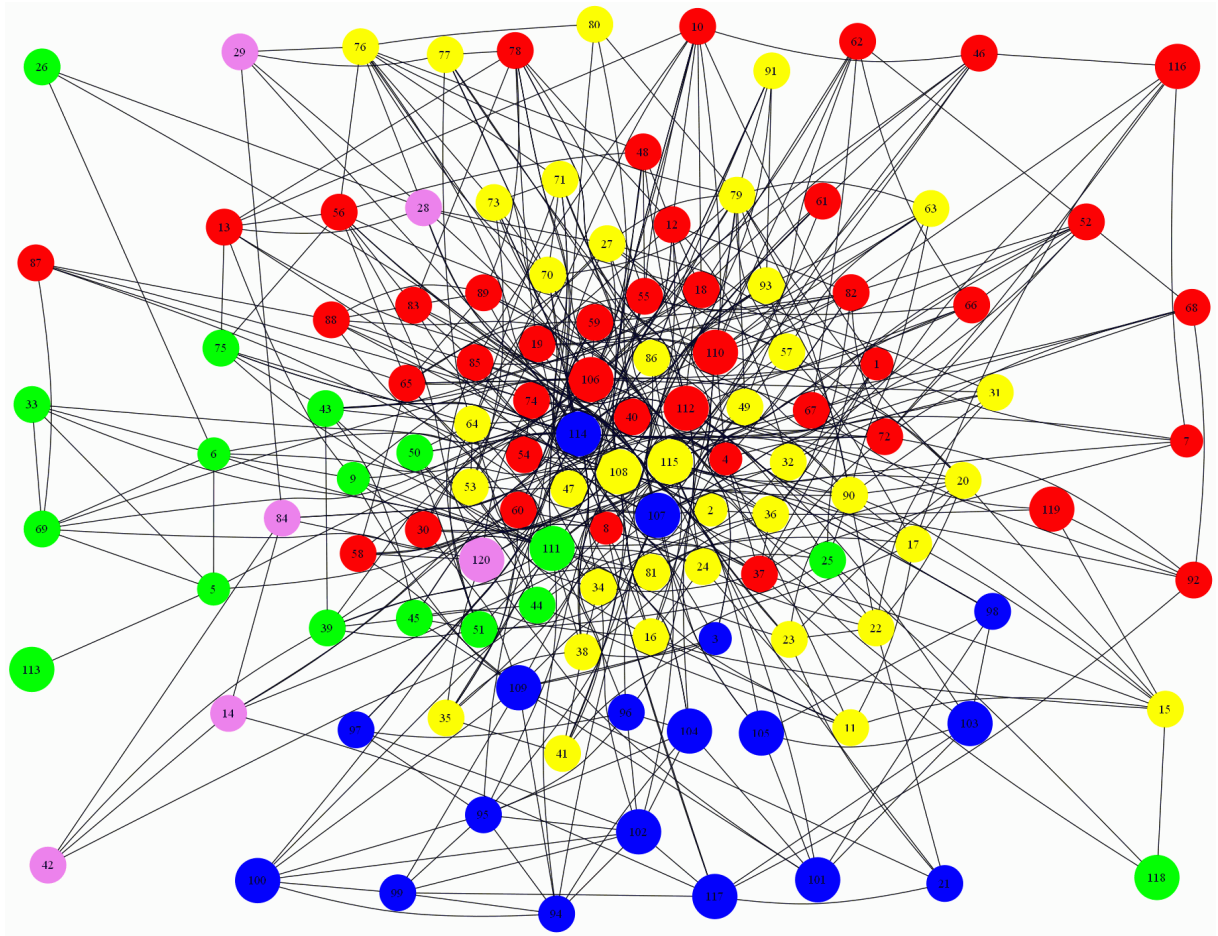


FIG. 4.1: Grafo Questionário e as 5 comunidades detectadas pelo algoritmo CNM

- As disciplinas Eletricidade Técnica, Física e Laboratório de Física foram consideradas um mesmo nó, com número 107.
- As disciplinas Resistência dos Materiais, Mecânica e Desenho Técnico foram consideradas um mesmo nó, com numeração 108.
- As disciplinas Introdução a Ciência dos Materiais e Química foram consideradas um mesmo nó, com numeração 109.
- As disciplinas Projeto de Inglês e Redação Científica foram consideradas um mesmo nó, com numeração 110.

O grafo resultante será denominado grafo Questionário Reduzido e é mostrado na figura 4.2. Assim, espera-se que alunos pertencentes a mesma comunidade que o nó 106 tenham um interesse maior pela $SE/2$, $SE/6$ ou $SE/8$. Assim como alunos que pertencem

a mesma comunidade que o nó 107 tenham interesse pela $SE/3$, alunos que pertencem a mesma comunidade que o nó 108 mostrem interesse pela $SE/4$, alunos que pertencem a mesma comunidade que o nó 109 tenham interesse pela $SE/5$ e alunos que pertencem a mesma comunidade que o nó 110 serão considerados não tendo uma intenção bem definida na escolha de sua engenharia.

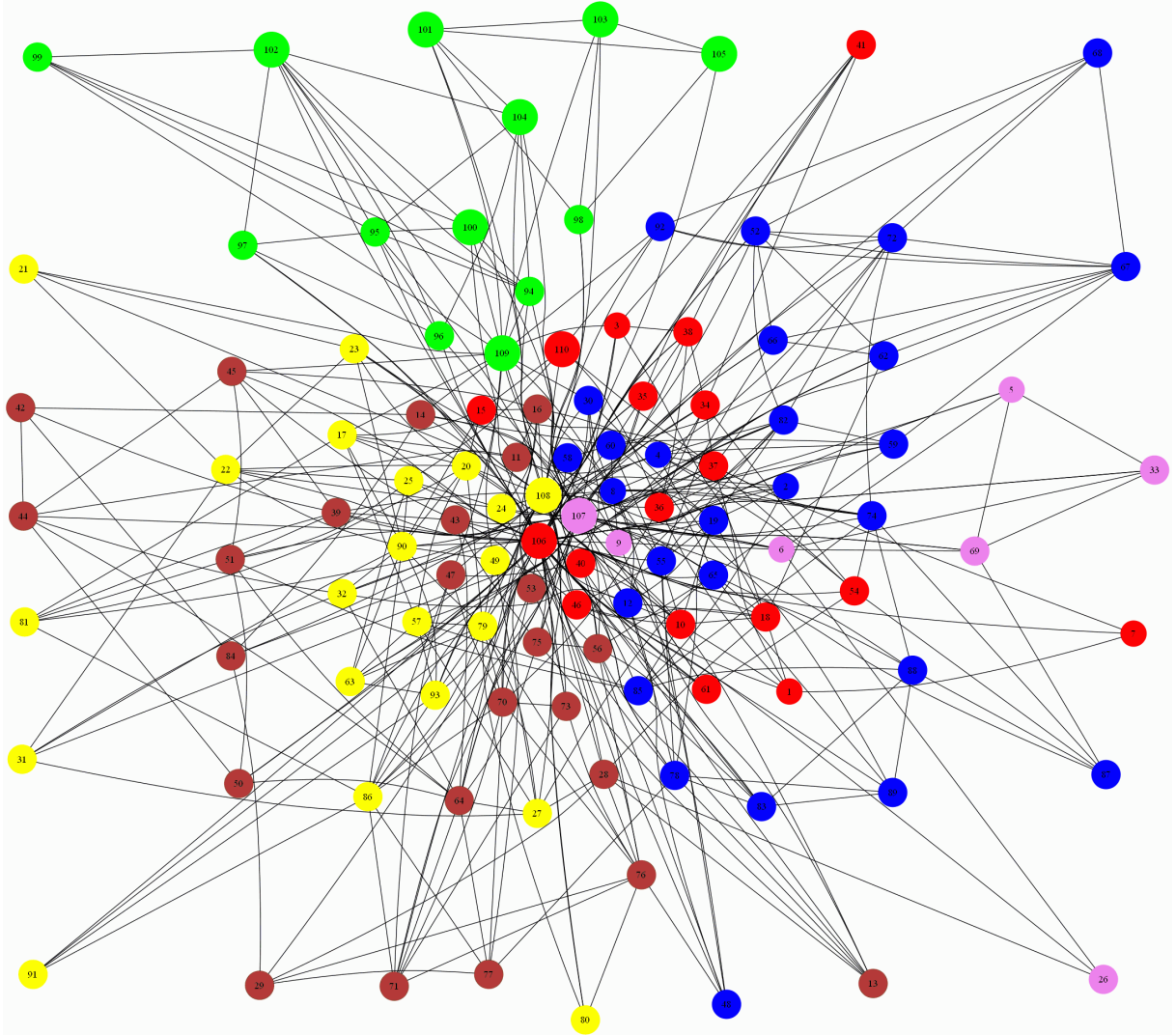


FIG. 4.2: Grafo Questionário Reduzido e as 6 comunidades detectadas pelo algoritmo CNM

4.2 GRAFO SEÇÃO

Numa segunda fase, após as escolhas de especialidades dos alunos, resolveu-se montar um novo grafo tendo os alunos e as engenharias de cada seção do IME como nós. Cada

aluno preencheu uma ficha no final do 2º ano (dezembro de 2011), onde ordena a prioridade que tem para escolher cada especialidade de engenharia e que deve ser consultada caso o mesmo falte no dia da escolha.

Construiu-se um grafo onde cada vértice representa um aluno ou uma das três especialidades citadas na ficha de escolha. Assim, se o aluno A citou em sua ficha, em ordem de preferência, as engenharias E_1 , E_2 e E_3 , então existem as arestas AE_1 , AE_2 e AE_3 no grafo Seção. A intenção desse grafo é realizar uma comparação das comunidades obtidas com o primeiro grafo, a fim de analisar a relação entre as comunidades baseadas em disciplinas e amizades entre os alunos e as preferências de especialidade.

Nessa fase, houve problemas devido a ausência da ficha de alguns alunos. Isso exclui a possibilidade de comparação, para esses alunos em particular, entre as escolhas e as comunidades sociais e como influíram na sua vida acadêmica. Existem 7 alunos do segundo ano que não preencheram a ficha de escolha de especialidade.

Pode ser que haja predominância de alguma turma em uma comunidade, embora não seja o normal, pois não necessariamente uma turma tem os mesmos interesses quando se trata de especialidade. Seria bom observar uma certa coincidência de comunidades em relação ao primeiro grafo, pois espera-se que alunos com as mesmas preferências de matérias e que estudam e compartilham seu tempo juntos tenham os mesmos interesses, ou que sejam pelo menos parecidos, quando se fala da engenharia pretendida.

O grafo obtido está na Figura 4.3. Da mesma forma como se fez na Seção 4.2, será reduzido o número de nós do Grafo Seção, de modo a obter menos comunidades e identificar as comunidades por seção e não mais por engenharia. Assim,

- Uniu-se as Engenharias Elétrica, Comunicações e Eletrônica no mesmo nó 107.
- Uniu-se as Engenharias de Materias, Mecânica de Automóveis e Mecânica de Armamento no nó 108.

A Figura 4.4 mostra o grafo reduzido resultante da ficha de escolha de engenharia e as comunidades obtidas pelo algoritmo CNM. Observa-se que os nós 94 a 105 (correspondente a alunos da turma do CG) não se encontram agrupados na mesma comunidade como no grafo anterior. Além disso, observa-se uma redução no número de comunidades.

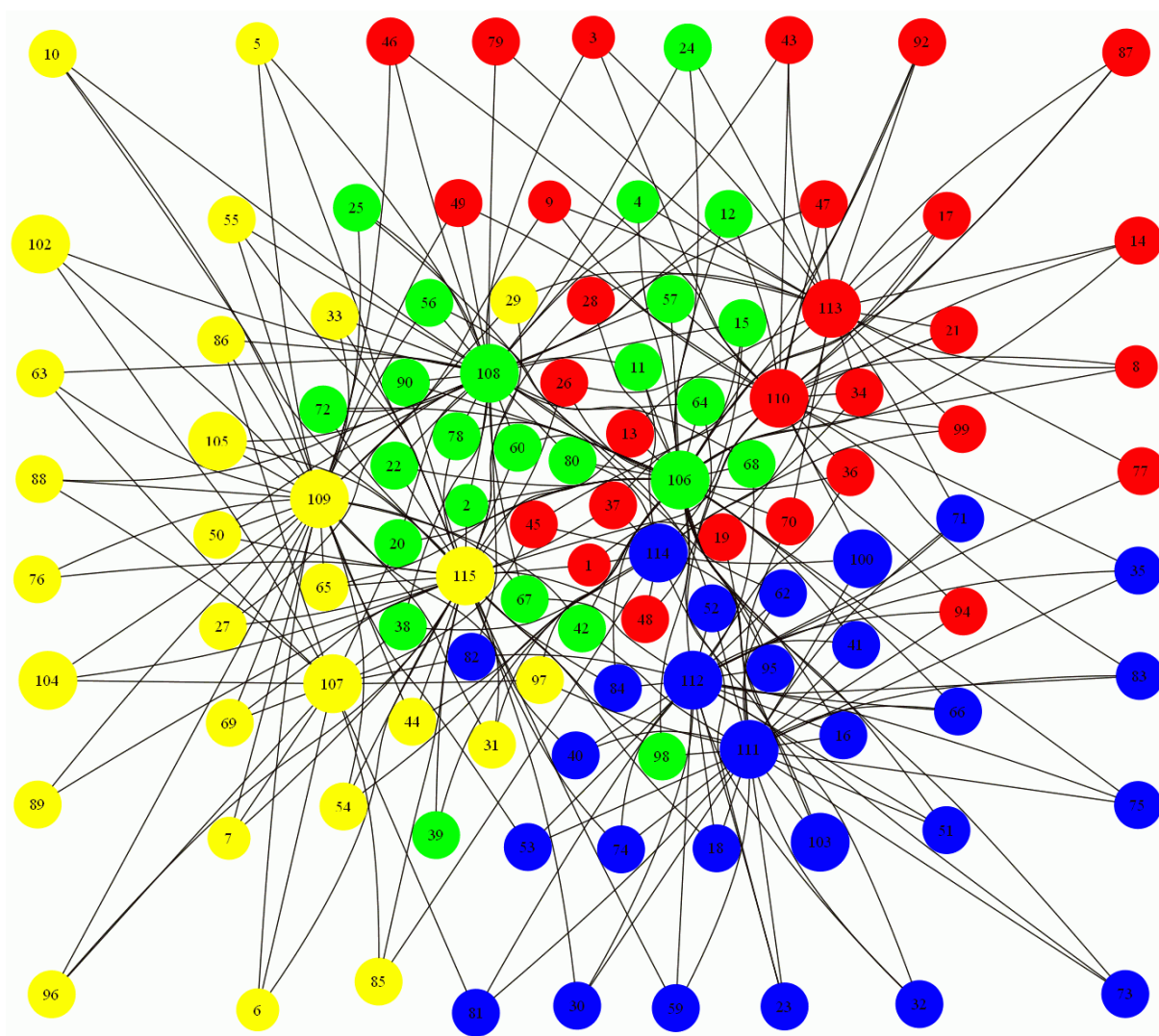


FIG. 4.3: Grafo Seção e as 4 comunidades detectadas pelo algoritmo CNM

4.3 COMPARAÇÃO DAS COMUNIDADES OBTIDAS POR CNM COM DADOS REAIS DO 3º ANO DE 2012

Nesta seção será apresentado uma análise dos resultados obtidos pelo algoritmo CNM (as comunidades representadas nas Figuras 4.2 e 4.4). A análise será baseada na comparação destes resultados com a distribuição dos alunos do 3º ano dos cursos de graduação em Engenharia do IME por seção de ensino.

O conteúdo desta seção está dividido em duas subseções. Na primeira delas será usado o **Grafo Questionário** reduzido. Na segunda será utilizado o **Grafo Seção** reduzido.

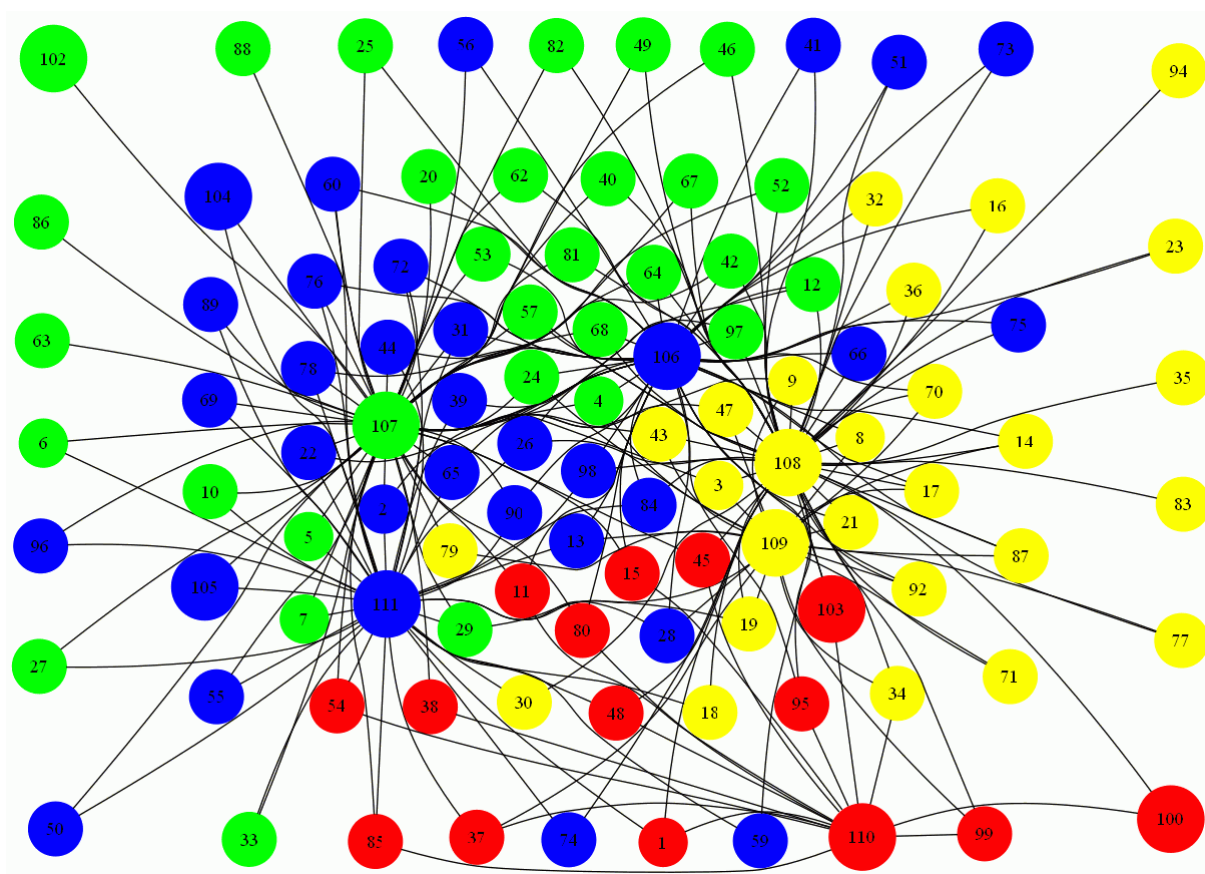


FIG. 4.4: Grafo Seção Reduzido e as 4 comunidades detectadas pela algoritmo CNM

4.3.1 COMPARAÇÃO COM O GRAFO QUESTIONÁRIO REDUZIDO

A Figura 4.5 representa a divisão em comunidades obtida da escolha dos alunos do 3º ano no ano 2012 sobre o **Grafo Seção** reduzido, na qual os vértices são coloridos segundo a distribuição do 3º ano nas seções de ensino do IME no início de 2012. Essa distribuição dos alunos foi realizada na cerimônia de escolha de especialidade em fevereiro de 2012 e do resultado dessa atividade surgiram as turmas de cada seção.

As cores utilizadas na Figura 4.5 representam:

- A cor rosa representa a SE/2.
- A cor verde representa a SE/3.
- A cor amarelo representa a SE/4.
- A cor marrom representa a SE/5.
- A cor vermelho representa a SE/6.

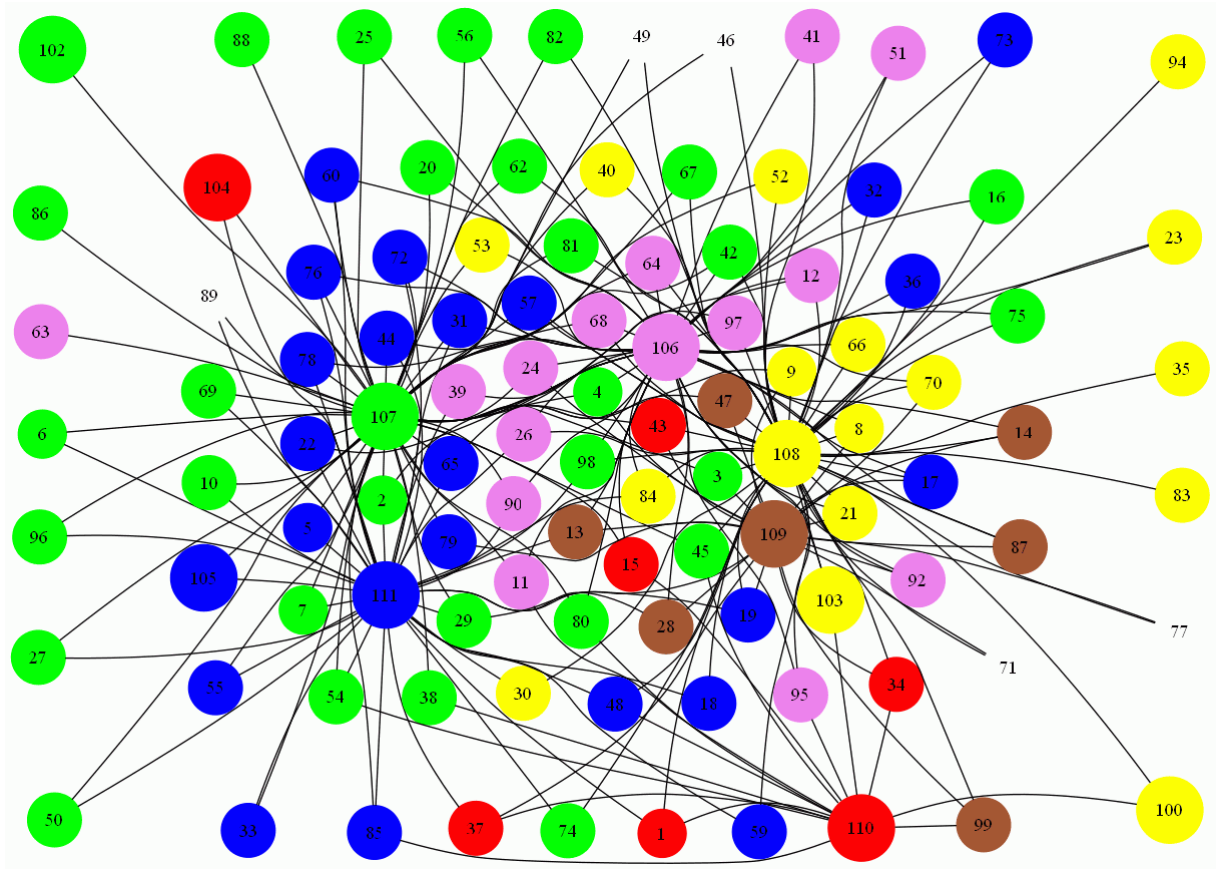


FIG. 4.5: Grafo Seção reduzido e comunidades definidas pela escolha de especialidade em fevereiro de 2012 dos atuais alunos do 3º ano

- A cor azul representa a SE/8.
- A cor branco representa os alunos que trancaram.

Se compararmos as comunidades obtidas no **Grafo Seção** reduzido, podemos observar uma interseção das comunidades obtidas pelo algoritmo CNM e pelos dados reais de escolha de especialidade. Essa observação é facilitada pela coloração adotada que está bastante similar nos dois grafos apesar do significado um pouco diferente, visto que o **Grafo Seção** reduzido com comunidades obtidas pelo algoritmo CNM agrupa algumas seções em uma mesma comunidade, como é o caso das SE/2 e SE/8 (na comunidade azul) e das SE/4 e SE/5 (na comunidade amarela) da Figura 4.4.

A SE/2 (representada pelo nó 106) faz parte da comunidade azul no **Grafo Seção** reduzido(Figura 4.4), sendo sua cor na Figura 4.5 o rosa. Houve 5 nós coincidentes entre essas comunidades e o total de alunos da SE/2 em fevereiro de 2012 é 14, ou seja, um pouco mais de 1/3 dos alunos foram alocados na comunidade correta pelo algoritmo CNM.

A SE/3 (representada pelo nó 107) é verde nas duas figuras (4.4 e 4.5). Uma coincidência de 16 nós entre as comunidades determinadas pelo algoritmo de CNM e os dados reais, sendo 30 o total de alunos na SE/3, ou seja, um pouco mais da metade foram corretamente alocados pelo algoritmo.

A SE/4 (representada pelo nó 108) é amarela nas duas figuras e são 9 os nós coincidentes, sendo o total de 16 alunos na seção; novamente, mais da metade dos alunos foram corretamente atribuídos a suas especialidades de engenharia. A SE/5 (representada pelo nó 109) é amarela na Figura 4.4 e marrom na Figura 4.5, nas quais 3 nós coincidem de um total de 6 alunos da seção, ou seja, exatamente a metade.

A SE/6 (representada pelo nó 110) tem 6 alunos e está colorida de vermelho na Figura 4.5. Na Figura 4.4, 3 desses nós estavam posicionados corretamente na comunidade correspondente, ou seja, novamente metade de acertos. Por fim, a SE/8 (representada pelo nó 111), com 23 alunos na Figura 4.5, é azul nas duas divisões e apresenta um total de 12 coincidências, ou pouco mais da metade dos nós.

Pode-se observar que metade dos nós foram corretamente atribuídos em cada SE pelo algoritmo CNM aplicado ao **Grafo Seção** reduzido, o que representa um resultado bom. Os erros podem ter sido causados pela falta de peso nas arestas, pois a lista de preferência de especialidade fornecida por cada aluno é de prioridade. Mais um fator que pode ter prejudicado nessa estimativa é o número limitado de vagas para os militares que resulta em alunos indo para SE's que não foram escolhidas na ficha ou que não aconteceria no caso da escolha ser livre.

Comparando com as comunidades obtidas no **Grafo Questionário** reduzido, as interseções não são tão claras como foram para o **Grafo Seção** reduzido. Esse resultado não foi satisfatório e se espera poder melhorar os acertos com algumas mudanças.

A primeira delas seria fazer uma divisão das disciplinas em áreas mais condizente com as comunidades observadas no **Grafo Questionário**. Um exemplo disso é a área 1 que no **Grafo Questionário** reduzido engloba as disciplinas Introdução a Computação, Cálculo Numérico, Probabilidade e Estatística, Cálculo e Álgebra Linear e poderia contribuir para uma divisão em comunidades mais condizente se dividida em duas áreas, uma com as disciplinas Introdução a Computação e Cálculo Numérico, e a outra com Probabilidade e Estatística, Cálculo e Álgebra Linear.

Outro fator observado na divisão em comunidades do **Grafo Questionário** pelo algoritmo CNM foi uma separação entre as turmas, sendo o exemplo mais destacado a

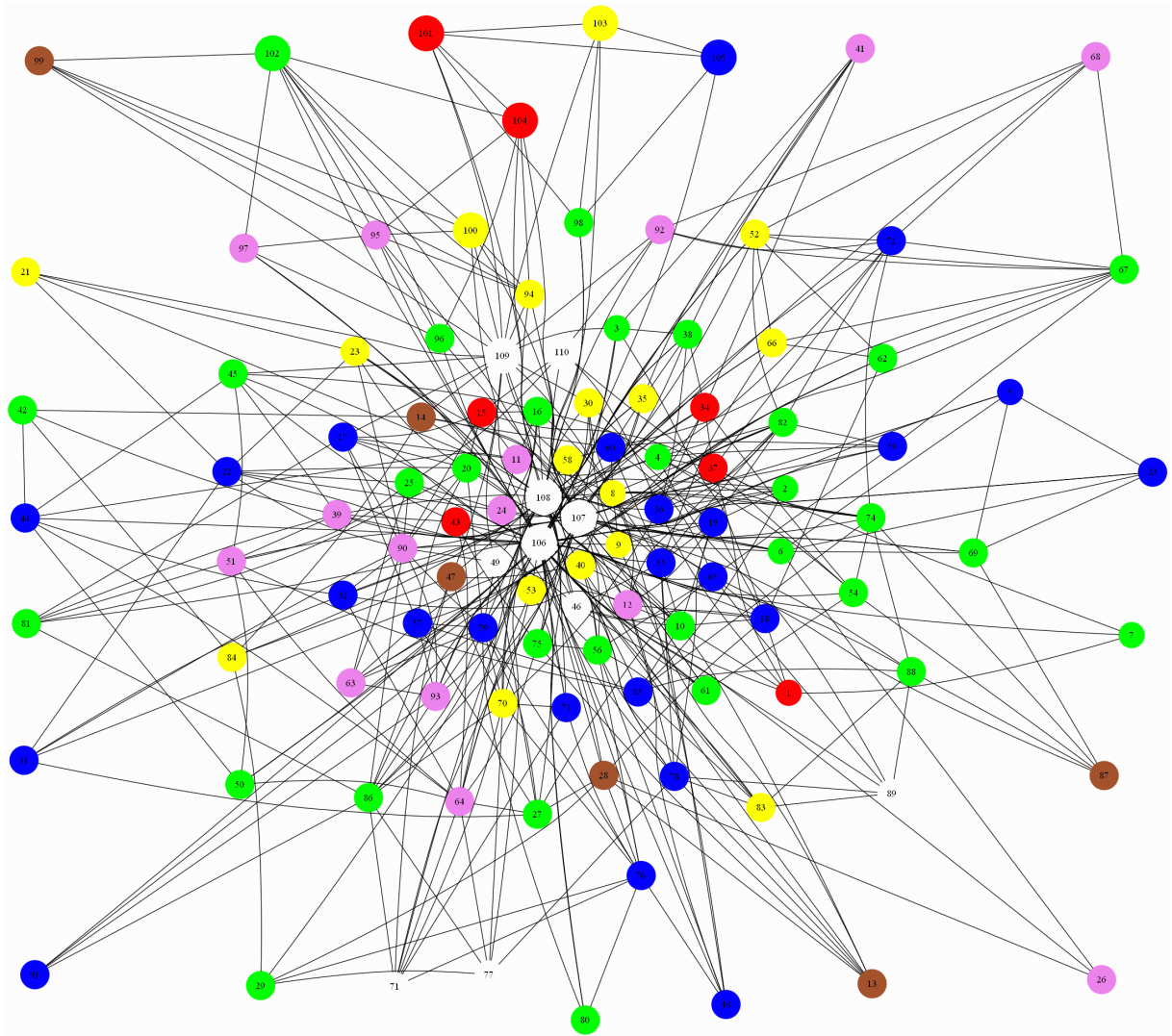


FIG. 4.6: Grafo Questionário reduzido e comunidades definidas pela escolha de especialidade em fevereiro de 2012 dos atuais alunos do 3ºano

turma do CG, que se agrupou toda em uma só comunidade. Isso pode ter decorrido do fato de duas perguntas no questionário envolverem outros alunos na resposta e apenas uma envolver disciplinas, o que ocasionou um número maior de arestas entre alunos do que para disciplinas e consequentemente uma divisão que favoreceu os relacionamentos e desfavoreceu os interesses acadêmicos individuais. Uma solução seria eliminar uma das perguntas sobre alunos, por exemplo a pergunta c) e teríamos um grafo mais equilibrado, com cada aluno indicando 3 outros alunos e 3 disciplinas.

4.4 CONSTRUÇÃO DO NOVO GRAFO QUESTIONÁRIO

Os resultados obtidos com os grafos da seção anterior não foram totalmente satisfatórios. Ainda espera-se encontrar um resultado melhor, mais condizente com uma previsão das escolhas.

A partir de idéias trocadas entre os membros do projeto e da banca de avaliação, foi resolvido que novos grafos deveriam ser montados. As conclusões provenientes das discussões foram:

- a) O grafo questionário deve ser reconstruído com a eliminação da pergunta sobre a convivência pessoal (pergunta c), a fim de equilibrar a quantidade de arestas que relacionam pessoas com pessoas e que relacionam pessoas com disciplinas.
- b) Deve ser realizada uma comparação entre as comunidades formadas no novo grafo questionário e as comunidades do grafo das escolhas de curso.
- c) Esses grafos devem ser construídas por uma ferramenta que torne o processo de construção transparente, de forma que não seja necessário um conhecimento sobre a estrutura do arquivo que representa o grafo para se usar o programa.

A explicação da ferramenta desenvolvida para tornar o processo transparente e de como ela foi projetada se encontra na Seção 5. Nesta seção também se encontra uma breve explicação do funcionamento da ferramenta.

Desenvolvida a ferramenta, um novo grafo questionário foi montado. O questionário foi atualizado e apenas duas perguntas foram consideradas na montagem do grafo:

- a) Quais são os colegas do 2º ano com os quais você mais estuda para as provas e trabalhos das disciplinas do IME?
- b) Dentre as disciplinas do 1º e 2º ano básicos, quais são as 3 que você mais se identifica?

O novo grafo questionário obtido está na Figura 4.7, a divisão em comunidades já foi realizada e se pode observar que elas são sete. Devido aos resultados confusos e não satisfatórios obtidos com o grafo questionário anterior, uma alteração na divisão de disciplinas foi feita. O nó que incluía as disciplinas Introdução à Computação, Cálculo Numérico, Álgebra Linear, Cálculo e Probabilidade e Estatística foi dividido em dois novos nós.

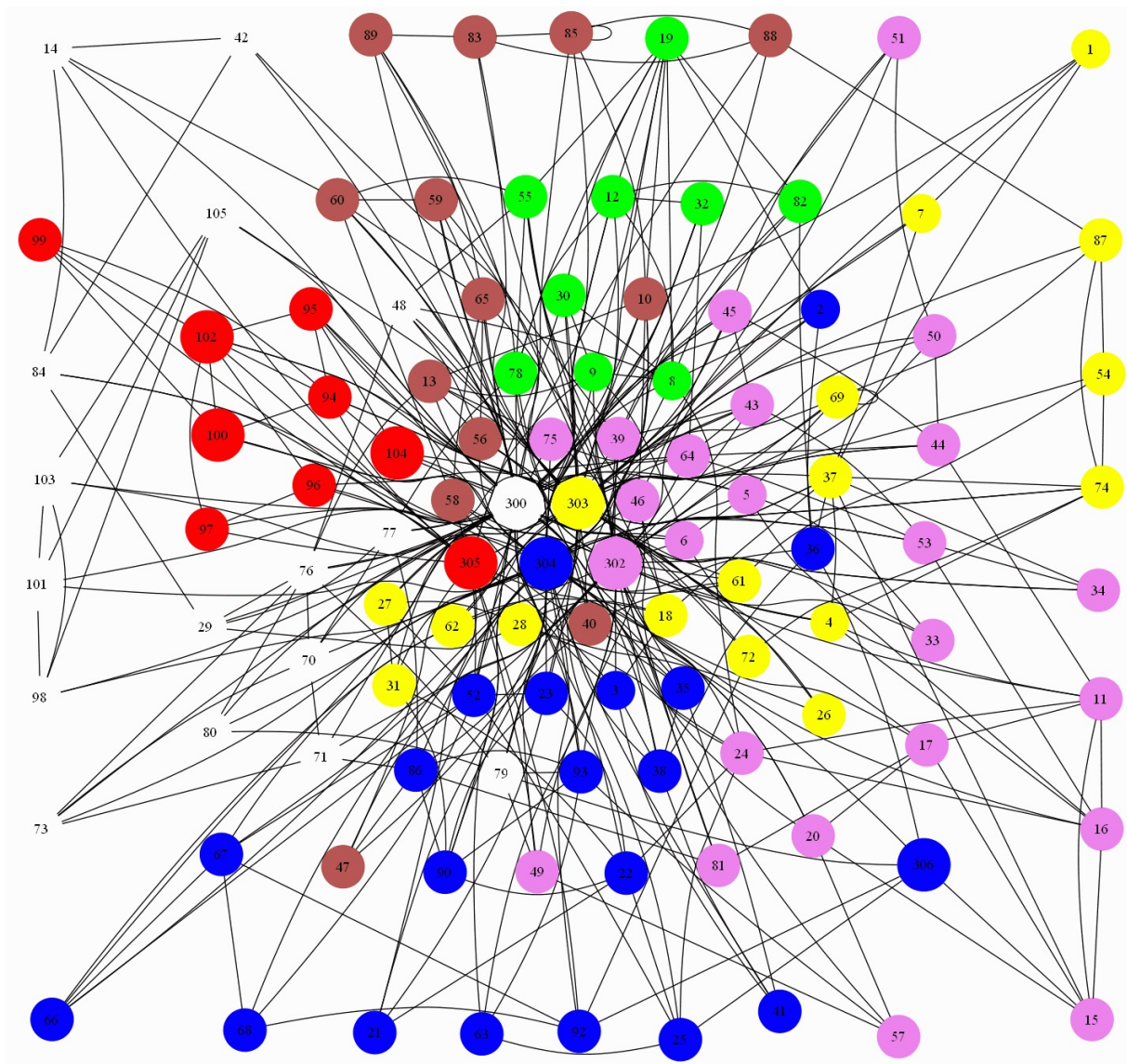


FIG. 4.7: Grafo do Novo Questionário montado com a ferramenta desenvolvida, já com as comunidades detectadas

- As disciplinas Álgebra Linear, Cálculo e Probabilidade e Estatística foram consideradas um mesmo nó, número 300.
- As disciplinas Eletricidade Técnica e Física/Laboratório de Física foram consideradas um mesmo nó, com número 302.
- As disciplinas Introdução à Computação e Cálculo Numérico foram consideradas um mesmo nó, com número 303.
- As disciplinas Resistência dos Materiais, Mecânica e Desenho Técnico foram con-

sideradas um mesmo nó, com numeração 304.

- As disciplinas Introdução a Ciência dos Materiais e Química foram consideradas um mesmo nó, com numeração 305.
- As disciplinas Projeto de Inglês e Redação Científica foram consideradas um mesmo nó, com numeração 306.

4.4.1 COMPARAÇÃO DO NOVO **GRAFO QUESTIONÁRIO** COM O **GRAFO SEÇÃO REDUZIDO**

Depois de fazer as novas divisões de disciplinas, em grupos mais condizentes com as respectivas engenharias, deve-se fazer uma nova comparação dos resultados. A nova comparação será feita entre os dois grafos montados pelas pesquisas realizadas com os alunos (**Grafo Seção** reduzido e o novo **Grafo Questionário**), com o objetivo de diminuir possíveis interferências do processo de escolha na distribuição dos alunos.

Isto é, o **Grafo Seção** reduzido representa as possíveis escolhas dos alunos independente da limitação de vagas imposta aos alunos militares, o que pode levar um aluno a uma SE não pretendida anteriormente e causar erros de previsão no novo **Grafo Questionário**. Dessa forma, usando o **Grafo Seção** reduzido pretende-se eliminar esse erro de previsão.

Quando se compara o **Grafo Seção** reduzido com o novo **Grafo Questionário**, pode-se agrupar os nós que se encontram nas mesmas comunidades nos dois grafos e isso nos dá um indicativo da validade das comunidades encontradas e de como os alunos se relacionam quando têm interesses em comum. A Tabela 4.1 quantifica as interseções.

Seção/Questionário	300	302	303	304/306	305	Com. 6	Com.7
SE/2 e SE/8	5	5	6	5	2	2	6
SE/3	2	11	4	6	2	2	3
SE/4 e SE5	5	4	2	8	2	3	2
SE/6	3	3	3	1	3	0	1

TAB. 4.1: Quantidade de pessoas que estão na interseção das duas Comunidades obtidas pelo Algoritmo CNM nos dois Grafos.

Observa-se que uma grande quantidade de alunos que pretende ir para a SE/3 estão na Com. 2 do novo **Grafo Questionário**, esse resultado era esperado pois a Com. 2 possui o nó 302 (Eletricidade Técnica e Física/Laboratório de Física).

O mesmo ocorre com os alunos da SE/4, que estão em maior quantidade na Com. 4, que possui o nó 304 (Resistência dos Materiais, Mecânica e Desenho Técnico). No **Grafo Seção** reduzido a SE/4 e a SE/5 se agruparam na mesma comunidade, isso pode ser resultado do interesse comum em matérias como Introdução a Ciência dos Materiais, Cálculo e Física.

Além disso, pode-se observar que a Com. 1 e a Com. 5 teve uma distribuição bem equilibrada entre todas as SE's. Este equilíbrio pode ser proveniente do fato dos alunos do IME serem selecionados com provas apenas de Matemática, Física e Química (além de Português e Inglês), isto é, os alunos já têm uma preferência natural pelas matérias da Com. 1, nó 300, e da Com. 5, nó 305, que englobam, entre outras, Cálculo, Álgebra Linear e Química.

Outro fato curioso foi a presença de comunidades sem nós de disciplinas, Com. 6 e Com. 7. A Com. 6 ficou dividida equilibradamente entre as SE /2, /3, /4, /5 e /8. Já a Com. 7 se concentrou nas SE /2 e /8, o que pode indicar um agrupamento de alunos com objetivos comuns de escolha de curso.

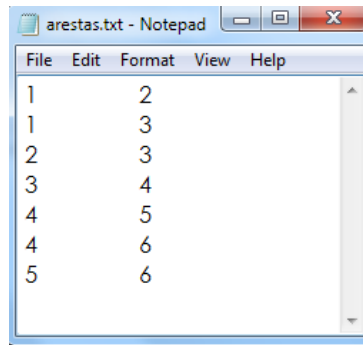
Os resultados obtidos podem ser considerados satisfatórios, principalmente para a SE/3 e também para as SE /4 e /5. No entanto, o estudo de uma turma apenas dificulta o trabalho de previsão do comportamento dos alunos e mais testes devem ser feitos para a consolidação e melhoria dos resultados já obtidos.

5 GERADOR DE GRAFOS

Neste capítulo, apresentamos uma descrição da página criada para inserir os dados reais que permitem obter um grafo, no qual será executado o algoritmo CNM.

Após os testes realizados com dados dos alunos do básico, se percebeu a necessidade de automatizar a montagem dos grafos. Os grafos utilizados neste trabalho são formados por uma média de 100 nós e uma quantidade de arestas que varia de 300 a 700.

O grafo de entrada do algoritmo CNM é representado por um arquivo de texto que descreve a lista de arestas do mesmo. A figura 5.2, mostra o grafo correspondente ao arquivo de texto contendo a lista de arestas na Figura 5.1.



1	2
1	3
2	3
3	4
4	5
4	6
5	6

FIG. 5.1: Exemplo de uma entrada para o software

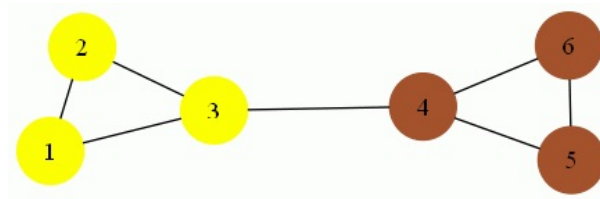


FIG. 5.2: Grafo de Exemplo

Devido a quantidade de arestas que cada grafo de alunos possui, o processo de digitação dos arquivos de texto é muito trabalhoso. Isso implicou na busca de uma alternativa para a construção desses grafos, para permitir a posterior utilização deste trabalho por outros usuários.

Para obter uma interface que permita gerar o grafo no formato de entrada do algoritmo CNM, será criada uma página html com um responsável por cadastrar os integrantes da

turma a ser pesquisada e, posteriormente, aplicar o questionário aos alunos daquela turma.

Esse processo de inserção de dados é dividido em duas fases. Na primeira, o responsável se loga na página e cadastra a turma de alunos. Depois disso, o responsável pode convocar os alunos a responderem a pesquisa. Nessa fase, o aluno escolhe seu nome dentre a lista de nomes de alunos cadastrados e responde ao questionário de afinidade, ou a lista de especialidades pretendidas, e a página adiciona as arestas correspondentes no grafo. É importante destacar que esse processo já resulta na construção do arquivo txt que é usado como entrada no programa de detecção de comunidades.

As condições para as respostas dos alunos ainda se mantêm, com um padrão de 3 respostas por aluno, tanto para as duas perguntas no questionário, quanto para a escolha de SE's. Esse foi mais um ponto que foi melhorado com a implementação do site, com o questionário escrito alguns alunos não obedeciam as regras estabelecidas e davam mais de 3 respostas a algumas perguntas. Isso produziu um desbalanceamento no grafo e contribuiu para erros na detecção de comunidades. Com a criação da página esse problema será evitado.

5.1 PROJETO

Nesta seção, será abordado um caso de uso, na Seção 5.1.1 relativo a implementação do gerador de grafos e, logo em seguida, na Seção 5.1.2, será apresentado o processo completo do projeto.

A discussão das funcionalidades de cada página foram importantes para direcionar o trabalho e dar início ao projeto. Os requisitos identificados e já citados para a construção do gerador de grafos são destacados abaixo:

- a) Facilidade de uso;
- b) Transparência na construção do arquivo de texto do grafo;
- c) Simplicidade na solução;
- d) Praticidade e agilidade no processo de aplicação das pesquisas.

5.1.1 PROCESSO COMPLETO

A Figura 5.3 mostra o processo completo do projeto. De uma forma geral, a partir de *Dados* de entrada, é produzido o resultado final, um conjunto de *Comunidades*.

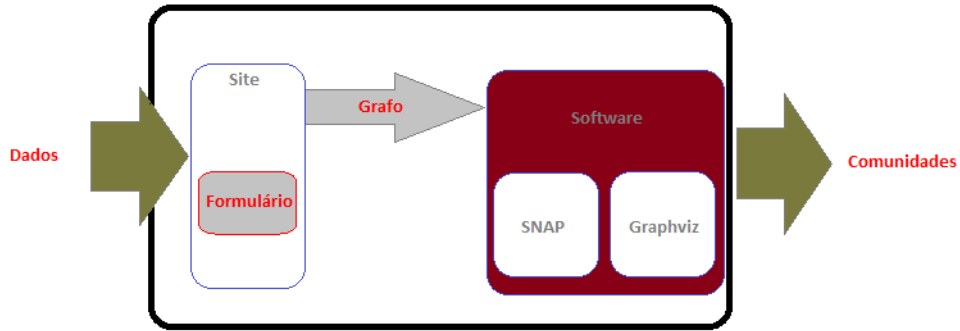


FIG. 5.3: Processo Completo

A seguir, será descrito o processo completo do projeto que é composto por 3 passos. A Figura 5.4 mostra o primeiro passo. O nome e o número de cada aluno constitui o conjunto de dados, que uma vez inseridos são disponibilizados em um arquivo *xml*.

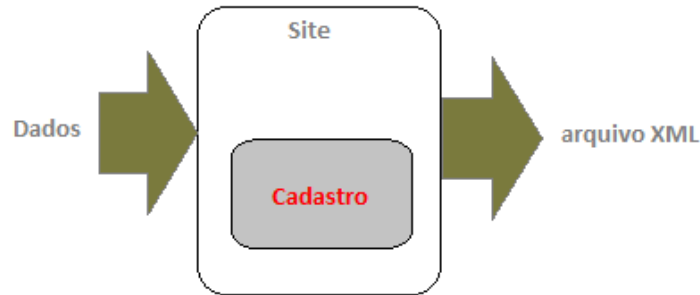


FIG. 5.4: Processo cadastro de alunos

O segundo passo é mostrado na Figura 5.5. A partir de um formulário disponibilizado na página *html*, que será respondido pelos próprios alunos, é gerado um arquivo *txt* (que corresponde a lista de arestas do grafo).

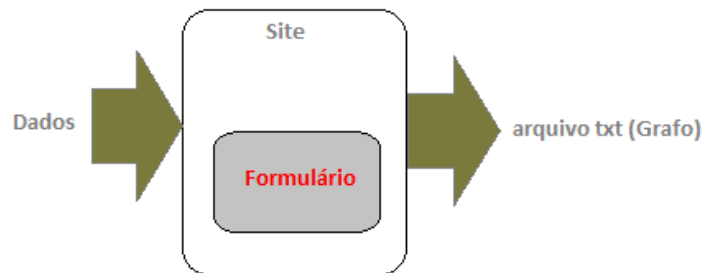


FIG. 5.5: Processo preencher formulário(responder questionário/ficha de escolha)

No último passo, que pode ser visto na Figura 5.3, a partir do grafo dado como

entrada no software , é gerado o resultado principal, que é o arquivo com as comunidades determinadas pelo algoritmo CNM da biblioteca SNAP, assim como um desenho deste grafo com as comunidades coloridas, que é gerado pelo Graphviz.

5.1.2 CASO DE USO

Com a definição das funcionalidades e se baseando nos requisitos, elaboramos um diagrama de casos de uso que está apresentado na Figura 5.6.

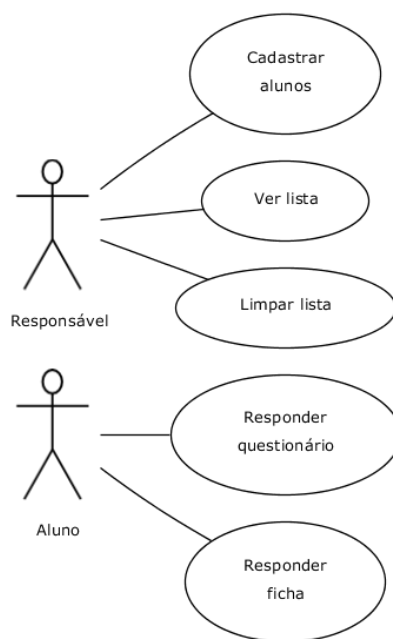


FIG. 5.6: Diagrama de casos de uso.

Um dos detalhamentos de caso de uso elaborados é apresentado a seguir, a fim de exemplificar como se deu o processo de montagem do site. Os outros casos de uso são mais simples do que o apresentado e seu objetivo pode ser entendido pelo nome.

Caso de uso: Cadastrar alunos

Ator: Responsável

Atores secundários: Não há

Pré-condições: Nenhuma

Fluxo básico de eventos

- a) Este caso de uso começa quando o responsável deseja cadastrar uma nova turma de alunos para realização das pesquisas.

- b) O responsável insere o nome e o número do aluno e confirma.
- c) O sistema mostra os dados inseridos.
- d) O sistema cadastra o aluno no arquivo e o caso de uso termina.

Fluxo alternativo de eventos

O RESPONSÁVEL DESEJA INSERIR MAIS UM ALUNO No passo [3. O sistema cadastra o aluno na lista e o caso de uso termina.(do fluxo básico de eventos)], se o responsável deseja inserir mais algum aluno:

- a) O responsável retorna para a página inicial e o caso de uso continua do passo 2 do fluxo básico de eventos.

Pós-condições:

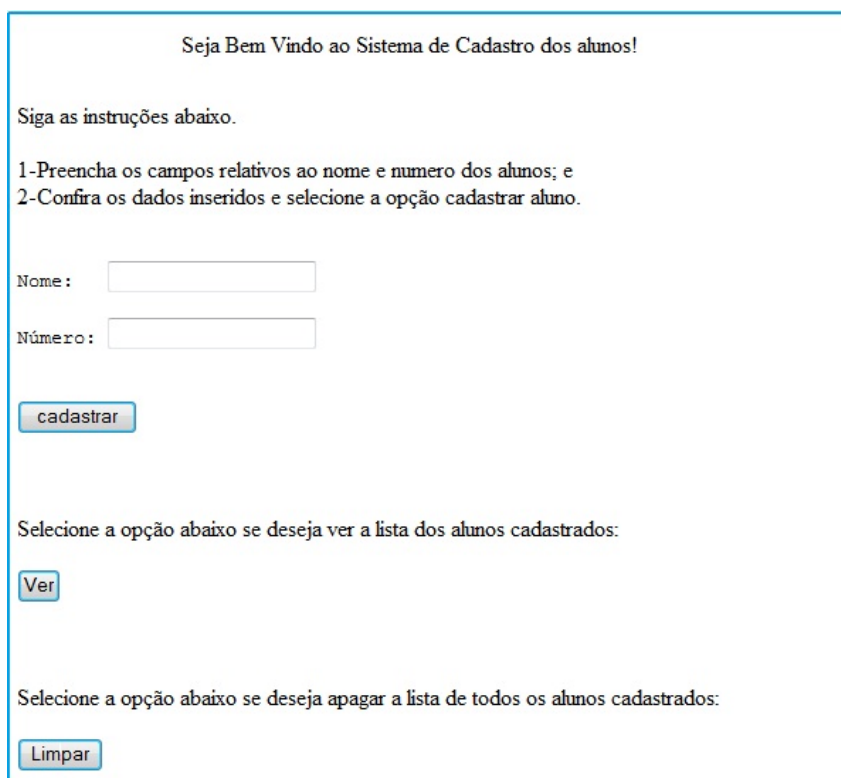
- O sistema registrou a lista de alunos no banco de dados.(para os fluxos básico e alternativo de eventos.)

A descrição de caso de uso apresentada é bem breve, devido à simplicidade objetivada na construção da ferramenta. Pode-se notar que o site é rapidamente entendido, sendo bastante usável. Os outros casos de uso não diferem muito e levam em conta as funcionalidades desenvolvidas. O objetivo de facilitar o processo de construção do grafo e torná-lo transparente para todos os usuários foi levado em conta.

O processo é de fácil entendimento, os alunos são cadastrados com seus respectivos números e a saída desse passo é um arquivo xml, com os nomes e números dos alunos, que representa os nós do grafo. Com os alunos cadastrados, os mesmos são chamados a responder as pesquisas e a página já registra as arestas correspondentes no arquivo txt que representa o grafo e serve de entrada para o programa de detecção de comunidades.

5.2 CADASTRANDO ALUNOS

O cadastro de alunos é iniciado com a tela mostrada na Figura 5.7. Depois de cadastrar cada aluno, o responsável pela inserção é redirecionado para outra página que permite verificar erros antes de salvar a inserção no arquivo *xml* (Figura 5.8).



Seja Bem Vindo ao Sistema de Cadastro dos alunos!

Siga as instruções abaixo.

1-Preencha os campos relativos ao nome e numero dos alunos; e
2-Confira os dados inseridos e selecione a opção cadastrar aluno.

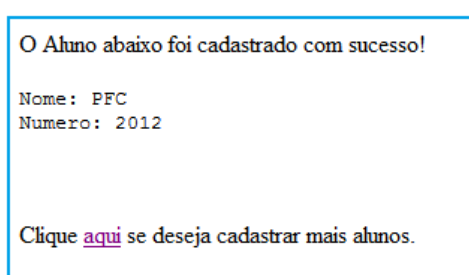
Nome:

Número:

Selecione a opção abaixo se deseja ver a lista dos alunos cadastrados:

Selecione a opção abaixo se deseja apagar a lista de todos os alunos cadastrados:

FIG. 5.7: Página inicial do sistema de Cadastro de alunos.



O Aluno abaixo foi cadastrado com sucesso!

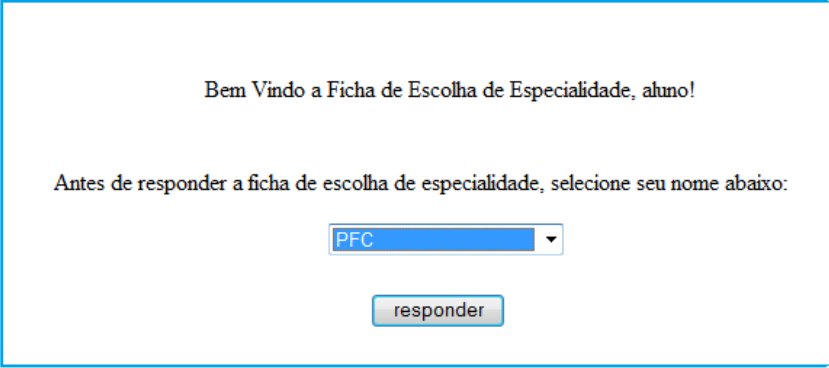
Nome: PFC
Numero: 2012

Clique [aqui](#) se deseja cadastrar mais alunos.

FIG. 5.8: Página de confirmação do cadastro de um aluno.

5.3 PREENCHENDO A FICHA DE ESCOLHA DE ESPECIALIDADE

Para inserir os dados correspondentes à escolha de especialidades, cada aluno deve informar o nome (que foi previamente cadastrado pelo responsável). A Figura 5.9 mostra a tela correspondente.



Bem Vindo a Ficha de Escolha de Especialidade, aluno!

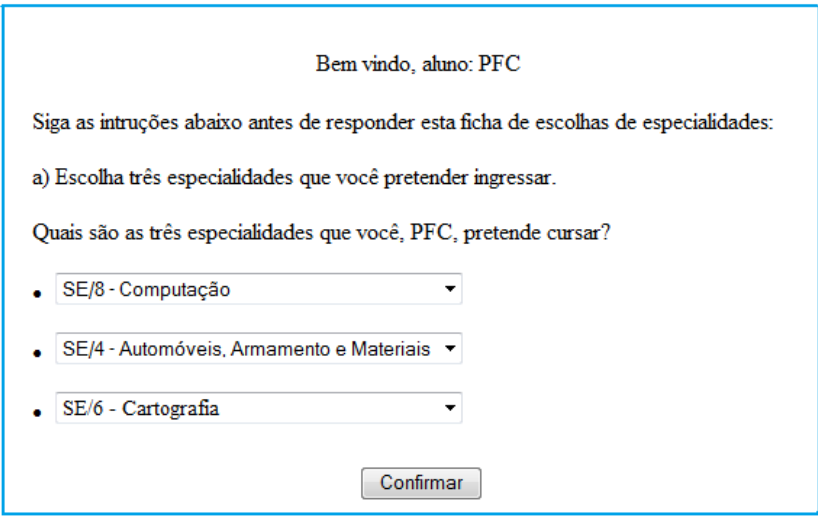
Antes de responder a ficha de escolha de especialidade, selecione seu nome abaixo:

PFC

responder

FIG. 5.9: Página inicial para responder a Ficha de Escolha de Especialidades.

Posteriormente, o aluno é redirecionado para a responder a ficha, mostrada na Figura 5.10



Bem vindo, aluno: PFC

Siga as instruções abaixo antes de responder esta ficha de escolhas de especialidades:

a) Escolha três especialidades que você pretender ingressar.

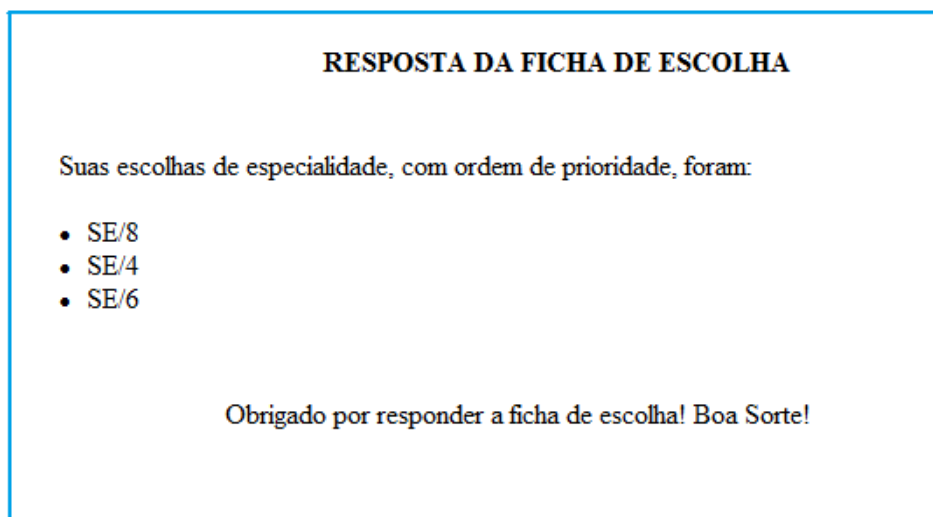
Quais são as três especialidades que você, PFC, pretende cursar?

- SE/8 - Computação
- SE/4 - Automóveis, Armamento e Materiais
- SE/6 - Cartografia

Confirmar

FIG. 5.10: Página onde o aluno seleciona suas escolhas de especialidades.

Uma vez que o preenchimento foi finalizado, o aluno confere suas escolhas, como mostra a Figura 5.11.



RESPOSTA DA FICHA DE ESCOLHA

Suas escolhas de especialidade, com ordem de prioridade, foram:

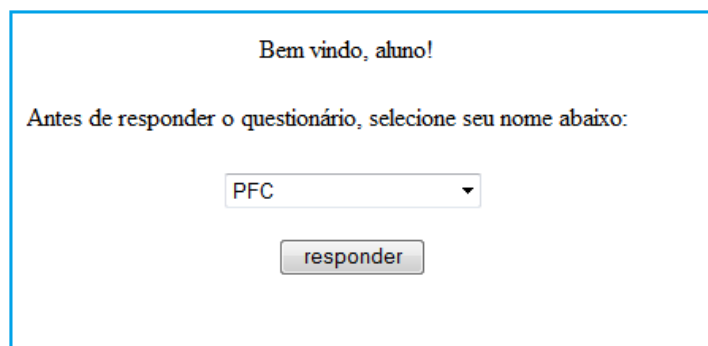
- SE/8
- SE/4
- SE/6

Obrigado por responder a ficha de escolha! Boa Sorte!

FIG. 5.11: Página de conferência das respostas.

5.4 RESPONDENDO O QUESTIONÁRIO

Da mesma forma que para preencher a escolha de especialidades, o aluno deve informar o nome para iniciar o processo de preenchimento do questionário. A Figura 5.12 mostra a página inicial para o Sistema de Resposta de Questionário pelo aluno.



Bem vindo, aluno!

Antes de responder o questionário, selecione seu nome abaixo:

PFC

responder

FIG. 5.12: Página inicial para o sistema de resposta do questionário.

Posteriormente, o aluno deve responder às perguntas (ver Figura 5.13). Após confirmar as respostas (Figura 5.14) as informações são salvas.

Siga as instruções abaixo antes de responder o questionário:

- a)Selecione no máximo três alunos da turma por questão;
- b)Caso deseje selecionar menos de 3 alunos em uma questão, selecione a opção "Nenhum" em alguns campos relativos a esta pergunta;
- c)Na última pergunta, selecione no máximo três disciplinas relativa a pergunta;
- d)Caso deseje selecionar menos de três disciplinas, selecione a opção "Nenhum"em alguns campos relativos a esta pergunta

QUESTIONÁRIO

Questão 1

Quais são os colegas do 2º ano com os quais você mais estuda para as provas e trabalhos das disciplinas do IME?

-
-
-

Questão 2

Dentre as disciplinas do 1º e 2º ano básicos, quais são as 3 que você mais se identifica?

-
-
-

FIG. 5.13: Página do questionário

RESPOSTA DO QUESTIONÁRIO

Suas respostas para as perguntas foram:

QUESTÃO 1:

Alunos: TELES, POLLYANA, RAPHAELA

QUESTÃO 2:

Disciplinas: Introdução a Computação, Cálculo Numérico, Álgebra Linear

FIG. 5.14: Página de conferência do questionário

6 CONCLUSÃO

O trabalho desenvolvido neste projeto de fim de curso iniciou-se com o estudo de três algoritmos para detectar comunidades em redes sociais baseadas nas idéias de CNM (3). Posteriormente, foi modelado um grafo de relacionamento a partir de dados reais de alunos do 2º ano do básico do IME em novembro de 2011. Com informações obtidas da ficha de escolha de especialidades dessa mesma turma foi possível obter 2 grafos: **Grafo Questionário** e **Grafo Seção** (4). Um site denominado Gerador de Grafos foi produzido no contexto deste trabalho que permite automatizar a coleta de dados para geração dos dois grafos propostos (5). Os grafos obtidos a partir de dados reais foram utilizados como entrada do algoritmo de CNM para detectar comunidades implementado na biblioteca SNAP (3.2) e o resultado pode ser visualizado utilizando o software Graphviz.

O **Grafo Questionário** reduzido apresentou seis comunidades. Foram identificadas as disciplinas em cada comunidade e tentou-se comparar com a real escolha dos alunos no ano de 2012. Os resultados não foram satisfatórios, pois para cada comunidade, além de existirem no mínimo quatro alunos de especialidades diferentes, três dessas especialidades apresentavam quantidades parecidas de alunos dentro dessa comunidade.

O **Grafo Seção** reduzido apresentou resultados satisfatórios, pois houve um acerto de mais de 50% dos resultados, ou seja, mais da metade dos alunos que escolheram as três engenharias na Ficha de Escolha de Especialidade realmente está na seção correspondente em 2012.

Depois de realizadas essas montagens, estudos e comparações, foi observada a necessidade de melhorar a estrutura do **Grafo Questionário** visando obter um resultado mais satisfatório. Além disso, foi constatado que o processo de montagem do grafo era muito trabalhoso e exigia um tempo de estudo para que fosse compreendido.

Dessa forma, um processo de automatização da montagem do grafo que fosse transparente ao usuário quanto a como o arquivo era constituído foi iniciado. Os requisitos resultaram no projeto de um site que facilita tanto a entrada da turma de alunos, que constituirão os nós do grafo, quanto a estruturação do arquivo de texto que representa as arestas do grafo e serve de entrada para o programa de detecção de comunidades.

Como era necessária uma reestruturação do **Grafo Questionário**, o site desenvolvido

já foi utilizado nessa fase do trabalho e isso foi um ótimo teste, que demonstrou que os requisitos do projeto foram atendidos. Montado o novo **Grafo Questionário**, o algoritmo CNM foi novamente aplicado e o resultado foi comparado com o **Grafo Seção**, que havia apresentado um resultado muito positivo quanto a constituição de suas comunidades. A comparação demonstrou que o novo **Grafo Questionário** era realmente mais correto do que o anterior, evidenciando o agrupamento de alunos por interesse em disciplinas e também pelo interesse em algum curso em particular.

Apesar da melhora nos resultados do **Grafo Questionário**, os mesmos ainda não podem ser considerados para uma previsão mais acertada da quantidade de pessoas que irão para cada especialidade, pois, como se trata de uma turma de acadêmicos de engenharia, a maioria tem um gosto maior pelas ciências exatas e uma das perguntas do questionário foca na afinidade dos alunos com as disciplinas, levando a uma maior tendência para as disciplinas de cálculo, física e química, o que dificultou uma separação em comunidades mais bem definidas. Espera-se que a aplicação dos algoritmos de Wakita ou Blondel, apresentem um resultado melhor devido a maior refinação nos resultados obtidos observada nos artigos em que foram publicados, [7] e [2] respectivamente. Esse é um ótimo trabalho futuro.

Propomos como trabalhos futuros, um refinamento das disciplinas usadas para o questionário com os alunos e realização de novos experimento com dados provenientes de outras turmas a fim de confirmar as previsões realizadas nesse trabalho e, possivelmente, fazer novas observações quanto a previsão da formação de turmas e o que o relacionamento entre os alunos influi nisso. Nesses novos experimentos, dividir a turma entre alunos civis e militares, devido ao problema de limitação de vagas para alunos militares que não existe para os alunos civis. Além disso, considerar a turma do CG como um grupo separado e analisar as comunidades nesse grupo também pode retornar resultados interessantes, visto que as vagas para a turma do CG são separadas e eles dificilmente se relacionam fora da própria turma.

Um trabalho mais diferenciado seria a modelagem dos grafos com pesos nas arestas, baseados em novas perguntas no questionário ou mesmo outros métodos julgados convenientes, a fim de diferenciar os tipos de relacionamentos entre alunos para detecção de comunidades em grafos com pesos. Todos os algoritmos estudados podem ser aplicados em grafos com pesos e isso seria mais um facilitador.

7 REFERÊNCIAS BIBLIOGRÁFICAS

- [1] BARBOSA, D. A, B. A.-L. D. S. R. F. O. C. J. C. Medidas de centralidade e detecção de comunidades em rede de co-autoria. **Anais do XLIII Simpósio Brasileiro de Pesquisa Operacional**, págs. 1–10, 2011.
- [2] BLONDEL, V. D., G. J.-L. L. R. L. E. Fast unfolding of communities in large networks. **J. Stat. Mech.** **P10008**, 2008.
- [3] CLAUSET, A., N. M. M.-C. O. C. Finding community structure in very large networks. **Physical Review E**, **70:066111**, 2004.
- [4] FORTUNATO. Community detection in graphs. **arxiv**, 2, 2010.
- [5] GIRVAN, M, N. M. Community structure in social and biological networks. **Proceedings of the National Academy of Sciences of the United States of America**, págs. 7821–7826, 2002.
- [6] KLEINBERG, J. EASLEY, D. Networks, crowds and markets. **Cambridge University Press**, 2010.
- [7] KWAK, H., C. Y. E.-Y. J. H. M. S. Mining communities in networks: A solution for consistency and its evaluation. **Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference (IMC'09)**, págs. 301–314, 2009.