Pós-Graduação em Ciência da Computação

“Caracterização de Grupos para Entendimento de Redes Sociais Educacionais”

Por

João Emanoel Ambrósio Gomes

Dissertação de Mestrado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE, JULHO/2013

UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

JOÃO EMANOEL AMBRÓSIO GOMES

"Caracterização de Grupos para Entendimento de Redes Sociais Educacionais"

ESTE TRABALHO FOI APRESENTADO À PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DE PERNAMBUCO COMO REQUISITO PARCIAL PARA OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIA DA COMPUTAÇÃO.

ORIENTADOR: Prof. Dr. Ricardo Bastos Cavalcante Prudêncio

RECIFE, JULHO/2013

Catalogação na fonte
Bibliotecária Jane Souto Maior, CRB4-571

Gomes, João Emanoel Ambrósio
Caracterização de grupos para entendimento de redes sociais educacionais / João Emanoel Ambrósio Gomes. - Recife: O Autor, 2013.
89 f.: il., fig., tab.

Orientador: Ricardo Bastos Cavalcante Prudêncio.
Dissertação (mestrado) - Universidade Federal de Pernambuco. CIn, Ciência da Computação, 2013.

Inclui referências e anexo.

1. Ciência da Computação. 2. Inteligência Artificial. I. Prudêncio, Ricardo Bastos Cavalcante (orientador). II. Título.

004 CDD (23. ed.) MEI2013 – 131

Dissertação de Mestrado apresentada por **João Emanoel Ambrósio Gomes** à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título "**Caracterização de Grupos para Entendimento de Redes Sociais Educacionais**" orientada pelo **Prof. Ricardo Bastos de Cavalcante Prudêncio** e aprovada pela Banca Examinadora formada pelos professores:

_____

Prof. Geber Lisboa Ramalho
Centro de Informática / UFPE

_____

Prof. Luciano Rogerio de Lemos Meira
Departamento de Psicologia / UFPE

_____

Prof. Ricardo Bastos Cavalcante Prudêncio
Centro de Informática / UFPE

Visto e permitida a impressão.
Recife, 17 de julho de 2013

_____

Profa. Edna Natividade da Silva Barros
Coordenadora da Pós-Graduação em Ciência da Computação do
Centro de Informática da Universidade Federal de Pernambuco.

Dedico este trabalho a minha família, professores e amigos
que me deram o apoio necessário a sua concretização.

Agradecimentos

Primeiramente agradeço a Deus pela saúde que me proporcionou ao longo de toda à minha vida, pela força e coragem que me fizeram nunca desistir e a cada dia mais superar novos desafios.

À minha família, especialmente aos meus pais, Francisco de Assis Gomes (Tuta) e Lúcia Ambrósio, aos meus irmãos Getúlio Kahlil e Kallyne Maria, e as minhas sobrinhas Iara Maria, Ana Júlia e Saha Maria, por todo o apoio proporcionado e por sempre me fazerem sentir que onde quer que eu esteja sempre teria uma casa para voltar. Vocês são uma luz na minha vida, eu nada seria sem vocês.

Ao meu orientador, professor Ricardo Prudêncio, pela confiança, competência, palavras de incentivo, dedicação e paciência em cada passo na construção deste trabalho. No momento de maior dificuldade no mestrado, se apresentou como um amigo e aceitou me orientar. Sem dúvidas, posso me considera um privilegiado em ser orientado por ele.

Aos meus amigos ("irmãos") Hilário Tomaz e Renê Gadelha pela grande convivência ao longo desses dois anos de mestrado, cujos momentos memoráveis lembrarei pelo resto da minha vida. Grandes aprendizados obtive com vocês, dentre eles, como se comportar em festas e como sair delas.

À empresa Joy Street por me dar a oportunidade de trabalho e por ter disponibilizado os dados para que este trabalho fosse realizado. Não esquecendo das amizades feitas no RHAE, em especial André Camara, Rafael Simplício, Alexandre, Ivan e Antônio.

Ao meu "irmão" Rodrigo Feitosa pelo acolhimento em João Pessoa em momentos difíceis, serei eternamente grato.

A todos os amigos do Brejo Santo em especial Pedro Ivo, Daniel, Davi, Armando, Antônio, Novinho, Moises, Joseph e Glauber que apesar da distância e do tempo nunca deixaram a amizade acabar.

Aos meus cunhados Ewerton e Luciana pela torcida e todo o apoio dado.

A minha namorada Regina por compreender minha ausência e por estar sempre ao meu lado.

A todos os amigos que tive a oportunidade conhecer ao longo do mestrado, em especial Jamilson, Airton, Lenin, Marco, Paulo Fernando, Kellyton, Janiel, Rafael e Alex.

Aos professores e amigos Ryan Ribeiro e Rodrigo Rocha por terem sempre me incentivado e encorajado o meu interesse pela vida acadêmica.

Ao Centro de Informática da Universidade Federal de Pernambuco - CIN-UFPE, pela excelente estrutura física e pessoal proporcionada a todos os seus alunos.

Eu sou de uma terra que o povo padece
Mas não esmorece e procura vencer.
Da terra querida, que a linda cabocla
De riso na boca zomba no sofrer.
Não nego meu sangue, não nego meu nome
Olho para a fome, pergunto o que há?
Eu sou brasileiro, filho do Nordeste,
Sou cabra da Peste, sou do Ceará.

—PATATIVA DO ASSARÉ

Resumo

Com o crescimento das redes sociais, diversas pesquisas vêm sendo realizadas para entendimento de suas estruturas. A detecção de comunidades nessas redes é importante para verificar especificidades que motivaram o agrupamento de usuários. Caracterizar comunidades pode auxiliar no entendimento das estruturas sociais existentes, como também na visualização, navegação e análise da rede. Caracterizar grupos em redes sociais no contexto educacional, denominadas Redes Sociais Educacionais (RSE), possibilita o fornecimento de conteúdos adaptáveis e um aprendizado direcionado aos grupos, auxiliando ainda, a tomada de decisão por parte de gestores e professores. As principais abordagens para caracterização de grupos são as baseadas em agregação e diferenciação. No entanto, relatos encontrados na literatura, indicam que os melhores perfis caracterizadores de grupos são obtidos por métodos baseados em diferenciação. Dessa maneira, este trabalho tem como objetivo avaliar a aplicação do teste estatístico *Wilcoxon Rank Sum* como método baseado em diferenciação para a caracterização de grupos em redes sociais. A partir dos atributos individuais dos usuários e comunidades identificadas, o método proposto busca caracterizar os grupos verificando características comuns que os diferenciem do restante da rede. Para avaliar o método proposto, foi realizado um estudo de caso com duas RSEs: as Olimpíadas de Jogos Digitais e Educação de Pernambuco (OJE-PE) e Acre (OJE-AC). Conforme verificado nos experimentos realizados, foram obtidos resultados promissores, dentre os quais se destacam a identificação de características descritivas para 80% dos grupos da rede OJE-PE e 70% para a OJE-AC.

Palavras-chave: Redes Sociais Educacionais, Detecção de Comunidades, Caracterização de Grupos

Abstract

Several researches have been conducted to understand the structures of social networks due to its increasing adoption. Community detection in these networks is important to identify features that motivated the users interconection. Characterize communities may assist the understanding of existing social structures, as well as, network navigation, visualization and analysis. Identify communities in social networks in the educational context, referred as Educational Social Networks (ESN), allows the sharing and adaptation of specific content to the groups, assisting managers and teachers to make better decisions. The main approaches for characterizing groups are based on aggregation and differentiation. Nevertheless, reports in the literature indicate that the best profiles to characterize groups are obtained by differentiation-based methods. Thus, this work aims to evaluate the application of the Wilcoxon Rank Sum test statistic as differentiation-based method to group profiling. Based on the individual user attributes and identified communities, the proposed method aims to characterize the groups, observing common characteristics that differentiate them from the rest of the network. In order to evaluate the proposed method, we conducted a case study with two ESNs: the Olympics Digital Games and Education of Pernambuco (OJE-PE) and Acre (OJE-AC). The performed experiments showed that the method was effective in identifying features to characterize the groups, pointing descriptive characteristics for 80% of groups in OJE-PE and 70% for OJE-AC.

Keywords: Educational Social Networks, Community Detection, Group Profiling

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

ARS	Análise de Redes Sociais
CGBA	Caracterização de Grupos Baseado em Agregação
CGBD	Caracterização de Grupos Baseado em Diferenciação
CGBDE	Caracterização de Grupos Baseado em Diferenciação Egocêntrica
MD	Mineração de Dados
MDE	Mineração de Dados Educacionais
MAM	*Multi-Level Aggregation Method*
OJE	Olimpíadas de Jogos Digitais e Educação
RSE	Redes Sociais Educacionais
WRS	*Wilcoxon Rank Sum*

1 Introdução

The first step towards getting somewhere is to decide
that you are not going to stay where you are.
—MORGAN, JOHN PIERPONT

1.1 Contexto e Motivação

Os desafios gerados pelas mudanças tecnológicas são significativos na sociedade, modificando até as formas como as pessoas interagem. Dentre essas tecnologias, as redes sociais têm apresentado grande destaque, tornando-se um fenômeno global muito presente na vida social humana. Diante desse destaque, surgem as Redes Socais Educacionais (RSE), serviço Web que oferecem aos alunos a oportunidade de interagir com outros alunos, professores, administradores, sem limitações geográficas (Boyd and Ellison, 2010).

Pesquisadores apoiam a aplicação das redes sociais na educação por sua capacidade de atrair, motivar e envolver os alunos em práticas significativas de comunicação, troca de conteúdo e colaboração; tornando o processo de aprendizagem mais dinâmico (Mills, 2011). No Brasil as RSE vêm crescendo e sendo bastante utilizadas, como exemplos: as Olimpíadas de Jogos Digitais e Educação (OJE[1]) e Redu Educational Technologies (REDU[2]).

Organizações que utilizam as RSEs são capazes de coletar grandes massas de dados em servidores web, bancos de dados e em logs de acesso, em formatos diversos. Esses repositórios de dados contêm informação que pode ser útil para modelagem e avaliação do processo de aprendizado, auxiliando ainda, a tomada de decisão por parte de gestores

[1] www.educacao.pe.gov.br/oje
[2] www.redu.com.br/

e administradores dos sistemas educacionais (Romero and Ventura, 2007; Mostow and Beck, 2006).

Paralelamente, pesquisas sobre Mineração de Dados Educacionais (MDE) vêm sendo realizadas. Essa é uma área emergente que envolve a aplicação de técnicas computacionais para identificar padrões em grandes repositórios de dados educacionais (Baradwaj and Pal, 2012). A aplicação de técnicas sofisticadas de Mineração de Dados (MD) em sistemas educacionais são especialmente úteis, uma vez que um processo informal de análise não é viável devido à complexidade, variedade e quantidade dos dados coletados (Castro et al., 2007).

A utilização das redes sociais nos sistemas educacionais proporcionam novos desafios ao processo de MDE. Na realidade, a MD atualmente lida com o problema de minerar estruturas com muitas informações (propriedades), conjuntos de dados heterogêneos (como por exemplo, redes sociais). Tais conjuntos de dados são melhores representados por redes ou grafos (Getoor and Diehl, 2005). Nesse contexto, estudos sobre mineração de *links* vêm sendo realizados para Análise de Redes Sociais (ARS), dentre esses destacam-se os trabalhos voltados para análise das estruturas modulares das redes, denominadas grupos/comunidades (Baumes et al., 2004; Sun et al., 2007; Tang and Liu, 2010a).

A homofilia foi uma das primeiras características estudas pelos pesquisadores de redes sociais (Almack, 1922; Bott, 1928; Wellman, 1926). Segundo o conceito de homofilia (McPherson et al., 2001), a probabilidade de pessoas semelhantes se relacionarem é superior a de pessoas com características distintas. A homofilia apresenta grande destaque nos estudos de ARS, sendo definida como o conceito base para o estudo das comunidades. Basicamente, uma comunidade é um conjunto de usuários que interagem uns com os outros com frequência (Wasserman and Faust, 1994).

A formação de grupos em ambientes educacionais é evidente, seja essa por interesses comuns e/ou afinidades (Baradwaj and Pal, 2012). Por exemplo, alguns usuários podem interagir, porque eles compartilham a mesma escola/sala de aula, estão envolvidos nas mesmas atividades ou estão interessados no mesmo objeto de estudo ou curso. Portanto, nesse universo dinâmico no qual essas redes estão inseridas, é de extrema importância que se procure entender os processos de formação de suas comunidades.

1.2 Problema Abordado e Justificativa

É no contexto dinâmico e evolutivo da rede que a *Caracterização de Grupos* se insere. Esse segmento de pesquisa se refere a uma subárea da detecção de comunidades, que

é responsável por delinear as estruturas da rede, em termos de grupos, descrevendo-os em perfis e buscando justificar as suas formações. É um problema bastante relevante, com diversas aplicações, tais como: entendimento das estruturas sociais, visualização e navegação da rede, acompanhamento a mudanças de temas de um grupo, casos alarmantes e marketing direto (Tang and Liu, 2010b). Todavia, algumas questões fundamentais permanecem sem solução, como: qual é o principal motivo que une os usuários do grupo, e como interpretar e compreender as comunidades de uma rede?

Vários métodos que realizam agrupamento têm como objetivo encontrar os recursos que são compartilhados por todo o grupo, dessa forma, uma abordagem natural e simples seria encontrar atributos que são mais prováveis de ocorrer dentro do grupo (Senot et al., 2010). Porém, em vez de agregar, pode-se selecionar características que diferenciam um grupo de outros na rede. O objetivo é descobrir as principais características discriminativas que são representativas para um grupo (Liu and Motoda, 1998). Ou seja, busca-se selecionar atributos que aparecem em um grupo e, dificilmente, aparecem nos outros.

Segundo Tang et al. (2011) e Tang and Liu (2010b), as abordagem baseadas na agregação de atributos individuais só são viáveis em um ambiente relativamente livre de ruído. Se os perfis são construídos a partir de atributos ruidosos, como posts de usuários, registro de atividades de usuários ou relatos de interesses, abordagens baseadas em diferenciação apresentam melhores perfis para caracterização dos grupos.

Nas redes sociais, muitas informações relevantes podem ser obtidas de atributos ruidoso como, postagens de usuários e registros de atividades; ou seja, a consideração desses dados em um estudo de ARS é fortemente aconselhado. Por exemplo, em uma RSE muitos atributos relevantes podem ser retirados de registro de atividade no site, tais como: quantidade de acessos a uma disciplina, maior acesso a um área educacional (humanas, exatas, etc), quantidade de questões respondidas no ambiente, entre outros. Apesar desses atributos terem grandes de chances de apresentarem ruídos, estes podem vir a ser excelentes caracterizadores em um estudo de criação de perfis de grupos sociais.

Outro ponto importante em dados coletados de redes sociais, é o fato dos atributos dos usuários não apresentarem valores equilibrados (em sua grande maioria), ou seja, alguns usuários apresentarem um maior "destaque" em relação a outros na rede. Continuando a análise em um RSE, são exemplos: alunos com grande número de links (*hubs*), alunos que participam ativamente dos jogos, alunos que acessam eventualmente, alunos que respondem muitas questões, etc; consequentemente, isso tende a gerar uma distribuição não normalizada entre os atributos dos usuários/grupos da rede.

Em um estudo estatístico, a escolha de qual método utilizar está diretamente rela-

cionada com forma pela qual os dados encontram-se distribuídos. Em caso dos dados não estarem em uma distribuição normal, métodos paramétricos, como o *t-student*, não são viáveis (Goulden, 1956). Logo, um método não-paramétrico é frequentemente uma escolha mais segura, sendo também mais robusto. Uma vez, que esses testes não estão condicionados por qualquer distribuição de probabilidades dos dados em análise, sendo também designados por "*distribution-free tests*" (Lehmann, 1975).

Partindo dos problemas apontados na utilização dos dados coletados de um rede social, dados não normalizados e presença de atributos ruidosos, a aplicação de um teste estatístico não paramétrico como um método de caracterização de grupos baseada em diferenciação, se apresenta com uma possível solução aos problemas supracitados.

Dessa forma, neste trabalho é proposto um método baseado em diferenciação para caracterização de grupos em redes sociais, aplicando-se o teste estatístico não-paramétrico *Wilcoxon Rank Sum* (Gomes et al., 2013). A partir dos atributos individuais dos usuários e comunidades identificadas, o método proposto busca caracterizar os grupos verificando interesses/características comuns aos grupos, que diferenciam esses do restante da rede. Almeja-se identificar o provável motivo de um conjunto de pessoas se conectarem ou interagirem umas com as outras, formando comunidades. Possibilitando o fornecimento de conteúdos adaptáveis e um aprendizado direcionado aos grupos, auxiliando ainda, na tomada de decisão por parte de gestores e professores.

Conforme apontado anteriormente, a realização de estudos voltados para o entendimento do processo de formação das comunidades em ambientes de RSEs são de grande relevância. Dessa forma, para avaliação do método proposto neste trabalho foi realizado um estudo de caso sobre duas RSEs: OJE Pernambuco (OJE-PE[3]) e Acre (OJE-AC[4]).

1.3 Objetivos

Com base no problema abordado, o presente trabalho tem como principal objetivo avaliar a aplicação do teste *Wilcoxon Rank Sum* como método baseado em diferenciação para a caracterização de grupos em redes sociais.

Para alcançar o objetivo principal de maneira satisfatória, alguns objetivos específicos foram estabelecidos:

- Realizar um levantamento teórico e uma análise dos estudos mais relevantes do estado da arte de mineração de dados educacionais (técnicas voltadas para análise

[3] www.educacao.pe.gov.br/oje
[4] www.oje.inf.br/acre/

de grupos e redes sociais) e caracterização de grupos, contribuindo assim para melhor entendimento dos fundamentos, aplicabilidade, limitações e tendências da área; bem como, identificação de métodos e técnicas já consolidadas pela comunidade acadêmica no tratamento de problemas relacionados;

- Identificação e extração dos atributos individuais dos usuários das redes sociais em estudo (OJE-PE e OJE-AC);
- Elaboração de um modelo para filtragem dos principais atributos caracterizadores dos usuários, representação dos usuários da rede;
- Validação dos atributos selecionados para representação dos usuários com especialista do domínio da rede em estudo;
- Implementação do algoritmo de detecção de comunidade *Multi-Level Aggregation Method*;
- Baseado nos atributos selecionados na representação dos usuário, aplicar o teste estatístico *Wilcoxon Rank Sum* para caracterização dos grupos identificados pelo algoritmo de detecção de comunidade.
- Avaliar o desempenho dos perfis caracterizadores dos grupos gerados, analisando as relações dos resultados entre as redes que o método proposto foi executado.

1.4 Estrutura da Dissertação

Em função da proposta de pesquisa e dos objetivos levantados, esta dissertação encontra-se organizada da maneira descrita a seguir.

- Capítulo 2: São apresentados os conceitos sobre a mineração de dados educacionais, mais precisamente as técnicas voltadas para análise de grupos e redes sociais;
- Capítulo 3: Detalha-se os conceitos sobre as redes sociais relacionados ao estudo de caracterização de grupos;
- Capítulo 4: Baseando-se no levantamento bibliográfico, o método proposto é descrito, a sua viabilidade é verificada através da realização de experimentos nas duas RSEs adotadas, e por fim apresenta uma análise dos resultados obtidos;

- Capítulo 5: Analisa as considerações finais deste trabalho, apresentando conclusões e contribuições, limitações do estudo e algumas sugestões de como este trabalho pode ser melhorado e continuado.

2

Mineração de Dados Educacionais

Do you wish to rise? Begin by descending. You plan a tower that will pierce the clouds? Lay first the foundation of humility.

—ST. AUGUSTINE

Grandes avanços foram observados nos sistemas educacionais nas últimas décadas como consequência da adoção de tecnologias diversas como plataformas de educação à distância, sistemas tutores inteligentes, jogos educacionais, redes sociais educacionais, dentre outras (Ha et al., 2000). Organizações que adotam tais sistemas são capazes de coletar grandes massas de dados em servidores web, bancos de dados e em logs de acesso, em formatos diversos. Esses repositórios de dados contêm informação que pode ser útil para modelagem e avaliação do processo de aprendizado, assim como auxiliar a tomada de decisão por parte de gestores e administradores dos sistemas educacionais (Romero and Ventura, 2007; Mostow and Beck, 2006).

Mesmo nos ambientes tradicionais de educação, pode ser observado o armazenamento de bases de dados (por exemplo: dados sobre alunos, professores, avaliações de disciplinas, informações curriculares) que devem ser analisadas de forma adequada (Markham et al., 2003). Dessa forma, a tomada de decisão sobre as vivências em salas de aula envolve observação dos comportamentos do aluno, a análise de histórico escolares e avaliação da eficácia das estratégias pedagógicas adotadas. Todavia, a realização desse monitoramento informal não é possível na análise de dados coletados de ambientes educacionais eletrônicos. Gerando, dessa forma, a necessidade do desenvolvimento de aplicações para coleta e análise de tais informações de forma eficaz.

O interesse sobre a análise automática de dados produzidos através da interação dos alunos com os ambientes de aprendizagem vem crescendo bastante recentemente

(Muehlenbrock, 2005). Técnicas de mineração de dados (do inglês, *Data Mining*), também conhecida como Descoberta de conhecimento em banco de dados (do inglês, *Knowledge Discovery in Databases - KDD*), são apontados como possíveis soluções para o desenvolvimento de ambientes de aprendizagem mais eficazes. Essa refere-se a disciplina que tem como objetivo descobrir "novas" informações através da análise de grandes quantidades de dados (Klösgen and Zytkow, 2002). Podendo ser usado para diversas tarefas em ambientes educacionais, tais como: aprender modelos do processo de aprendizagem (Hämäläinen et al., 2004), modelagens de estudantes (Tang and McCalla, 2002), avaliação e aperfeiçoamento de sistemas de *e-learning* (Zaïane, 2001), dentre outras.

Neste contexto, surge a Mineração de Dados Educacionais (MDE), uma área emergente que envolve a aplicação de técnicas computacionais para identificar padrões em grades repositórios de dados educacionais. Técnicas sofisticadas de Mineração de Dados (MD) são especialmente úteis no contexto de sistemas educacionais eletrônicos, uma vez que um processo informal de análise não é viável devido à complexidade, variedade e quantidade dos dados coletados(Castro et al., 2007). Dentre as técnicas de mineração mais usadas no contexto educacional pode-se citar modelos preditivos para classificação e regressão, técnicas de agrupamento de dados, técnicas de visualização e algoritmos de aprendizado de regras de associação (Baker et al., 2010).

Generalizando, o processo de MDE converte os dados brutos, provenientes de sistemas educacionais, em informações úteis que podem ter grandes impactos nas pesquisas e práticas educativas. Este processo não difere muito da aplicação de MD em outras áreas, tais como: negócio, genética, medicina, dentre outras. Pois realiza os mesmos passos do processo de MD geral (Romero et al., 2004): pré-processamento, MD, e pós-processamento. De acordo com Romero and Ventura (2007), as principais questões que diferenciam a aplicação MD na educação dos demais domínios, são:

1. *Objetivo*: O objetivo do MD difere para cada área de aplicação. Por exemplo, na educação, além do propósito da pesquisa, na busca de melhorias no processo de aprendizagem e orientação dos alunos; objetiva-se a investigação pura, almejando uma compreensão mais profunda dos fenômenos educacionais. Esses objetivos são, por si só, difíceis de se quantificar, exigindo seus próprios conjuntos de técnicas de medição;

2. *Dados*: São específicos para a área educacional, e portanto têm informações semânticas intrínsecas, relacionamento com outros dados, e múltiplos níveis de uma hierarquia de sentidos. Alguns exemplos são os modelos de domínio usados em

Sistemas de Tutoria Inteligentes (*Intelligent tutoring system* - ITS) e Sistemas Educacionais de Hipermídia adaptativa (*adaptive educational hypermedia system* - AEHS), os quais representam relacionamentos entre conceitos de assuntos específicos por meio de grafos ou uma hierarquia (por exemplo, um curso consiste de vários capítulos organizados em lições e cada lição inclui alguns conceitos). Além disso, também é necessário ter aspectos pedagógicos do aluno e do sistema em consideração;

3. *Técnicas*: dados e problemas educacionais apresentam algumas características especiais que exigem a aplicação da mineração para serem tratados de forma diferente. Embora a maioria das técnicas tradicionais de MD possam ser aplicadas diretamente a dados educacionais, outras não, gerando a necessidade de se adaptar manualmente tais técnicas aos problemas específico de ensino. Além disso, essas técnicas específicos de MD podem ser utilizados para resolução de problemas intrínsecos de ensino.

Neste contexto, os trabalhos relacionados à área de MDE apresentam uma grande diversidade em termos de objetivos, tarefas e técnicas empregadas (Romero and Ventura, 2007). Segundo Zorrilla et al. (2005), a MDE pode ser orientada a três tipos de atores: (1) estudantes: em trabalhos que realizam recomendações de atividades para alunos a partir das tarefas passadas já realizadas, observando-se os níveis de desempenho (Heraud et al., 2004) e a predição de engajamento individual (Lu, 2004); (2) educadores: como exemplo, pode-se citar a avaliação de dificuldade de problemas (Ha et al., 2000) e a personalização de cursos (Tang et al., 2000); e (3) administradores: como a avaliação do impacto de programas e currículos educacionais (Beck and Woolf, 2000) e a construção de sistemas de suporte à decisão (Grob et al., 2004). De acordo com a Conferência Internacional sobre Mineração de Dados Educacionais (*International Conference on Educational Data Mining*), as aplicações primárias da MDE são (Yacef et al., 2012):

- Predição de performance estudantil
- Modelagem do estudante
- Detectar comportamentos indesejáveis dos alunos
- Análise e visualização de dados
- Fornecimento de feedback para apoiar os instrutores

- Construção de material didático
- Planejamento e programação
- Recomendações para estudantes
- Categorização de alunos
- Análise de redes sociais
- Desenvolvimento de mapas conceituais

Como o objetivo deste trabalho é entender o comportamento das RSEs a partir da caracterização dos grupos (comunidades), a seguir são detalhadas as aplicações de MDE em **categorização de alunos** e **análise de redes sociais**.

2.1 Aplicação de MDE para Categorização de Alunos

O objetivo é a criação de grupos de alunos a partir da personalização de suas informações e/ou análises de características pessoais. Dessa forma, grupos de estudantes identificados podem ser utilizados na construção de sistemas personalizados, promovendo um aprendizado direcionado aos grupos, fornecendo conteúdos adaptáveis, entre outras aplicações. As técnicas de MD aplicadas nessa tarefa são: classificação (aprendizado supervisionado[1]) e agrupamento (aprendizado não supervisionado[2]). A análise dos agrupamentos são realizados sobre os subconjuntos de membros identificados, denominados *clusters*, atribuindo-se um conjunto de observações para classificação desses.

Diferentes algoritmos de agrupamento já foram utilizados para tarefa de identificação de grupos de estudantes, tais como: algoritmo de agrupamento aglomerativo hierárquico (*Hierarchical Agglomerative Clustering*), K-means e agrupamento baseado em modelo (*Model-Based Clustering*), para identificação de grupos com perfis de competências semelhantes (Ayers et al., 2009); Tang and McCalla (2002) aplicaram um algoritmo de agrupamento baseado em grandes sequências generalizadas, para identificação de grupos de alunos com características de aprendizado semelhantes, baseando-se em seus padrões de caminho de passagem e no conteúdo de cada página que visitaram; e Talavera

[1] Visa identificar a qual grupo um determinado usuário pertence. Nessa tarefa, o modelo analisa o conjunto de usuários fornecidos, com cada usuários já contendo a indicação à qual grupos pertence, a fim de "aprender" como classificar um novo usuário.

[2] Visa identificar e aproximar os usuários similares. Essa tarefa difere da classificação pois não necessita que os usuários sejam previamente categorizados.

and Gaudioso (2004) realizaram um estudo de agrupamento baseado em modelo, para identificação automática de grupos úteis a partir de dados de Sistemas de Gestão de Aprendizagem (do inglês, *Learning Management System*), visando a obtenção de grupos a partir dos comportamentos dos alunos (mensagens respondidas, discussões, atividades colaborativas, entre outros).

Não diferentemente, vários algoritmos de classificação foram aplicados, para categorização de alunos, tais como: análise discriminante, redes neurais (*Neural Networks*), as florestas randômicas (*Random Forests*) e árvores de decisão (*Decision Trees*), para classificação de estudantes universitários (Superby et al., 2006); classificação e árvore de regressão , *chi-squared automatic interaction detection*, e aplicação do algoritmo C4.5 para a identificação automática de estilos cognitivos dos alunos (Lee et al., 2009); e a aplicação do algoritmo de classificação *K-nearest neighbor* (K-NN) combinado com algoritmos genéticos para identificação e classificação de formas de aprendizagem (Chang et al., 2009).

Como pode-se observar diversos trabalhos já foram realizados no contexto de categorização de alunos em ambientes educacionais virtuais, na Tabela 2.1 são sumarizados cada um desses trabalhos. Neste trabalho, diferentemente dos supracitados, analisa-se dados coletados de uma RSE (grafos), o que difere da análise exclusiva dos dados. Para identificação dos grupos utilizou-se um algoritmos de detecção de comunidades, como pode ser visto em detalhe no Capítulo 4.

Tabela 2.1: Sumarização dos trabalhos de MDE para categorização de alunos

Trabalho	Tipo de Caracterização	Objetivo
Ayers et al., (2009)	Agrupamento	Identificação de grupos de alunos com perfis de competências semelhantes
Tang and McCalla (2002)	Agrupamento	Identificação de grupos de alunos com características de aprendizado semelhantes
Talavera and Gaudioso (2004)	Agrupamento	Obtenção de grupos a partir dos comportamentos dos alunos
Superby et al., (2006)	Classficação	Classificação de estudantes universitários
Lee et al., (2009)	Classficação	Identificação automática de estilos cognitivos dos alunos
Chang et al., (2009)	Classficação	Identificação e classificação de formas de aprendizagem

2.2 Aplicação de MDE para Análise de Redes Sociais

Na analise de redes sociais (ARS), ou análise estrutural, os objetivos vão além da análise direta dos dados (atributos ou propriedades individuais), analisando-se as relações entre as pessoas (no caso das RSE entre alunos). Como já visto, uma rede social é considerada um grupo de pessoas, uma organização ou indivíduos sociais que estão ligados por relações sociais, como amizade, relações cooperativas, ou troca de informações (Freeman, 2006). Diferentes técnicas de MD têm sido aplicadas na extração das redes sociais em ambientes educacionais, porem a filtragem colaborativa é a mais comum.

Filtragem colaborativa ou filtragem social, é um método para realização de previsões automáticas (filtragem) dos interesse de um usuário, a partir das preferências de um grupo de usuários (colaboradores) (Solomonoff and Rapoport, 1951). Em ambientes educacionais virtuais, sistemas de filtragem colaborativa podem produzir recomendações pessoais calculando-se as similaridades entre as preferências dos alunos. Dessa forma, esta tarefa está diretamente relacionada com a atividade de recomendações para os alunos (Kotsiantis et al., 2003). Como exemplo de estudos que aplicaram filtragem colaborativa como técnica de MDE para ARS, podemos citar: Lemire et al. (2005), que aplicou filtragem colaborativa para criação de listas de recomendações em repositórios de objetos de aprendizagem, Drachsler et al. (2008) no desenvolvimento de um sistema de recomendação pessoal e Khribi et al. (2008) para recomendação de links relevantes a um aluno ativo.

Existem algumas outras técnicas de MD que foram aplicados para ARS, tais como apresentados nos trabalhos: De Chen et al. (2008), esses propuseram a mineração de redes sociais interativas para recomendação de parceiros de aprendizagem apropriados; e Farzan and Brusilovsky (2006) que aplicaram técnicas de análise de redes sociais e MD para análise de comunicações entre os alunos a partir de ferramentas de fóruns. Todavia, nenhum desses se relacionou tanto com a pesquisa realizada quanto o trabalho realizado por Reffay and Chanier (2003). Nesse foi medida a coesão de pequenos grupos na aprendizagem colaborativa a distância, aplicando visualização e agrupamento sobre os grafos gerados a partir de fórum de discussão.

No presente trabalho, baseando-se nos atributos individuais de cada usuário procurou-se caracterizar os grupos sociais identificados, após a execução do algoritmo de detecção de comunidade. A partir dos resultados obtidos tentou-se compreender o comportamento das RSEs em estudo, buscando justificar a formação dos grupos e caracterizando cada um deles.

2.3 Considerações Finais

Nesse capítulo foi detalhado um pouco sobre a MDE, mais precisamente as técnicas voltadas para análise de grupos e de redes sociais. Foi visto que a MDE tem como principal foco no desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. E que através dessa, pode-se compreender de forma mais eficaz e adequada os alunos, como eles aprendem, o contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem.

A comunidade de MDE vem crescendo rapidamente. Em 2008 criou-se a Conferência Internacional sobre Mineração de Dados Educacionais (International Conference on Educational Data Mining), após uma sequência de worshops bem sucedidos realizados anualmente desde 2004. Contudo, no Brasil ainda são poucos os trabalhos publicados nesta área de pesquisa (Baker et al., 2011). Um dos trabalhos pioneiros no uso de mineração de dados na educação foi publicada por Brandão et al. (2006) analisando dados do programa nacional de informática na educação. Um outro trabalho pioneiro no Brasil que analisou dados da avaliação de alunos é apresentada por Pimentel and Omar (2006).

No presente trabalho, técnicas de mineração de dados são utilizadas no contexto de uma RSE. Destaca-se que o trabalho se insere dentro da convergência de duas tendências da área de Informática na Educação: a mineração de dados educacionais e análise de redes sociais educacionais. Salientando-se, ainda, que o entendimento das estruturas sociais em um RSE pode auxiliar na navegação da rede, visualização e análise, tornando possível um aprendizado direcionado aos grupos, assim como a realização de repasse de conteúdos adaptáveis.

3
Redes Sociais

Do you wish to rise? Begin by descending. You plan a tower that will pierce the clouds? Lay first the foundation of humility.

—ST. AUGUSTINE

Este capítulo tem com objetivo apresentar alguns conceitos sobre as redes sociais, tais como: evolução história, principais modelos, medidas de caracterização, algoritmos de detecção de comunidades e abordagens de caracterização de grupos. Fundamentos relevantes para o entendimento do trabalho de modo geral.

3.1 Grafos

Em 1736, ao propor uma rigorosa demonstração matemática para o problema das pontes de Kögnisberg[1], o matemático suíço Leonhard Euler deu início ao ramo da matemática conhecido como Teoria dos Grafos, que constitui a base para estudos acerca das redes em geral (Biggs, 1976; Barabási, 2009). Segundo Bollobas (1998), Euler teria criado o primeiro grafo da história.

Em teoria dos grafos, um grafo é uma representação abstrata de um conjunto finito de objetos, no qual alguns deles estão conectados uns aos outros por meio de um conjunto finito de links (ver Figura 3.1). A representação de um grafo é dada por:

$$G = (V, A) \tag{3.1}$$

[1] O problema é baseado na cidade de Kögnisberg (território da Prússia até 1945, atual Kaliningrado). No início do século XVII, o rio Pregel - que se estendia por toda a cidade - possuía sete pontes que o cruzavam. Essas serviram de base para a formulação do seguinte questionamento: seria possível atravessar todas as pontes sem passar mais de uma vez em cada ponte?

onde V é um conjunto de vértices ou nós não vazio, enquanto A é o conjunto de pares não ordenados de V, denominados arestas ou links (Biggs, 1976).

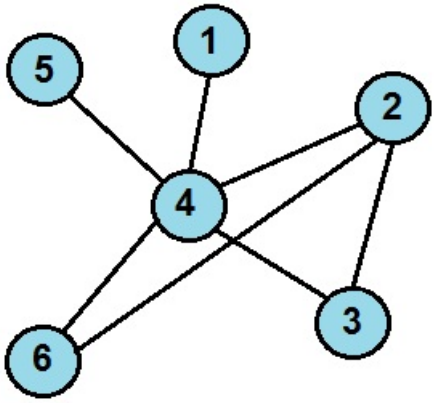

Figura 3.1: Representação gráfica de um grafo. Os círculos representam os nós e as linhas que os conectam, as arestas.

Redes geralmente são modeladas por meio de grafos, visto que esses possuem muitas características que podem ser associadas a diferentes propriedades estruturais de uma rede, além de possuírem um arcabouço matemático com o qual muitas dessas propriedades podem ser mensuradas e provadas (Wasserman and Faust, 1994).

Quando representadas por meio de grafos, os nós representam as entidades (membros ou atores) da rede e as arestas indicam como essas entidades se relacionam entre si (Deo, 1974). Por exemplo, em uma rede social educacional, os nós são os alunos e professores que fazem parte dela e as arestas são os laços gerados a partir da interação social (amizade, relacionamento por interesses educacionais, consultorias, relacionamentos de trabalhos, entre outros) entre eles.

Nos grafos, diversas propriedades podem ser associadas aos vértices e as arestas, tais como peso, sentido das ligações e tipos de vértices. Dessa forma, as arestas podem ou não ser direcionadas, onde grafos com conexão direcionadas são chamadas de dígrafos (por exemplo, uma rede que representa citações de trabalhos científicos). Existem também grafos com mais um tipo de vértice, como exemplo nas RSE, onde podemos ter alunos e professores. Por fim, as arestas podem ter valoração ou peso. Relações que consideram apenas a ausência ou presença de relacionamento são chamadas de dicotômicas, já quando se associa valores discretos ou contínuos as arestas, são denominadas valoradas (Silva et al., 2007).

Outro ponto de estudo da teoria dos grafos é o particionamento de grafos. Seja $G = (V,A)$ um grafo, $G' = (V',A')$ é dito um sub-grafo de G se $V' \subset V$ e $A' \subset A$, ou seja,

um subgrafo trata-se de um grafo cujo seus vértices e arestas estão contidos em outro grafo. Um particionamento de grafo, representado por um vetor P, é o conjunto de subgrafos distintos que unidos formam um grafo. Os subgrafos que formam o particionamento são chamados de partições, e o valor P_i refere-se a partição a qual o vértice i pertence. São denominadas arestas de *corte*, as arestas que ligam vértices de diferentes partições. O corte de um partição, $EC(P)$, é igual a soma dos pesos das arestas de corte que ligam vértices de P a outras partições (Almeida, 2009).

A representação gráfica de um grafo é útil para entendimento e visualização de algumas de sua propriedades. Todavia, quando o grafo é muito grande e complexo, esse tipo de representação não é muito adequada. Salientando, ainda, a necessidade de outras representações para que o processo de análise de redes seja realizado automaticamente e corretamente por um computador.

Dentre as principais representações utilizadas para armazenar dados de um grafo, tem-se: matriz adjacência e lista de adjacências. Em uma matriz de adjacência, o valor de uma célula a_{ij} representa a ligação entre os vértices i e j. Para redes sem peso, o valor de $a_{ij} = 1$, quando exite um relacionamento (aresta) entre os vértices i e j, ou $a_{ij} = 0$ caso contrário. Em caso de redes ponderadas os valores dos pesos são armazenados em uma matriz de pesos. Quando o grafo apresenta um número muito elevado de vértices em relação as arestas, a matriz adjacência tende a ser esparsa (muitas células a_{ij} têm valor 0), resultando em uma alocação de memória desnecessária. Como alternativa para essas situações utiliza-se listas de adjacências; essas permitem uma economia considerável de espaço, pois só são armazenadas as ligações existentes entre os pares de vértices. Todavia, o custo de acesso é maior devido a necessidade de busca-las em uma estrutura de lista.

Recentemente uma nova área de pesquisa surgiu para analisar propriedades estatísticas de "grandes" grafos (milhares e até milhões de vértices). Essa nova área é conhecida como análise de redes complexas, e é delineada a seguir.

3.2 Análise de Redes Complexas

A teoria das redes complexas não só estende o formalismo da teoria dos grafos, mas principalmente, propõe medidas e métodos fundamentados em propriedades reais do sistema (Costa et al., 2007). As redes complexas apresentam um comportamento multidisciplinar e estão presentes em aplicações de áreas distintas, utilizando a computação com ferramenta para modelagem, tratamento e análise dos dados. A seguir são apresentadas algumas categorias de rede:

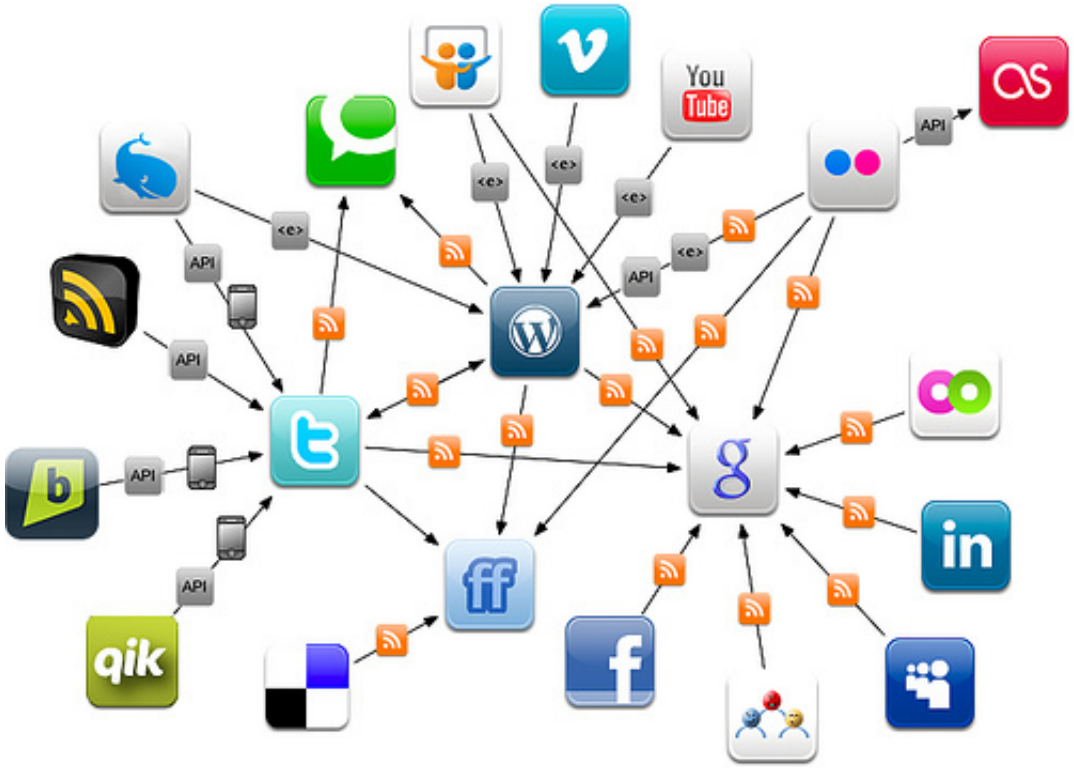

Figura 3.2: Exemplos de redes sociais existentes e suas interações

- **Sociais**: Redes que representam relações entre pessoas. Onde as pessoas são representadas por vértices e as relações entre elas por arestas. A Figura 3.2 apresenta um grafo referenciado as relações entre as pessoas nas diversas redes sociais existes, caracterizando o aumento das interações e a elevação do fluxo de informações proporcionado pelas redes sociais;

- **Tecnológicas**: Redes desenvolvidas para distribuição de algum bem ou recurso, tais com redes de distribuição de água, eletricidade e telefonia;

- **Biológicas**: São redes construídas para modelar sistemas biológicos. Uma exemplo são as rede metabólica, que modela o mecanismo de interação entre os substratos do metabolismo celular, tais como ATP, ADP e H_2O (nós) por meio de reações químicas das quais eles participam (arestas);

- **de Informação:** Redes obtidas a partir de bases de conhecimento formal, tais como as citações de artigos científicos e a *World Wide Web*. Como exemplo, tem-se a rede de citação, onde os nós representam trabalhos científicos publicados e as arestas evidenciam a referência de um artigo a outro previamente publicado.

Medidas de Caracterização

As redes do mundo real são geralmente estruturas complexas, isto é, não são governadas por fórmulas determinísticas (redes regulares). Essas podem apresentar diversas topologias e características conforme o domínio estudado, aplicando-se algumas medidas, as redes podem ser analisadas, caracterizadas e modeladas (Costa et al., 2007). Nesta seção, serão apresentadas medidas utilizadas para quantificar algumas das principais propriedades locais e globais de uma rede/grafo.

- **Centralidade:** A análise de centralidade facilita o entendimento sobre o fluxo de informação dentro de um grafo, pois através dela é possível verificar quais os vértices têm maior importância dentro do grafo, ou seja, os nós mais importante para o fluxo de informação. De acordo com Freeman (1979), a centralidade apresenta três características que levam a três conceitos diferentes de centralidade, a seguir será apresentado cada um individualmente.

  - **Grau:** Também chamado de valência, o grau de um nó representa o número de arestas incidentes a ele. Na Figura 3.1, por exemplo, o vértice 4 apresenta o maior grau (igual a 5), pois ele está diretamente conectado a todos os outros vértices da rede, nós 1,2,3,5 e 6. Em outras palavras, o grau de um nó nada mais é do que o número de nós diretamente conectados a ele (nós vizinhos), conforme a equação:

$$grau(v_i) = |\Gamma(v_i)| \tag{3.2}$$

    Onde $\Gamma(v_i)$ representa o conjunto de nós vizinhos ao nó v_i e $|\Gamma(v_i)|$ mede a cardinalidade desse conjunto, ou seja, representa o número de vizinhos do nó v_i. Para análise de grau de toda a rede tem-se, *grau médio*, essa trata-se de uma medida simples para caracterização de redes, onde é determinado o número de médio de ligações entre o vértices de uma rede (ver Equação 3.3).

$$\langle deg \rangle = \frac{1}{n} \sum_{i=1}^{n} grau(v_i) \tag{3.3}$$

    A análise do grau e de sua distribuição permite determinar se a construção da rede obedece à alguma lei de formação ou tem caráter aleatório.

  - **Proximidade:** O grau de proximidade (do inglês, *closeness centrality*) de um vértice v é a média dos caminhos mínimos desse vértice para qualquer

outro vértice da rede (Clauset et al., 2008). Ou seja, essa medida indica quão próximo um nó está do restante da rede. As proximidades mais baixas são observadas nos vértices centrais, pois esses possuem menor distância em média para os demais. É calculado pela soma dos inversos das Distâncias Geodésicas (*dist*) de um nó em relação aos demais nós do grafo. A centralidade de proximidade, portanto pode ser descrita pela equação:

$$proximidade(v_i)^{-1} = \sum_{v_j \in V, v_j \neq v_i} dist(v_i, v_j) \tag{3.4}$$

Onde $dist(v_i, v_j)$ representa a distância geodésica entre os nós v_i e v_j.

– **Intermediação:** O grau de intermediação (do inglês, *betweenness centrality*) de um vértece v é dado pela quantidade de caminhos mínimos, entre qualquer par de vértice da rede, que passa por ele. Portanto, essa medida analisa a frequência com que um nó aparece no caminho geodésico entre dois outros nós quaisquer, ou seja, procura identificar nós com grande potencial de controle do fluxo de informação na rede. Pode ser descrito pela equação:

$$intermediacao(v_i) = \sum_{j<k} \frac{g_{jk}(v_i)}{g_{jk}} \tag{3.5}$$

Onde g_{jk} é o número de caminhos geodésicos que ligam os nós v_j e v_k e $g_{jk}(v_i)$ é o número de caminhos geodésicos do total g_{jk} que unem os nós v_j e v_k passando pelo nó v_i em algum ponto.

- **Diâmetro:** O diâmetro é a maior distância entre dois nós de um grafo. Essa medida é identificada calculando-se a menor distância entre todos os pares possíveis de vértices (Wasserman and Faust, 1994). Após identificação dos dados, o diâmetro é o par de vértices que possuir maior distância.

- **Densidade:** A densidade é uma medida bastante utilizada no estudo de comunidades. Ela indica o grau de coesão entre os membros da rede, isto é, o quão conectadas estão as entidades com relação ao número máximo de possíveis conexões na rede. Portando, quanto mais densa (conectada) for a rede, melhor o fluxo de informação entre as entidades que a compõem (Wasserman and Faust, 1994). A equação da densidade de um grafo (G) é dada por:

$$densidade(G) = \frac{2l}{n(n-1)} \tag{3.6}$$

onde l e n são respectivamente, os números de *links* e de nós da rede. Caso o grafo seja direcionado, as arestas só são contabilizadas em uma direção, por isso, não é feita a multiplicação por 2 na equação acima.

Uma rede com densidade máxima, isto é, todos os seus membros estão conectados entre si, é denominada *Clique*. É importante ressaltar que mesmo que uma rede esteja distante de ser um clique, essa pode conter diversos subgrafos (comunidades) que são cliques, caracterizando um distinto grupo de nós (altamente conectados) em relação ao restante da rede.

- **Coeficiente de Agrupamento:** Também conhecido como **coeficiente de aglomeração** (do inglês, *clustering coefficient*), introduzido por Watts and Strogatz (1998), essa mediada verifica quão próximos os vizinhos de um nó estão de formar cliques, expressando a presença de ciclos mínimos de uma rede, ou seja, ciclos com apenas três vértices formando triângulos. Em outras palavras, mede a transitividade[2] ao redor de um vértice. Em uma rede real, de amizades por exemplo, esse medida seria basicamente a probabilidade de dois amigos quaisquer, terem um amigo em comum. O coeficiente de agrupamento CC_v é dado pela Equação 3.7.

$$cc_v = \frac{2e_i}{deg_v(deg_v - 1)} \tag{3.7}$$

onde v é um vértice qualquer que possui deg_v vizinhos, e e_i é o número de arestas que realmente existe entre os deg_v vizinhos. Caso esses deg_v formassem um clique, o número de arestas entre eles seria: $deg_v(deg_v - 1)/2$.

Para cálculo de média de agrupamento entre todos os vértices tem, o **coeficiente de agrupamento médio** ($\langle cc \rangle$), representado na Equação 3.8.

$$\langle cc \rangle = \frac{1}{n} \sum_{i=1}^{n} cc_i \tag{3.8}$$

Na literatura podem ser encontradas outras medidas de caracterização, apresentadas em detalhes nos trabalhos de Newman (2003) e Barabási (2009).

[2] A transitividade ocorre quando um vértice A está conectado a um vértice B, e este, por sua vez, está conectado a um vértice C, aumentando as chances de A também estar conectado a C. No caso particular de um grafo direcionado, a verificação de transitividade deve obedecer a orientação das arestas.

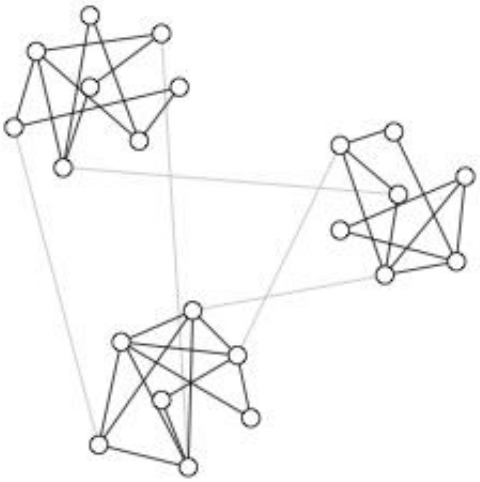

Figura 3.3: Formação de comunidades. Os traços mais escuros indicam as ligações entre membros da comunidade (conexões entre nós de um mesmo aglomerado). Os traços claros indicam como as comunidades se relacionam entre si. [Fonte: (Girvan and Newman, 2002)]

3.3 Detecção de Comunidades

Uma propriedade bastante comum em redes complexas é a presença de estruturas modulares chamadas de grupos/comunidades. As comunidades são aglomerados formados por grupos de nós altamente relacionados com outros nós pertencentes ao mesmo aglomerado, mas pouco densos com relação aos relacionamentos que possuem com nós de aglomerados vizinhos (Girvan and Newman, 2002) (ver Figura 3.3). Hoje, sabe-se que a clusterização está presente na Web, nas redes físicas de computadores na Internet, nas cadeias alimentares, nas redes moleculares, nas redes sociais, entre outras; ou seja, trata-se de uma propriedade genérica das redes complexas (Barabási, 2009).

Embora as comunidades possam estar presentes em distintas redes, a sua conotação varia com o tipo da rede em questão. Em uma rede social, podem conotar subgrupos sociais com interesses em comum por exemplo (Girvan and Newman, 2002); enquanto que em uma rede metabólica, está mais relacionada a formação de grupos funcionais (Yuruk et al., 2007).

De acordo com Wasserman and Faust (1994), um grupo (comunidade) é um conjunto de usuários que interagem uns com os outros com frequência. Diversos estudos e aplicações já foram desenvolvidas para a identificação de grupos, incluindo: visualização de rede, análise de inteligência (Baumes et al., 2004), Compreensão de rede (Sun et al.,

2007),estudo comportamental (Tang and Liu, 2010a), segmentação e recomendação (Wang et al., 2010), e filtragem colaborativa (Chen et al., 2009).

Paralelamente também vem sendo desenvolvidos diversos métodos para detecção de comunidades em redes complexas (Wasserman and Faust, 1994), dentre eles podemos citar: Algoritmo de Girvan and Newman (2002), *Fast greedy modularity optimization* de Clauset et al. (2008), *Fast modularity optimization* de Blondel et al. (2008), entre outros. A seguir o algoritmo *Fast modularity optimization* de Blondel et al. (2008), também conhecido como algoritmo *Multi-Level Aggregation Method*, é detalhado. Esse foi método utilizado no estudo para identificação das comunidades.

3.3.1 *Multi-Level Aggregation Method*

O *Multi-Level Aggregation Method* (MAM) é um método de passos múltiplos, baseado na modularidade local otimizada (Girvan and Newman, 2002). De acordo com Newman (2006), a modularidade é uma função benefício utilizada na análise de redes ou grafos, tais como as redes de computadores ou as redes sociais. Ela quantifica a qualidade das partições de uma rede em módulos ou comunidades. Boas partições, com valores elevados de modularidade (a modularidade de uma partição é um valor escalar entre -1 e 1), são aqueles que apresentam conexões internas densas entre os nós dentro dos módulos, mas apenas ligações esparsas entre diferentes módulos. Segundo Newman (2003), a modularidade também pode ser usada como uma função de otimização.

O algoritmo de detecção de comunidade MAM é dividido em duas fases, que são repetidas iterativamente. Na primeira fase a partir de uma rede de N nós, como primeiro passo, cada nó da rede recebe uma comunidade diferente. Logo, na partição inicial a quantidade de comunidades é igual ao número de nós existentes. Então, para cada nó i são considerados os seus j nós vizinhos, para avaliação do ganho de modularidade que ocorreria na remoção de i da sua comunidade, e colocando-o na comunidade de j. Dessa forma o nó i é colocado na comunidade que obteve o maior ganho positivo. Todavia, se nenhum ganho positivo é possível, i fica em sua comunidade de origem. Esse processo é aplicado repetidamente e sequencialmente para todos os nós até que nenhuma melhoria possa ser alcançada e a primeira fase é, então, concluída.

A segunda fase do método MAM consiste na construção de uma nova rede cujos nós agora são as comunidades encontradas durante a primeira fase. Para realização dessa, novos pesos são definidos para as ligações entre os nós, dado pela soma dos pesos das ligações entre os nós correspondentes de ambas as comunidades (Blondel et al., 2008). Após a conclusão da segunda fase, é possível reaplicar a primeira fase do algoritmo para

a rede ponderada resultante, e iterar novamente. A combinação dessas duas fases será denominada por "passagem". Por construção, a quantidade de comunidades diminui em cada passagem, e como consequência a maior parte do tempo de processamento é usada na primeira fase. As passagens são iteradas (ver Figura 3.4) até que não haja mais mudanças e a máxima modularidade é atingida. O algoritmo de MAM, ainda, resulta em uma hierarquia de comunidades. A altura da hierarquia gerada é determinada pelo número de passagens (geralmente é um número pequeno).

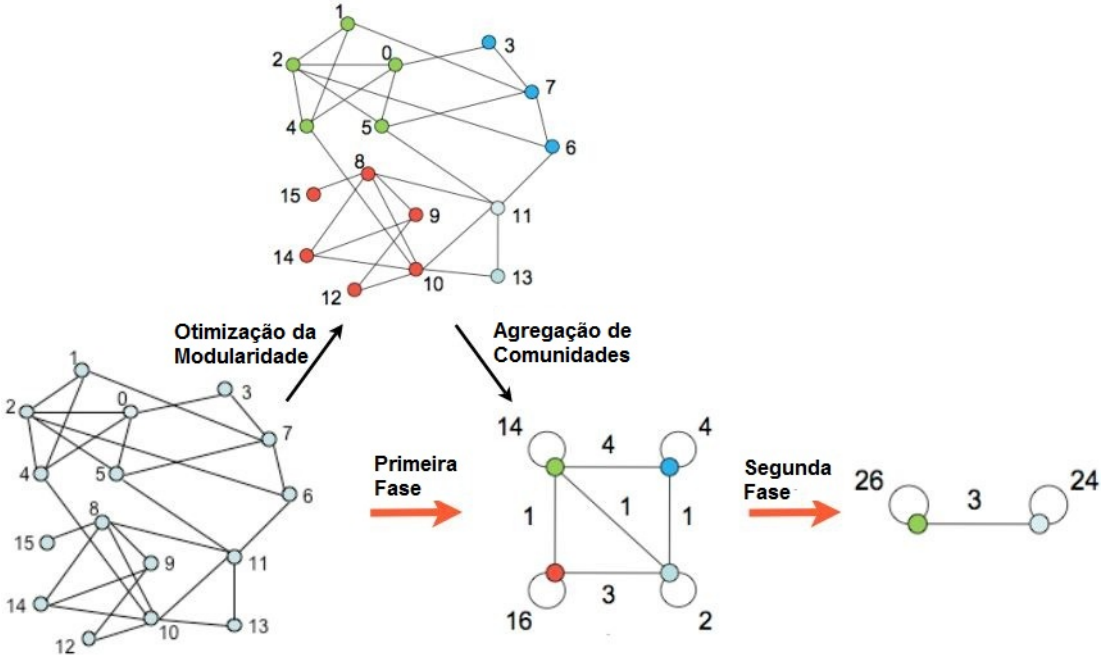

Figura 3.4: Visualização dos passos do Método MAM. Cada passo é realizado em duas fases: na primeira a modularidade é otimizada, permitindo somente alterações locais das comunidades; já na segunda as comunidades encontradas são agregadas para a construção de uma nova rede de comunidades. Esses passos são repetidos iterativamente até que nenhuma aumento de modularidade seja possível e uma hierarquia de comunidades seja produzida. [Fonte: (Blondel et al., 2008)]

Quando o valor máximo de modularidade local é alcançado durante a optimização, isso corresponde ao número de agrupamentos a partir das comunidades formadas. Uma vez que geralmente existem vários valores de máxima modularidade local disponíveis durante a fase de optimização, esses números de fragmentação sob diferentes valores de modularidade, dessa forma, podem ser considerados níveis de agrupamento (resoluções) diferentes. Portanto, o número aproximado de comunidade pode ser encontrado a partir dos diferentes níveis de agrupamento, pois o valor máximo de modularidade global pode ser encontrado entre os valores máximos de modularidades locais. Dessa forma, o

MAM fornece um esquema de heurística que possibilita a identificação do número de comunidades automaticamente.

Em um estudo comparativo realizado por Fortunato and Lancichinetti (2009), foram analisados diversos métodos de detecção de comunidades. De acordo com autor, o método de MAM oferece um compromisso equilibrado entre a precisão da estimativa da máxima modularidade, que é melhor que técnicas gananciosos como *Fast greedy modularity optimization* (Clauset et al., 2008), e uma complexidade computacional, que é basicamente linear ao número de Links do grafo. O autor ainda conclui que o MAM está entre os métodos que apresentaram melhor desempenho sobre redes não direcionadas e não ponderadas (como é o caso da OJE, RSE utilizado no presente trabalho). Além de ser muito rápido, essencialmente em redes de tamanho linear. De fato, apesar de simples esse algoritmo tem várias vantagens, dentre elas podem ser referenciadas, os passos intuitivos e de fácil implementação, e o resultado ser não supervisionado (Blondel et al., 2008).

3.3.2 Caracterização de Grupos Sociais

De acordo com Tang et al. (2011), o estudo de caracterização de grupos sociais (do inglês *Group Profiling*) visa a construção de um perfil descritivo para um dado grupo fornecido. Diversas abordagens veem sendo desenvolvidas para realização desse estudo.

Alguns trabalhos pioneiros tentam compreender a formação de grupos com base na análise estatística da estrutura da rede. Backstrom et al. (2006) estudaram algumas questões básicas sobre a evolução dos grupos, tais como: quais são as características estruturais que determinam qual grupo um indivíduo vai aderir? Os autores descobriram que a quantidade de amigos no grupo é um fator relevante para determinar que um novo usuário junte-se ao grupo. Isso proporciona um nível global de análise estrutural, facilitando o entendimento sobre como as comunidade atraem novos usuários. Já Leskovec et al. (2008) observaram que o agrupamento espectral (um método popular usado para a detecção da comunidade) sempre encontra, mesmo que em pequena escala, comunidades triviais, ou seja, comunidades que estão conectadas ao restante da rede através de uma única aresta. Os dois trabalhos focam em uma imagem global (estatística) das comunidades. Dessa forma, mais pesquisas são necessárias para entender a formação de grupos específicos.

A caracterização de grupos sociais pode facilitar a compreensão das comunidades implícitas extraídas com base na estrutura da rede, bem como as comunidades explícitas formadas por assinaturas de usuários. Diversas são as aplicações para caracterização de grupos, dentre essas podem ser citadas: entendimento das estruturas sociais, visualização e

navegação da rede, acompanhamento a mudanças de temas de um grupo, casos alarmantes e marketing direto (Tang et al., 2011).

Analisando-se a aplicação em *marketing* direto, é possível que os consumidores (clientes) *online* formem diversos grupos, e que cada grupos possua um conjunto de diferentes mensagens, comentários e opiniões sobre os produtos. Uma vez que um perfil pode ser gerado para cada grupo, as empresas podem conceder novos produtos de acordo com as características dos grupos, sugerir produtos direcionados aos grupos, entre outras oportunidades (Wang et al., 2010).

De acordo com Kumar et al. (2006) uma rede pode ser dividida em três regiões: *singletons* (nós que não interagem com outras pessoas), comunidades isoladas, e um componente gigante conectado. Comunidades isoladas realmente ocupam uma porção muito estável em toda a rede, e consequentemente a probabilidade de duas comunidades isoladas se unirem com a evolução da rede é baixíssima. Todavia, com a caracterização dos grupos disponíveis, é possível que uma comunidade isolada ou um *singleton* identifique outros grupos semelhantes, e realizem interação (conexões) de comunidades isoladas de interesses semelhantes. Baseando-se nesses possibilidades algumas abordagens veem sendo desenvolvidas para a realização da caracterização de grupos socais. A seguir serão detalhadas as abordagens mais relacionadas ao estudo realizado nesta dissertação.

Principais Abordagens

De acordo com Tang et al. (2011), é desejável que todo método de caracterização de grupos satisfaça as seguintes propriedades:

- Descritivo: Os atributos selecionados para caracterizar um grupo devem refletir o perfil do grupo, ou seja, o interesse comum ou afiliação.
- Robusto: Montanhas de dados são produzidas a cada dia em mídias sociais. Estes dados tendem a ser muito ruidoso. O método de caracterização de grupo deve ser robusto aos ruídos.
- Escalável: Em mídias sociais, uma rede com grandes dimensões é normal. A cada dia, novos usuários são inseridos na rede, e novas conexões ocorrem entre os nós existentes. Dessa forma, os usuários envolvem-se em várias atividades e interações, produzindo uma grande quantidade de conteúdos ricos em informações relevantes sobre os usuários. Indiretamente isso também representa um grande desafio para os estudos de caracterização de grupos.

A seguir serão apresentadas as principais abordagens de caracterização de grupos relacionadas a pesquisa realizada.

- **Caracterização de grupos baseado em agregação**

  Sabendo-se que o estudo de caracterização de grupos tem como objetivo encontrar as características que são compartilhadas por todo o grupo, uma abordagem natural e simples seria encontrar as características (atributos) que são mais prováveis de ocorrer dentro do grupo. Esse é o objetivo essencialmente da abordagem de caracterização de grupos baseada em agregação.

  A abordagem de caracterização de grupos baseado em agregação (CGBA) é simples, basicamente atributos individuais do grupo são agregados e um *ranking* desses é gerado, escolhendo os top-k (k trata-se do número limite de atributos, ou seja, k=10 seria selecionados os 10 principais atributos) atributos mais frequentes (Tang et al., 2011). Observando-se os atuais sistemas de etiquetagem, é verificado que a abordagem de CGBA é amplamente utilizada, na forma de nuvens de *tags*. Nuvens de *tags* são amplamente utilizados em mídias sociais para mostrar a popularidade de uma *tag* (atributos, marca, produto, palavras mais frequentes em fóruns, entre outros) pelo seu tamanho da fonte. Considerando-se toda a rede, como um grupo, em seguida, uma nuvem de marcação é produzida com base na agregação das características do grupo.

  Todavia, esse método pode ser sensível em algumas situações. Por exemplo, em um fórum de rede social algumas palavras que não contribuem para a caracterização de um grupo, pode vir a serem muito utilizadas. Isso resultaria em uma caracterização equivocado para o grupo analisado, não satisfazendo a propriedade em que os atributos selecionados para caracterização dos grupos devem ser "Descritivo".

- **Caracterização de grupos baseado em diferenciação**

  Em vez de agregar, podemos selecionar as características que diferenciam um grupo de outras pessoas na rede (restante da rede). Assim, a abordagem de caracterização de grupos baseado em diferenciação (CGBD) converte o problemas de caracterização de grupos em um problema de classificação com duas classes (Liu and Motoda, 1998). Dessa forma, os nós que fazem parte do grupo ficam na classe positiva e os demais ficam na classe negativa.

  Portanto, o objetivo da abordagem CGBD é descobrir as principais (top-k) características discriminativas que são representativas para um grupo, o diferenciando

do restante da rede. Ou seja, essencialmente a abordagem seleciona atributos que aparecem em um grupo e dificilmente aparece nos outros.

- **Caracterização de grupos baseado em diferenciação egocêntrica**

  Na abordagem de diferenciação anterior, todos os nós de fora de um grupo são considerados como pertencendo à classe negativa. No entanto, o cálculo baseado em diferenciação de grupos aplicado a uma visão global (considerando-se todos os nós de uma rede), pode ser considerado um desperdício de processamento. A escalabilidade também pode ser uma preocupação. Redes sociais online mais populares são enormes, com centenas de milhões de nós. Dessa forma, é necessário um elevado tempo de processamento para recuperação de todas as informações de uma rede do mundo real. Ressaltando que em algumas aplicações, somente uma vista egocêntrica está disponível. Em outras palavras, pessoas têm informações sobre os seus amigos, mas pouco conhecimento sobre as pessoas que não conhecem.

  Nesse contexto, Tang et al. (2011) propuseram a abordagem caracterização de grupos baseado em diferenciação egocêntrica (CGBDE), que em vez de diferenciar um grupo de toda a rede, o diferencia dos vizinhos[3] dos membros do grupo, ou seja, perfis de grupo com base na visão egocêntrica.

Tang et al. (2011) realizaram um estudo comparando as três abordagens de caracterização de grupos supracitadas. Para realização do estudo comparativo selecionaram como bases de dados os sites de mídia social: BlogCatalog[4] e LiveJournal[5]. Ao final do estudo foi descoberto que as abordagem baseadas em diferenciação (CGBD e CGBDE) sempre superavam a baseada em agregação (CGBA). Isso tornou-se mais perceptível quando os atributos individuais utilizados para caracterização apresentavam ruídos (como por exemplo: atributos coletados postagens de usuários, registros de atividades e relatos de usuários). Ou seja, a aplicação da abordagem de CGBA aplicada a dados ruidosos, selecionou uma grande quantidade de atributos que não descreviam o grupo corretamente.

Dessa forma, Tang et al. (2011) concluíram que a abordagem de CGBA somente é aplicável em um ambiente relativamente livre de ruído. Pois, se os perfis são construídos a partir de atributos ruidosos, como posts de usuários, registro de atividades de usuários ou relatos de interesses, abordagens baseadas em diferenciação, apresentam melhores

[3] Vizinhos de um grupo se referem aos nós fora de um grupo, que são conectados a pelo menos um membro do grupo.

[4] http://www.blogcatalog.com/

[5] http://www.livejournal.com/

perfis para caracterização dos grupos. Baseado nisso, este trabalho propõe um método de CGBD, a partir da aplicação do teste estatístico *Wilcoxon Rank Sum*.

3.4 Redes Sociais Educacionais

O termo Web 2.0 ou Web Social refere-se ao novo panorama para o funcionamento da internet, centrado, principalmente, na atividade colaborativa, entre os internautas, mediada por sistemas muitas vezes interligados (Silva and Pereira, 2008). Neste contexto, surgiram as Redes Sociais online, serviço Web que permite aos indivíduos: (1) construir perfis total ou parcialmente públicos dentro de um sistema, (2) interagir com outros usuários com os quais se está conectado e (3) visualizar e percorrer listas de conexões feitas por usuários do sistema (Boyd and Ellison, 2010). As Redes Sociais online consistem em um tipo de software colaborativo (*groupware*), no qual as entidades são pessoas e as conexões representam os relacionamentos entre elas (Regueiro, 2009).

Esse tipo de aplicação vem se popularizando bastante. Como por exemplo, o *Facebook*[6], que segundo Mui and Whoriskey (2010) é atualmente o site mais visitado da web, superando o *Google*[7]. De acordo com NIELSEN (2009) as redes sociais online ultrapassaram o *e-mail* como serviço mais utilizado na Web. Motivados por essa popularização, a comunidade científica tem demostrado grande interesse na realização de estudos mais aprofundados sobre as redes sociais, abrindo espaço para diversos trabalhos sobre esses ambientes.

A análise e a extração de informação de redes sociais vêm sendo amplamente utilizadas em várias áreas, incluindo Ciências Sociais e Comportamentais, Economia e Marketing, em que a compreensão do comportamento da sociedade é estratégica (Freitas et al., 2008). De acordo com um estudo divulgado em 2010, noventa por cento das empresas no Brasil, por exemplo, usam as Redes Sociais para cativar novos negócios. O estudo revela que no nível global a média de empresas que já descobriram o potencial das redes sociais para gerar negócios é um pouco menor, em torno dos setenta e cinco por cento. Tais empresas utilizam as Redes Sociais para se relacionar com os clientes, coletar dados sobre sua opinião e preferências e realizar divulgação de produtos e serviços (TERRA, 2010)

Com o sucesso das mídias sociais, muitos serviços de redes sociais foram desenvolvidos, entre esses, as Redes Sociais Educacionais (RSE). Analisando-se bem, a estrutura

[6]www.facebook.com
[7]www.google.com

social do ensino se adéqua perfeitamente ao conceito das redes sociais. As entidades são compostas por professores e estudantes; as conexões, por relações educacionais como cursos, consultorias, grupos de trabalho interdisciplinares, entre outras. A sala de aula é uma pequena sociedade formada pelo professor e seus alunos. Portanto, deve ser um bom lugar para colaboração e trabalho conjunto (Purcell, 2010).

As RSE têm como principal proposito fazer uso das tecnologias de redes sociais para fins educacionais (Purcell, 2010). Diferentemente das redes sociais, os sites de RSE são locais mais seguros pois a grande maioria são privados (apenas membros convidados podem visualizar), alem de ser usualmente gratuitos. Oferecendo aos alunos a oportunidade de interagir com outros alunos, professores, administradores, sem limitações geográficas.

Pesquisadores apoiam as redes sociais pela sua capacidade de atrair, motivar e envolver os alunos em práticas significativas de comunicação, como troca de conteúdos e colaboração (Mills, 2011). No Brasil as RSE veem crescendo e sendo bastante utilizadas, como exemplos: as Olimpíadas de Jogos Digitais e Educação (OJE[8]) e *Redu Educational Technologies* (REDU[9]).

Inúmeros são os benefícios obtidos com a integração de tecnologias de redes sociais a educação para alcançar objetivos de aprendizagem, incluindo, mas não limitados a eles seguem (Purcell, 2010):

- **Conhecimento e desenvolvimento de habilidades:** Atividades planejadas incorporadas as RSE podem aumentar a profundidade do conhecimento nos conteúdos das áreas relacionadas, elevando ainda, as habilidade no manuseio de tecnologias.

- **Motivação:** Sites de RSE atraem alunos a completarem as atividades aumentando seus interesses. Trabalhos realizados pelos alunos serão visto por um público autêntico, em vez de serem simplesmente entregues em forma de trabalhos aos professores. Ter esse público tão grande pode incentivar os alunos a fazerem o seu melhor para garantir bons trabalhos.

- **Conectividade:** As redes sociais possibilitam estudantes se conectarem com outros estudantes e especialistas em qualquer parte do mundo.

- **Familiaridade:** As redes sociais já são familiares aos alunos, assim, o uso dessas ferramentas podem facilitar a passagem de conhecimento.

[8] www.joystreet.com.br/products/oje
[9] www.redu.com.br/

- **Custo benefício:** Redes sociais apresentam um bom custo benefício, pois muitas são gratuitas, necessitando apenas de acesso a internet. Evitando que organizações com um menor poder aquisitivo, gastem dinheiro com planos de aulas caros e recursos.

- **Conveniência:** Sites de redes sociais são acessados a qualquer momento através da internet, dessa forma o processo de aprendizagem pode continuar após as aulas.

- **Aumento do uso apropriado:** Ensinando-se os alunos a usarem as redes sociais de forma a adquirirem mais conhecimento, consequentemente estes terão mais responsabilidades no uso pessoal das redes sociais.

- **Aumento da eficiência:** Os alunos têm acesso às informações quando necessário. Assim podem acessar as informações quando estão ausentes ou depois do horário da escola.

- **Melhora a forma de como se Expressar:** Estudantes gostam de ser ouvidos, e a *web* torna possível a divulgação em uma escala de audiência mundial. Além das redes sociais possibilitarem as conexões entre grandes comunidades de usuários que apresentam interesses comuns, apresentando-se como público autêntico.

- **Aumento do trabalho em equipe e cooperação:** Realizar as tarefas em grupos ou correspondentes, com estudantes em locais diferentes pode melhorar as habilidades de trabalho em grupo e a cooperação.

3.5 Considerações Finais

Neste capítulo, foram apresentados os principais conceitos sobre as redes complexas. A análise de redes complexas é uma das áreas que surgiram recentemente com o intuito de estudar grandes volumes de dados modelados como redes (grafos). Algumas de suas principais medidas de caracterização foram discutidas nesse capítulo.

Uma propriedade relevante em diversas redes complexas é a existência de grupos de nós densamente conectados, denominados comunidades. O estudo de métodos para caracterização desses grupos tem despertados interesses da comunidade científica e trabalhos interessantes têm sido realizados. Neste capítulo foram descritas as principais abordagens de caracterização de grupos, entre elas a baseada em diferenciação (CGBD), a qual é um dos focos de estudo desse trabalho.

Este capítulo abordou os conceitos sobre as redes sociais relacionados ao estudo de caracterização de grupos sociais. Tais conceitos, são a base para desenvolvimento do método de CGBD, aplicando o teste estatístico de WRS. O próximo capítulo por sua vez se concentrará na apresentação e avaliação do método proposto neste trabalho.

4

Caracterização de grupos na rede OJE

Se você quer ser bem sucedido, precisa ter dedicação total, buscar seu último limite e dar o melhor de si mesmo. No que diz respeito ao empenho, ao compromisso, ao esforço, à dedicação, não existe meio termo. Ou você faz uma coisa bem feita ou não faz.
Trabalhei muito para chegar ao sucesso. Mas não conseguiria nada se Deus não ajudasse.

—AYRTON SENNA DA SILVA

Este capítulo apresenta os artefatos produzidos nesse trabalho e as considerações sobre eles. Ressaltando, este trabalho tem como objetivos principais:

- Realizar um levantamento teórico e uma análise dos estudos mais relevantes do estado da arte de mineração de dados educacionais (técnicas voltadas para análise de grupos e redes sociais) e caracterização de grupos, contribuindo assim para melhor entendimento dos fundamentos, aplicabilidade, limitações e tendências da área; bem como, identificação de métodos e técnicas já consolidadas pela comunidade acadêmica no tratamento de problemas relacionados;

- Identificação e extração dos atributos individuais dos usuários das redes sociais em estudo (OJE-PE e OJE-AC);

- Elaboração de um modelo para filtragem dos principais atributos caracterizadores dos usuários, representação dos usuários da rede;

- Validação dos atributos selecionados para representação dos usuários com especialista do domínio da rede em estudo;

- Implementação do algoritmo de detecção de comunidade *Multi-Level Aggregation Method*;
- Baseado nos atributos selecionados na representação dos usuário, aplicar o teste estatístico *Wilcoxon Rank Sum* para caracterização dos grupos identificados pelo algoritmo de detecção de comunidade.
- Avaliar o desempenho dos perfis caracterizadores dos grupos gerados, analisando as relações dos resultados entre as redes que o método proposto foi executado.

Como visto nos capítulos anteriores, um dos objetivos da análise de redes sociais (ARS) é compreender o processo de formação e evolução das redes, mais especificamente aquelas de caráter social. Caracterização de grupos, portanto tem papel fundamental nesse processo investigativo, pois será responsável por delinear as estruturas da rede, em termos de grupos, descrevendo-os em perfis e buscando justificar as suas formações.

Grande parte dos trabalhos da área estão embasados na identificação de caracterizadores de grupos através de métodos baseados em agregação, em sua grande maioria aplicando a abordagem de formação de nuvens de tags em mídias sociais. Conforme apontado por Tang et al. (2011), esses métodos não são viáveis em ambientes com ruído. Em tais situações, abordagens baseadas em diferenciação apresentam melhores resultados.

Neste contexto, foi desenvolvido um método baseada em diferenciação para geração de perfis de grupos, aplicando o teste estatístico *Wilcoxon Rank Sum* (Gomes et al., 2013). Focado na caracterização descritiva de grupos de pessoas a partir dos seus atributos individuais, o método é delineado em detalhes mais adiante. Na próxima seção a arquitetura do método é apresentada e todos os seus módulos são detalhados. Apresentando, ainda, as redes adotadas, experimentos e resultados obtidos pelo método proposto.

4.1 Arquitetura Geral

Nesta seção, é apresentado o método desenvolvido para a caracterização dos grupos sociais. O método é definido em três etapas: Inicialmente, é realizado o pré-processamento sobre a base de dados, identificando os atributos individuais dos usuários. Após o pré-processamento da base, os principais atributos dos usuários são selecionados: (1) fase de **Representação dos Usuários**. A partir desta representação é gerada a estrutura da rede, com seus nós e arestas. Posteriormente um algoritmo de detecção de comunidades é aplicado sobre a rede gerada para a identificação das comunidades, (2) fase de **Detecção**

de Comunidades. Finalmente, é iniciada a (3) fase de **Caracterização dos Grupos**, com a aplicação do teste estatístico *Wilcoxon Rank Sum* sobre a rede pré-processada na fase anterior. Na Figura 4.1, a arquitetura desenvolvida é apresentada. Nas subseções a seguir, serão apresentados e detalhados cada um dos seus módulos.

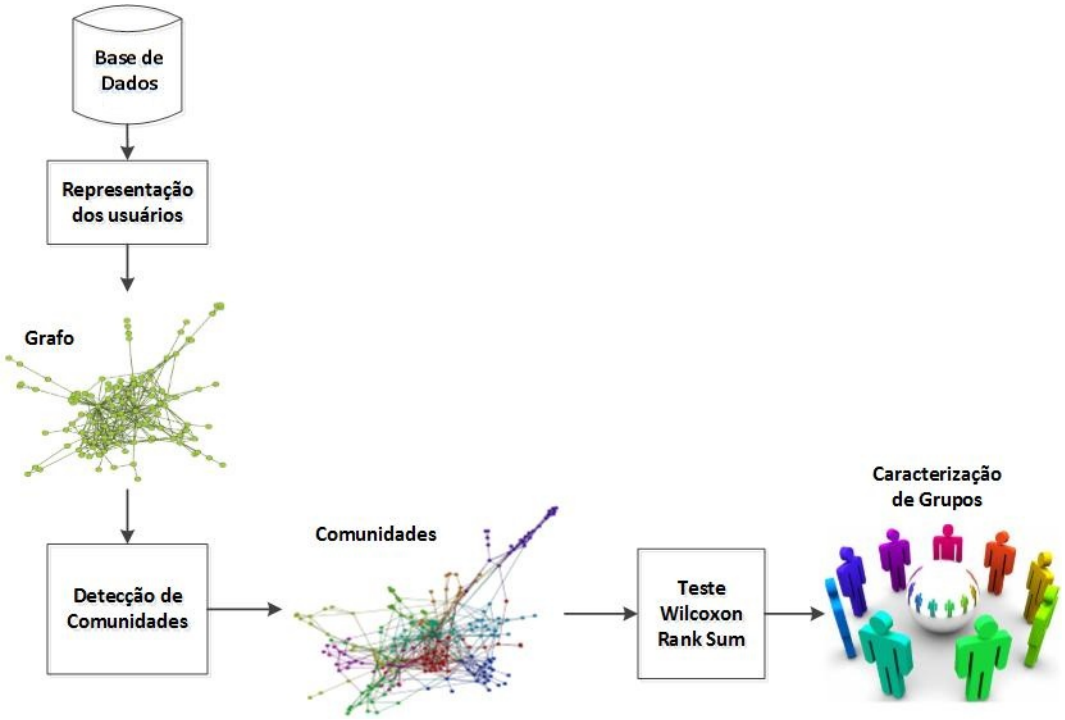

Figura 4.1: Arquitetura do método proposto

4.2 Redes Utilizadas

Para avaliar o desempenho do método proposto, duas redes sociais educacionais foram utilizadas nos experimentos realizados no contexto deste trabalho. Optou-se por esse tipo de rede porque, sendo uma rede social, é uma estrutura inerentemente evolutiva, o que faz dela uma fonte de dados ideal para a investigação da formações de grupos. Além disso, as RSE são propícias à avaliação de diversas tarefas relacionadas a ARS (dentre elas caracterização de grupos), pois trata-se de uma rede onde os usuários interagem, em sua grande maioria, motivados por interesse educacionais comuns. Conforme descrito na Seção 3.4, em uma RSE as entidades (nós) são compostos por professores e estudantes; e as conexões, por relações educacionais como cursos, consultorias, grupos de trabalho interdisciplinares, entre outras.

Para realização dos experimentos, selecionou-se como base de dados a Olimpíada de Jogos Digitais e Educação (OJE). Essa trata-se de um ambiente de RSE, que conecta alunos e professores através de vários desafios oferecidos aos usuários, na forma de jogos casuais e enigmas baseados nos conteúdos escolares do Ensino Básico. De acordo com Meira et al. (2009), a OJE foi desenvolvida inicialmente para o Estado de Pernambuco (Secretaria de Educação) atendendo inicialmente 18 mil alunos da rede pública de ensino do Estado de Pernambuco. Além de acesso à internet o sistema funciona como Lan Houses com o intuito de estimular a atenção dos jovens, assim, possibilitando os alunos gerarem suas próprias redes sociais. Algumas condições foram elaboradas para implementação da OJE:

- Inclusão dos conteúdos curriculares em um ambiente atraente para entrosamento dos alunos;
- Integração dos jogos entre si e dos jogos com práticas que pudessem ser realizadas sem o computador;
- Estimular a competição e colaboração entre alunos, escolar, comunidades;
- Utilização de mídias e tecnologias que atendam as limitações escolares;
- Os jogadores são adequados aos grupos relacionados à sua série;
- Gerenciar ritmo e gênero de acordo com a aprendizagem dos alunos;
- O professor tem toda mediação de solução e utilização pelos alunos;
- Avaliar a familiaridade dos professores com a tecnologia da informação;
- Focar em alunos da 8ª e 9ª do ensino Fundamental e todo ensino médio.

Desenvolvida pela empresa *Joy Street*[1], atualmente a OJE está integrada a dois estados brasileiros, Pernambuco e Acre. Apesar das duas redes apresentarem as mesma funcionalidades, são independentes; ou seja, os usuários de uma rede não se relacionam com os usuários da outra. A OJE pernambuco (OJE-PE)[2] é formada por 49784 usuário, dos quais 5656 são alunos ativos com 20802 relacionamentos. Já a rede do Acre (OJE-AC) [3], possui 5652 usuários, dos quais 5240 são alunos ativos com 9374 relacionamentos (dados coletados em maio de 2013). Na Tabela 4.1, temos os informações descritivas sobre as duas redes.

[1] www.oje.inf.br
[2] www.educacao.pe.gov.br/oje
[3] www.oje.inf.br/acre/

Tabela 4.1: Estatísticas das redes OJE-PE e OJE-AC

	OJE-PE	OJE-AC
Alunos Validados	5656	5240
Links	20802	9374
Densidade da Rede	0,001	0,001
Grau Médio	7,356	3,578
Diâmetro da Rede	11	12

Analisando as estatísticas exibidas na tabela, é possível verificar que embora as duas redes apresentem as mesmas funcionalidades, apresentam diferenças fundamentais: a OJE-AC apresenta praticamente o mesmo número de usuários ativos da OJE-PE, todavia essa apresenta uma maior interação entre usuários, possuindo mais que o dobro de links. Resultando, assim, em um grau médio superior para a OJE-PE.

4.3 Representação dos Usuários

Nesta fase, os atributos relevantes para a caracterização da rede e posteriormente dos grupos, são selecionados a partir do conjunto de dados. Basicamente é realizado um filtro sobre os dados coletados da base, selecionando apenas os atributos mais relevantes para representação do domínio da rede em estudo. Por exemplo, em uma RSE os atributos podem ser: sexo, idade, escola, registros de atividades em disciplinas, áreas de maior interesse, entre outros.

O módulo de representação de usuários é realizado em dois passos. Inicialmente são (a) coletados da base de dados os atributos individuais dos usuários, e posteriormente sobre esses é realizado o (b) processo de filtragem (seleção), resultando no conjunto de atributos "mais representativos" para caracterização dos usuários da rede social em estudo. A filtragem é realizada objetivando a seleção dos atributos mais relevantes, que possam justificar a formação das comunidades.

Salientando que a realização da filtragem dos atributos representativos da rede não é uma tarefa trivial, se faz necessário o acompanhamento e validação por especialistas do domínio da rede em estudo. Do contrário, poderá resultar em resultados incompletos ou até mesmo incorretos. Como por exemplo: a exclusão indevida de um atributo representativo.

Para representação dos usuários da OJE, contamos com a validação de dois pedagogos (funcionários da *Joy Street*) especialistas em educação. Esses acompanharam todo o

processo, modificando e certificando que a representação dos usuários foi realizada corretamente.

Ao final da coleta de dados da base de dados, foi reunido um conjunto de 39 atributos, dentre eles 14 nominais e 25 numéricos, exibidos na Tabela 4.2.

Tabela 4.2: Atributos dos usuários extraídos da base de dados da OJE

Nome	Quantidade de Enigmas Respondidos	Quantidade de Acessos Jogos Linguagem	Quantiadade de Acesso Terceiro Enigma
Idade	Quantidade de Enigmas Corretas	Quantidade de Acessos Jogos Humanas	Quantidade de Acessos Enigma Natureza
Sexo	Quantidade de Enigmas Erradas	Quantidade de Pontos Primeiro Jogo	Quantidade de Acessos Enigma Linguagem
Série	Primeiro Jogo mais Acessado	Quantidade de Pontos Segundo Jogo	Quantidade de Acessos Enigma Humanas
Ensino	Segundo Jogo mais Acessado	Quantidade de Pontos Terceiro Jogo	Quantidade de Medalhas de Bronze
Cidade	Terceiro Jogo mais Acessado	Primeiro Enigma mais Acessado	Quantidade de Medalhas de Prata
Escola	Quantidade de Acesso Primeiro Jogo	Segundo Enigma mais Acessado	Quantidade de Medalhas de Ouro
Quantidade de Acessos Rede	Quantidade Acesso Segundo Jogo	Terceiro Enigma mais Acessado	Classificação Jogos
Quantidade de Acessos Jogos	Quantidade de Acesso Terceiro Jogo	Quantidade de Acesso Primeiro Enigma	Classificação Enigmas
Quantidade de Acessos Enigmas	Quantidade de Acessos Jogos Natureza	Quantidade de Acesso Segundo Enigma	_____

Dos atributos inicialmente coletados da base de dados, foram secionados e validados pelos especialistas do domínio, um conjunto de 16 atributos. Salientando que atributos que não traziam nenhuma informação nova (obvias) em relação a formação dos grupos foram descartados. Como por exemplo: escola, série e cidade. Na Tabela 4.3 são apresentados os atributos selecionados para representarem os usuários da rede OJE (Pernambuco e Acre), a seguir cada um deles são descritos para um melhor entendimento.

- **Idade:** Este atributo foi utilizado para verificar a existência de grupos por faixas etárias.

Tabela 4.3: Atributos selecionados no módulo de representação dos usuários

Idade	Quantidade de Enigmas Respondidos	Quantidade de Acessos Jogos Linguagem	Quantidade de Acessos Enigma Humanas
Quantidade de Acessos Rede	Quantidade de Enigmas Corretas	Quantidade de Acessos Jogos Humanas	Quantidade de Medalhas de Bronze
Quantidade de Acessos Jogos	Quantidade de Enigmas Erradas	Quantidade de Acessos Enigma Natureza	Quantidade de Medalhas de Prata
Quantidade de Acessos Enigmas	Quantidade de Acessos Jogos Natureza	Quantidade de Acessos Enigma Linguagem	Quantidade de Medalhas de Ouro

- **Acessos:** Para verificação do acesso as atividades da OJE por parte dos alunos, se estabeleceu 3 atributos: Quantidade de Acesso ao Site, Jogos e Enigmas.
- **Participação enigmas:** com o objetivo de analisar a participação dos alunos em enigmas, foram definidos 3 atributos: Quantidade de Enigmas Respondidos, Acertados e Errados.
- **Classificação de jogos por área educacional relacionada:** essa classificação foi aplicada baseada no trabalho realizada pela equipe de pedagogos da *JoyStreet*, para identificação da área educacional dos jogos da OJE. Foram criados 3 atributos relacionado as áreas definidas pelos pedagogos: Quantidade de Acesso Jogos: Natureza, Linguagem e Humanas.
- **Classificação de enigmas por área relacionada:** realizada com o mesmo propósito dos jogos. Foram criados 3 atributos: Quantidade de Acessos Enigmas: Natureza, Linguagem e Humanas.
- **Quantidade de Medalhas Obtidas:** para avaliação das medalhas conquistadas (premiações) com a interação entre os alunos, definiu-se 3 atributos Quantidade de Medalhas de: Ouro, Prata e Bronze.

Ao final da fase de representação dos usuários, a API do Gephi[4] é utilizada para geração da rede. Os nós (usuários) possuem todas as informações selecionadas na fase de representação dos usuários, e as arestas representam os relacionamentos entre os nós (por

[4]www.gephi.org/

exemplo, o usuário A e o usuário B são amigos). Após estar com a toda a estrutura das duas redes utilizadas definidas, o método avança a fase de identificação das comunidades, detalhada a seguir.

4.4 Detecção de Comunidades

Foi verificado que a OJE não possui nenhuma aplicação no contexto de grupos, não havendo comunidades explícitas. Dessa forma, aplicou-se o algoritmos de detecção de comunidades MAM para identificação dos grupos. Esse é um método de passos múltiplos baseado na modularidade local otimizada (Girvan and Newman, 2002). Segundo Fortunato and Lancichinetti (2009), o MAM está entre os métodos que apresentaram melhor desempenho sobre redes não direcionadas e não ponderada, tais como a OJE. Além de ser muito rápido, essencialmente em redes de tamanho linear. Maiores detalhes sobre implementação do algoritmo MAM podem ser obtidos na Seção 3.3.1.

Antes da aplicação do algoritmo MAM, para a identificação das comunidades nas redes, realizou-se o pré-processamento sobre os dados coletados. Como o objetivo do estudo é caracterizar grupos sociais a partir dos atributos individuais dos usuários, os nós isolados (*singletons*) foram removidos, uma vez que esses alunos não poderiam estar presentes em nenhuma comunidade. Em seguida, o algoritmo de detecção de comunidades foi executado sobre as duas redes, obtendo 45 grupos na OJE-PE e 32 na OJE-AC.

Com o objetivo de selecionar somente as comunidades mais conectadas para o experimento de caracterização de grupos, realizou-se uma seleção sobre as mesmas. Grupos com menos de 10 usuários foram removidos, pois foram considerados muito pequenos para o estudo. Para os grupos restantes, calculou-se a densidade de cada um, como visto na Seção 3.2, essa é uma medidas comum que verifica quão bem conectada a rede é, ou seja, no nosso caso o quão unido é o grupo (Tang and Liu, 2010a).

Baseado nos valores de densidade calculados, realizou-se um *ranking* dos 10 grupos de maior densidade (TOP-10). As estatísticas dos conjuntos de dados resultante para as duas redes (OJE-PE e OJE-AC) são sumarizadas na Tabela 4.4, onde são apresentados: o número de usuários e *links*, a densidade, grau médio, diâmetro e o número de grupos selecionados.

Tabela 4.4: Estatísticas das redes OJE-PE e OJE-AC pré-processadas para o experimento

	OJE-PE	OJE-AC
Usuários	167	228
Links	894	795
Densidade da Rede	0,064	0,031
Grau Médio da Rede	10,707	6,974
Diâmetro da Rede	5	8
Número de Grupos	10	10

Analisando-se a Tabela 4.4, pode-se observar que a rede resultante (TOP-10) OJE-PE é mais conectada que a OJE-AC, isso é facilmente identificado verificando seus valores de densidade, grau médio e diâmetro. A seguir, são apresentados os detalhes do resultado da detecção de comunidades para as redes OJE-PE e OJE-AC.

4.4.1 Rede OJE-PE

Conforme sumarizado na Tabela 4.4, para realização dos experimentos na rede OJE-PE, foram selecionados um conjunto de 167 usuários (todos alunos ativos e validados) com 894 relacionamentos, distribuídos entre 10 grupos. Na Figura 4.2, a rede OJE-PE resultante de todo o pré-processamento, citado anteriormente, pode ser visualizada. Na legenda ao lado da figura, os rótulos de cada grupo são apresentados, através desses é possível identificar cada grupo individualmente. Na sequência da figura, na Tabela 4.5, as estatísticas de cada grupo são apresentadas; detalhando os seus tamanhos, densidades e graus médios.

Analisando-se a Figura 4.2, inicialmente chama atenção a presença de duas comunidades isoladas, grupo 32 e 34, onde os usuários desses grupos apenas se relacionam com nós do próprio grupo. Ainda é possível verificar que o grupo 10 é o grupo com maior número de usuários, além de apresentar uma boa quantidade de links. Em contra partida, o grupo 13, apesar de ter poucos usuários mostra-se bem conectado.

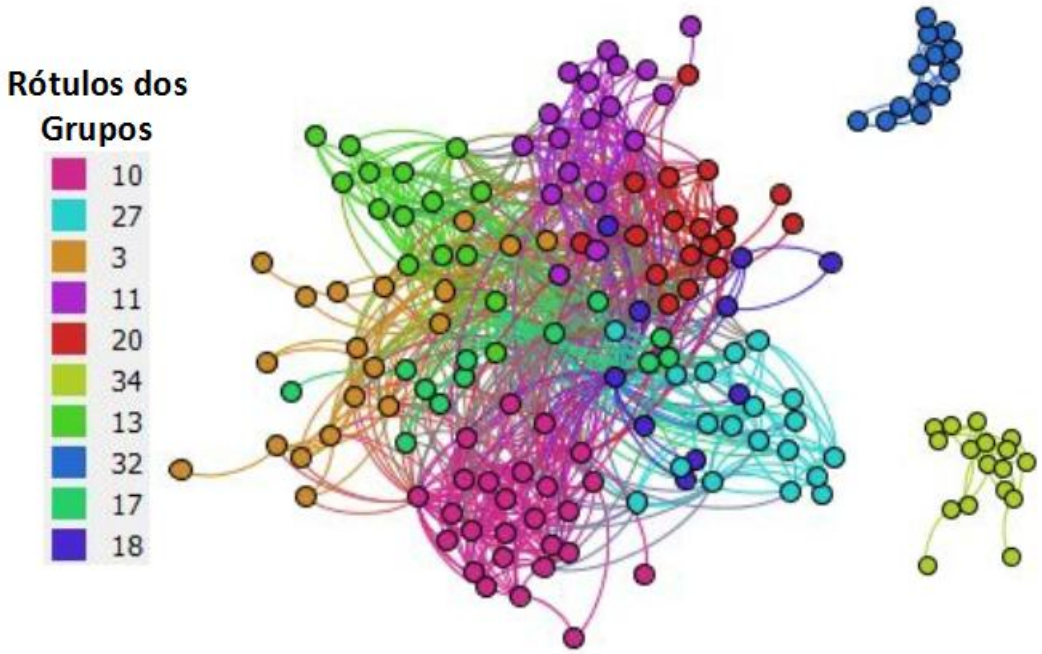

Figura 4.2: Rede OJE-PE do experimento

Tabela 4.5: Estatísticas TOP-10 grupos OJE-PE

Grupo	Tamanho	Grau Médio	Densidade
3	19	5,158	28,7%
10	26	8,846	35,4%
11	18	6	35,3%
13	15	7,867	56,2%
17	12	3	27,3%
18	10	2,600	28,9%
20	18	5,667	33,3%
27	19	5,684	31,6%
32	13	5,231	43,6%
34	17	4,588	28,7%

Baseado nos dados apresentados na Tabela 4.5, podemos verificar as análises iniciais realizadas a partir da Figura 4.2. Pode-se ver que realmente o grupo 10 trata-se do grupo com maior número de membros, e como havia sido apontado apresenta um bom número de links, apresentando também o maior grau médio dentre todos os grupos. Conforme apontado o grupo 13, apesar de pequeno é bastante conectado, apresentado a maior densidade. Com relação as comunidades isoladas, grupos 32 e 34, são comunidade com bons valores de grau médio e densidade comparados aos grupos de valores menores. Todavia, o grupo 32 apresenta maior destaque, possuindo a segunda maior densidade. A seguir na Figura 4.3, são apresentados os quatro grupos isoladamente para facilitar a visualização e entendimento dessas comunidades.

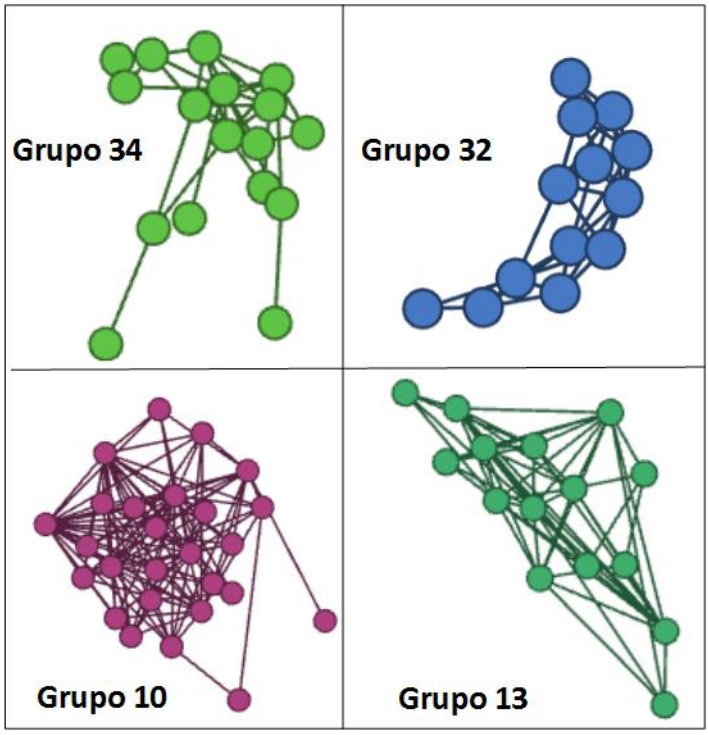

Figura 4.3: Visualização dos grupos 34, 32, 10 e 13 isoladamente.

Para uma visualização geral da distribuição do conjunto de dados da OJE-PE, na Figura 4.4, são apresentados 4 *boxplots*. Os *boxplots* são gerados a partir dos valores dos atributos Idade (esquerda superior), Quantidade de Acessos a Rede (direita superior), Quantidade de acesso a Jogos (esquerda inferior) e Quantidade de Acessos a Enigmas (direita inferior). Esses foram selecionados para dá uma visão geral da distribuição desse conjunto de dados.

Observado-se os *boxplots*, é verificado que nenhum dos atributos selecionados encontra-se em uma distribuição normal. Conforme discutido anteriormente, isso ocorre

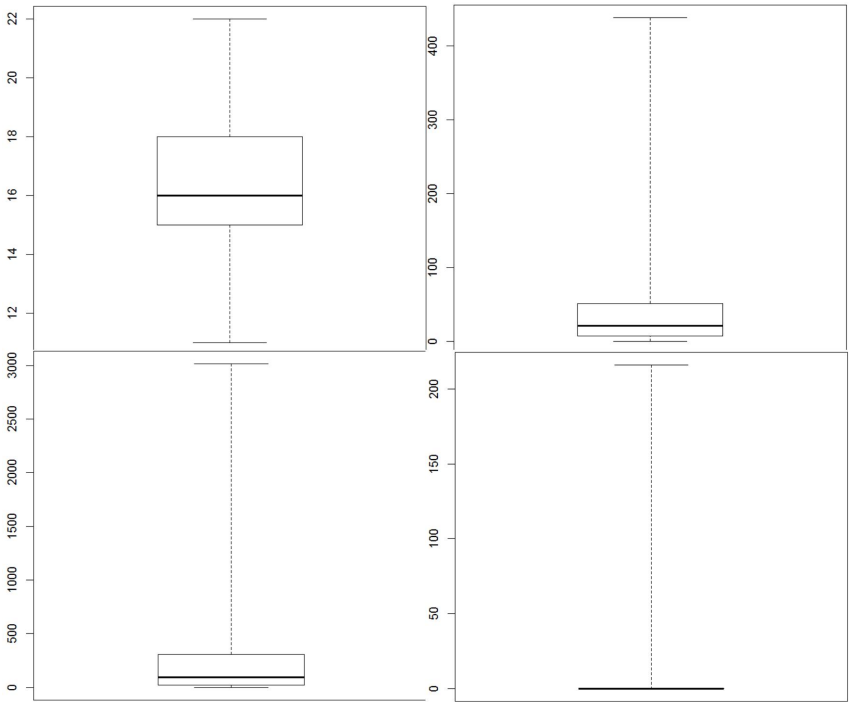

Figura 4.4: *Boxplots* da distribuição dos atributos (OJE-PE): Quantidade de Acesso Rede e Jogos;Quantidade de Enigmas Corretos e Errados

pelo fato dos usuários apresentarem diferentes níveis de acessos às funcionalidades das redes sociais (em sua grande maioria).

Dessa forma, o teste *Shapiro Wilk* (Shapiro and Wilk, 1965) foi executado para verificar a normalidade das distribuições dos atributos do conjunto de dados (maiores detalhes do teste no Anexo A.2). Para todas as distribuições, exceto idade (p valor = 0.0003372), um p valor menor que 2,2e-16 foi obtido. Como o p-valor é menor que o nível de significância (α=5%), a hipótese nula, de que os dados são não normais, é a aceita. Justificando, assim, a aplicação do teste não-paramétrico *Wilcoxon Rank Sum* (WRS). Na Seção 4.5.2, são apresentados e analisados os resultados obtidos com a aplicação do teste WRS para caracterização das comunidades na rede OJE-PE.

4.4.2 Rede OJE-AC

Para realização dos experimentos na rede OJE-AC, selecionou-se um conjunto de 228 usuários (todos alunos ativos e validados) com 795 relacionamentos, distribuídos entre 10 grupos (maiores detalhes na Tabela 4.4). Na Figura 4.5, a rede OJE-AC resultante da seleção dos TOP-10 grupos por densidade, é visualizada. É possível identificar os grupos individualmente na legenda ao lado da figura. Após a figura, na Tabela 4.5, as estatísticas

de cada grupo também são apresentadas.

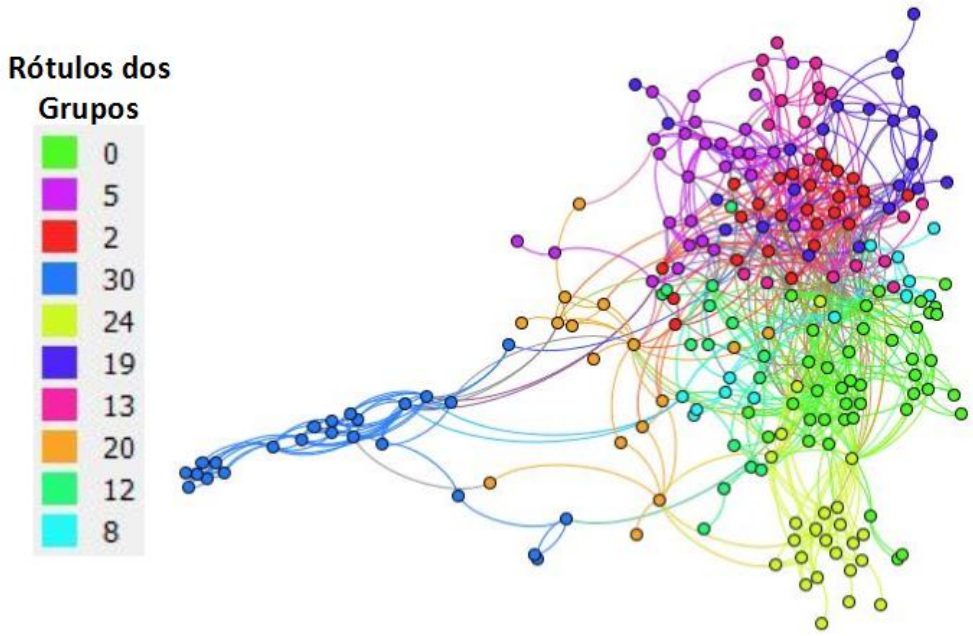

Figura 4.5: Rede OJE-AC do experimento

Analisando-se a Figura 4.5, inicialmente é notado que o grupo 30 demostra ser o mais deslocado das demais comunidades, verificando que a grande maioria dos seus relacionamentos são entre seus próprios membros, além de ser um grupo visivelmente bem conectado. Já o grupo 0, é facilmente visualizado como o grupo com maior número de membros. Outro grupo que chama a atenção é o grupo 2, esse demostra ter bom número de links entre seus usuários, se apresentando como uma comunidade bem conectada.

Tabela 4.6: Estatísticas TOP-10 grupos OJE-AC

Grupo	Tamanho	Grau Médio	Densidade
0	41	6,293	15,7%
2	26	6,385	25,5%
5	27	4,074	15,7%
8	13	3,538	29,5%
12	15	2,533	18,1%
13	19	3,474	19,3%
19	23	3,913	17,8%
20	16	2,625	17,5%
24	24	4,750	20,7%
30	24	6	26,1%

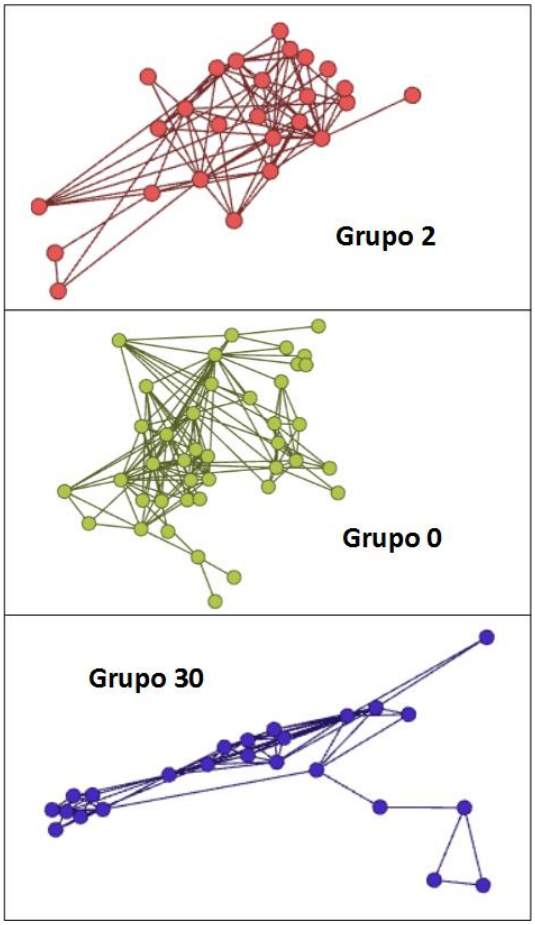

Figura 4.6: Visualização dos grupos 2, 0 e 30 isoladamente.

A partir dos dados apresentados na Tabela 4.6, é possível verificar as análises realizadas inicialmente sobre a Figura 4.5. O grupo 0 de fato é o grupo com maior número de usuários, apresentando uma quantidade considerável de usuários em relação os demais grupos. Como analisado o grupo 30 se diferencia dos demais, verificando seus dados nota-se que se trata de um grupo com uma boa conectividade em relação aos demais, apresentando a segunda maior densidade (superado apenas pelo grupo 8). Indiferente, o grupo 2, que apresenta o maior grau médio. Na Figura 4.6, são apresentados os três grupos isoladamente para facilitar a visualização e entendimento dessas comunidades.

Para ter uma visão geral da distribuição desse conjunto de dados, na Figura 4.7, são apresentados quatro *boxplots*. Os *boxplots* são gerados a partir dos valores dos atributos Idade (esquerda superior), Quantidade de Acessos a Rede (direita superior), Quantidade de acesso a Jogos(esquerda inferior) e Quantidade de Acessos a Enigmas(direita inferior).

Observado-se os *boxplots* dos quatro atributos, é verificado que nenhum desses se encontra em uma distribuição normal. Para o conjunto de dados da OJE-AC, também foi

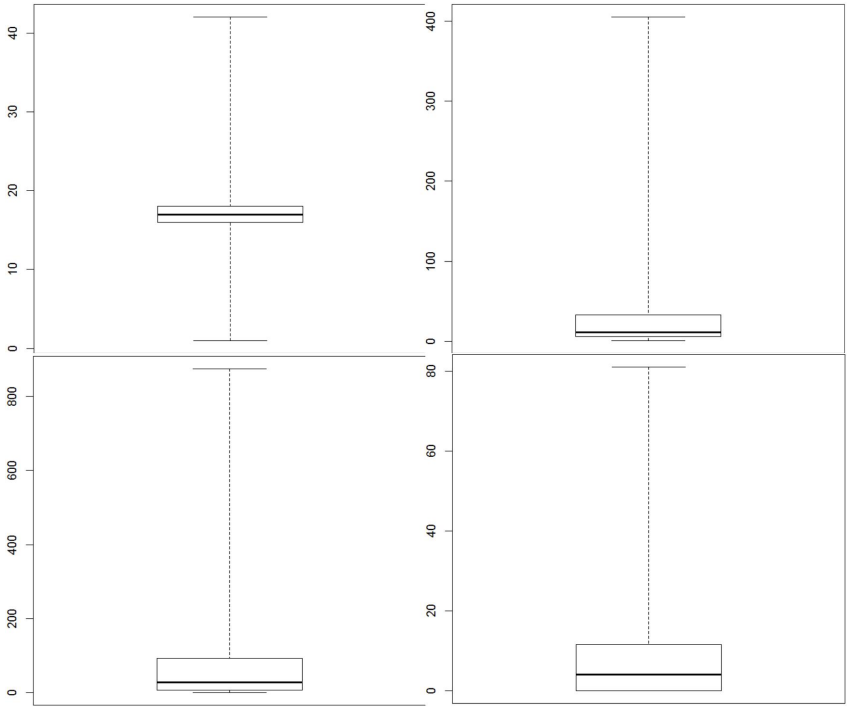

Figura 4.7: *Boxplots* da distribuição dos atributos (OJE-AC): Quantidade de Acesso Rede e Jogos;Quantidade de Enigmas Corretos e Errados

executado o teste *Shapiro Wilk* (Shapiro and Wilk, 1965) para verificar a normalidade das distribuições dos atributos (maiores detalhes do teste no Anexo A.2). Para todas as distribuições obteve-se um p-valor menor que 2,2e-16 foi obtido. Como o p-valor é menor que o nível de significância (α=5%), a hipótese nula, de que os dados são não normais, é a aceita. Dessa forma, justificando também para o conjunto de dados da OJE-AC a aplicação do teste não-paramétrico WRS. Na Seção 4.5.3, os resultados obtidos com aplicação do teste WRS na rede OJE-AC são apresentados e analisados.

4.5 Caracterização de Grupos

Nesta seção, é apresentado o módulo de caracterização de grupos, sintetizado na aplicação do teste WRS. Em seguida, os resultados obtidos com a aplicação do teste WRS nas redes OJE-PE e OJE-AC, são analisados e discutidos.

4.5.1 Aplicação do Teste WRS

Vários métodos que realizam agrupamento têm como objetivo encontrar os recursos que são compartilhados por todo o grupo, uma abordagem natural e simples seria encontrar

atributos que são mais prováveis de ocorrer dentro do grupo. Porém, em vez de agregar, pode-se selecionar características que diferenciam um grupo de outros na rede. O objetivo é descobrir as top-k (principais) características discriminativas que são representativas para um grupo. Ou seja, busca-se selecionar atributos que aparecem em um grupo e, dificilmente, aparecem nos outros.

Conforme visto na Seção 3.3.2, as abordagem baseadas na agregação de atributos individuais só são viáveis em um ambiente relativamente livre de ruído. Se os perfis são construídos a partir de atributos ruidosos, como posts de usuários, registro de atividades de usuários ou relatos de interesses, abordagens baseadas em diferenciação, apresentam melhores perfis para caracterização dos grupos (Tang et al., 2011).

Todavia, muitas informações relevantes podem ser coletadas de dados ruidosos, tais como os supracitados; ou seja, a consideração dessas informações em um estudo de ARS, é fortemente aconselhada. Por exemplo, em um estudo de caracterização de grupos, esses dados podem se tornar excelentes caracterizadores.

Outro ponto importante em dados coletados de redes sociais, é o fato dos atributos dos usuários não apresentarem valores equilibrados (em sua grande maioria), ou seja, alguns usuários apresentarem um maior "destaque" (maiores acessos as funcionalidades) em relação a outros na rede. Consequentemente, isso gera uma distribuição não normalizada entre os atributos dos usuários/grupos da rede.

Em um estudo estatístico, a escolha de qual método utilizar está diretamente relacionada a forma pela qual os dados encontram-se distribuídos. Em caso dos dados não estarem em uma distribuição normal, métodos paramétricos, como o *t-student*, não são viáveis (Goulden, 1956). Logo, um método não-paramétrico é frequentemente uma escolha mais segura, sendo também mais robusto.

Partindo dos problemas apontados na utilização dos dados coletados de um rede social, dados não normalizados e presença de atributos ruidosos, é proposto a adaptação do teste estatístico não-paramétrico WRS a uma abordagem baseada em diferenciação para geração de perfis caracterizadores de grupos (Gomes et al., 2013).

Proposto por Wilcoxon (1945), o teste WRS, em síntese, testa a hipótese nula (H_o) de que, dadas amostras observadas a partir de duas populações A, com n_a observações e B, com n_b observações, tenham a mesma distribuição, isto é, que não há uma diferença estatisticamente relevante nos dados para afirmar que as duas populações são diferentes (Wild, 2000) (maiores detalhes da implementação do teste WRS no Anexo A.1). O Teste é uma alternativa ao teste *t-student*, para comparar as médias de duas amostras independentes quando os dados não encontram-se em uma distribuição normal (Mei et al.,

2008). A seguir, são apresentados os passos realizados na aplicação do teste WRS para caracterização dos grupos:

1. Especifique um nível de significância limite α(por exemplo, 0.05), para indicar os atributos caracterizadores de um grupo;

2. Divide-se os dados coletados da rede social em duas amostras: grupo e restante da rede. Sobre as amostras, calcule o método estatístico WRS para cada atributo representativo dos usuários (todos os atributos selecionados na fase de *representação dos usuário*);

3. Use as estatísticas para calcular o valor correspondente de p (exemplo: função *ranksum* do MATLAB);

4. Selecionar os atributos representativos cujos valores de p são inferiores ao α estabelecido, o que significa que a diferença entre o grupo e o restante da rede para o atributo em estudo é estatisticamente significante, apontando dessa forma, o atributo como um caracterizador do grupo.

As vantagens da aplicação do método estatístico WRS para construção de perfis de grupos sociais, está no fato de conseguir torna a abordagem independente da distribuição dos dados. Ou seja, o método desenvolvido independente da forma como os dados estão distribuídos (em redes sociais os dados tendem a não estarem normalizados). Salientando, ainda, que em casos de distribuições normais, o teste WRS é quase tão eficiente quanto o teste *t-Student* (Mei et al., 2008).

4.5.2 Resultados Rede OJE-PE

Como já mencionado, antes da execução do método estatístico WRS para identificação dos atributos caracterizadores dos grupos, divide-se a rede em duas amostras: grupo e restante da rede. Em seguida, para cada atributo selecionado no módulo de representação dos usuários o teste WRS é executado. Na Tabela 4.7, são apresentados os p valores retornados pelo teste WRS, para cada atributo dos grupos.

Os atributos com p valores marcados de verde ou vermelho indicam que foram selecionados como caracterizador para o grupo. Onde os verdes, são os que apresentaram uma diferença estatisticamente relevantes para uma medida de significância $\alpha = 0.05$. E os vermelhos os que apresentaram as maiores diferenças estatísticas, possuindo um p valor inferior ao α_{max} (0.01). Os demais são os atributos que não apresentaram diferença

Tabela 4.7: *P* Valores retornados pelo teste WRS para a rede OJE-PE

Atributos/Grupos	3	10	11	13	17	18	20	27	32	34
Idade	0.04e-5	0.048	0.4	0.6	0.8	0.4	0.6	0.1	0.06e-2	0.03e-6
Acessos Rede	0.60	0.73	0.048	0.004	0.9	0.4	0.8	0.4	0.04e-2	0.061
Acessos Jogos	0.63	0.9	0.10	0.1	0.57	0.8	0.9	0.5	0.007	0.007
Acessos Enigmas	0.65	0.5	0.42	0.4	0.041	0.6	0.1	0.1	0.03	0.063e-14
Enigmas Respondidos	0.66	0.6	0.5	0.3	0.3	0.6	0.2	0.6	0.015	0.04e-8
Enigmas Corretas	0.56	0.8	0.6	0.1	0.4	0.3	0.2	0.8	0.023	0.063e-9
Enigmas Erradas	0.69	0.6	0.5	0.3	0.38	0.6	0.3	0.5	0.015	0.074e-8
Jogos Natureza	0.90	0.9	0.057	0.1	0.7	0.5	0.7	0.5	0.009	0.016
Jogos Linguagem	0.31	0.6	0.7	0.4	0.73	0.6	0.8	0.9	0.057	0.1
Jogos Humanas	0.14	0.56	0.9	0.08	0.3	0.4	0.6	0.3	0.07	0.002
Enigma Natureza	0.56	0.2	0.8	0.3	0.07	0.9	0.11	0.2	0.065	0.044e-17
Enigma Linguagem	0.87	0.9	0.8	0.1	0.07	0.3	0.2	0.07	0.062	0.012e-15
Enigma Humanas	0.93	0.5	0.1	0.1	0.07	0.3	0.027	0.08	0.065	0.065e-16
Medalhas de Bronze	0.97	0.6	0.1	0.1	0.59	0.8	0.84	0.2	0.08e-2	0.4
Medalhas de Prata	0.71	0.5	0.1	0.058	0.81	0.82	0.8	0.3	0.001	0.3
Medalhas de Ouro	0.52	0.3	0.1	0.07	0.81	0.81	0.4	0.5	0.07e-2	0.3

estatisticamente relevante (p valor superior a 0.05), ou seja, não foram selecionados como caracterizadores para os grupos.

Analisando-se as médias dos valores dos atributos selecionados como caracterizadores para o estudo, em grupo individual pelo restante da rede, é verificado se o grupo se diferencia do restante da rede por ter maiores valores ou menores. Na Figura 4.8, os atributos caracterizadores selecionados para cada grupo são exibidos, de acordo com os resultados obtidos pelo teste WRS. Atributos marcados de azul, indicam que o valor médio no grupo é maior que o valor médio para o atributo no restante da rede. Os de vermelho, todavia, indicam que os atributos dentro do grupo tem uma média inferior em relação ao restante da rede.

Analisando-se os dados, observa-se que as combinações dos caracterizadores identificados para cada grupo é bastante distinta, o que demostra o potencial do método proposto. Para obter-se uma melhor compreensão dos resultados, a seguir serão discutidos três exemplos concretos: os grupos 27, 32 e 34.

Conforme apontado na Figura 4.8, descobriu-se que o grupo 34 é o mais proeminente da rede. A Figura 4.9 apresenta um gráfico de barras contendo o valor médio de cada atributo selecionado para caracterizar o grupo 34 em comparação com o valor médio para o restante da rede. Ao final da análise, conclui-se que o grupo é formado por alunos mais

Grupos	Atributos Caracterizadores
18 e 27	Nenhum Atributo Selecionado
3 e 10	Idade
11 e 13	Acessos a Rede
17	Acessos a Enigmas
20	Enigmas Humanas
32	Idade, Acessos a Rede, Acessos a Jogos, Acessos a Enigmas, Enigmas Respondidas, Enigmas Corretos, Enigmas Errados, Jogos Natureza, Medalhas Bronze, Medalhas Prata, Medalhas Ouro
34	Idade, Acessos a Jogos, Acessos a Enigmas, Jogos Natureza, Jogos Humanas, Enigmas Natureza, Enigmas Linguagem, Enigmas Humanas, Enigmas Respondidas, Enigmas Corretas, Enigmas Erradas

Figura 4.8: Atributos caracterizadores de cada grupo

jovens (média de idade 14 anos), diferenciando-se dos demais por um elevado acesso aos enigmas e por não apresentar uma boa participação em jogos.

Esse destaque em enigmas é bastante relevante para análises de influências no aprendizado proporcionadas pela OJE. Verifica-se que alunos estão interagindo e formando comunidades para uma atividade a fim, no caso, para resolução de enigmas. E como apresentado, na Figura 4.9, os alunos do grupo 34 se destacam nas 3 áreas de enigmas: Natureza, Humanas e Linguagem. Salientando, ainda, que esse grupo trata-se de um comunidade isolada (verificar comunidade na Figura 4.2), o que também justifica o comportamento diferenciado em relação ao demais grupos.

Na Figura 4.10 é apresentado o gráfico de barras com os valores médios dos atributos caracterizadores do grupo 32, em relação ao restante da rede. Nesse verifica-se um comportamento oposto ao grupo 34, todos os atributos caraterizadores do grupo 32 apresentam um valor médio que é inferior a média observada para o restante da rede. Este grupo de alunos estão unidos em comunidades, não só pela estrutura identificada pelo método de detecção de comunidades, mas também por um comportamento comum. O

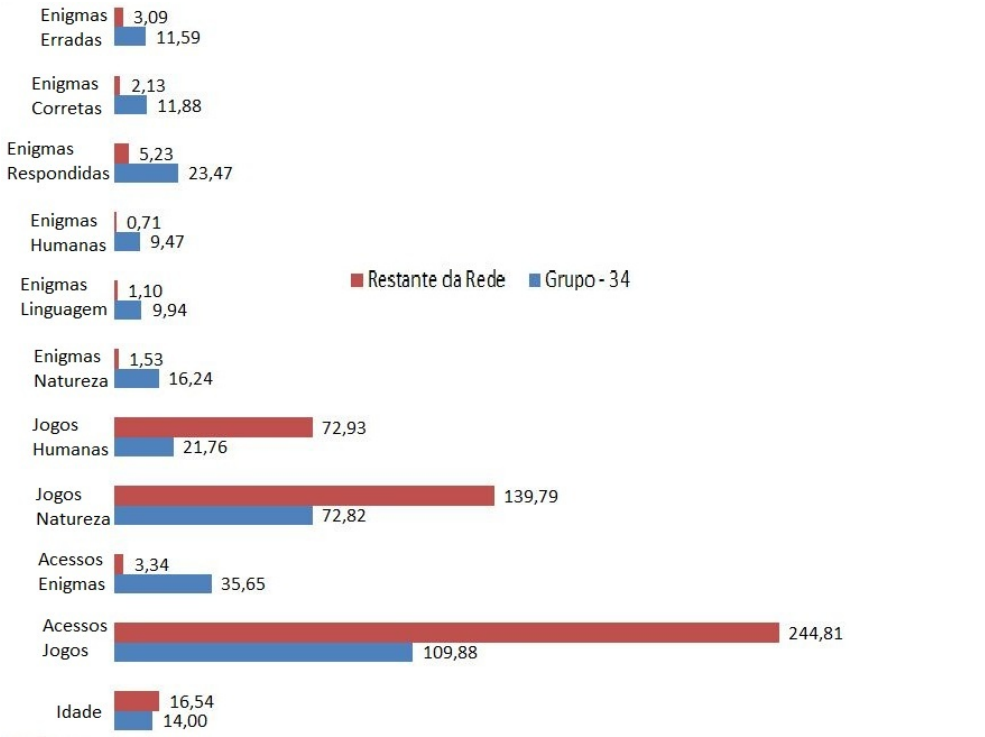

Figura 4.9: Gráfico de barras com as médias de cada atributo caracterizador selecionado para o grupo 34

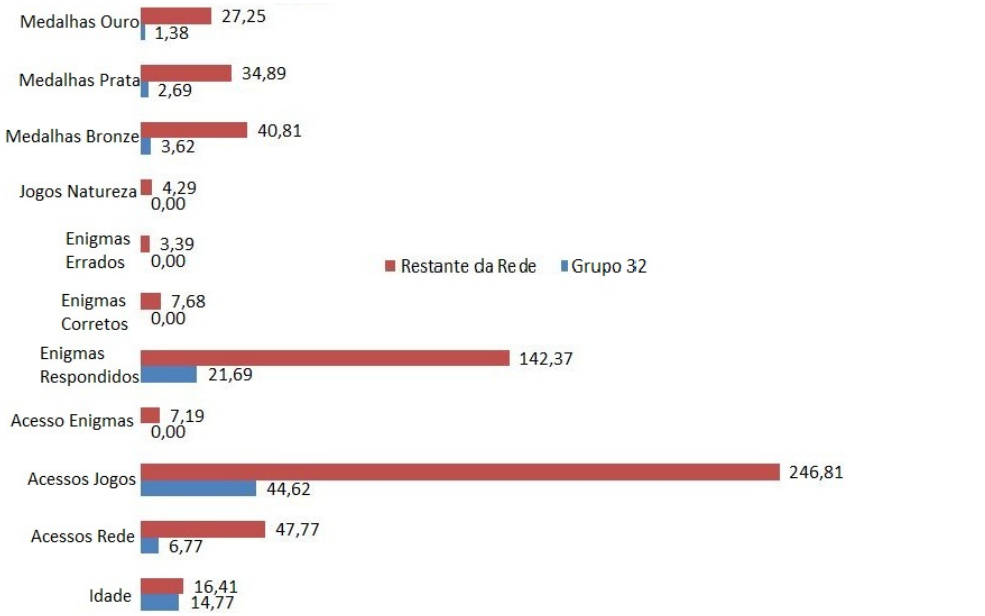

Figura 4.10: Gráfico de barras com as médias de cada atributo caracterizador selecionado para o grupo 32

grupo apresenta um baixo acesso a rede social e aos jogos, e o mais agravante, nenhum acesso aos enigmas, o que não é bom para os objetivos educacionais da OJE. Refletindo, ainda, em um número baixíssimo de premiações (medalhas) obtidas pelos alunos desse grupo. Outras investigações poderiam ser analisadas sobre esse conjunto de usuários, para explicar tais comportamentos. Salientando, ainda, que esse grupo assim como o anterior, trata-se de um comunidade isolada (verificar comunidade na Figura 4.2).

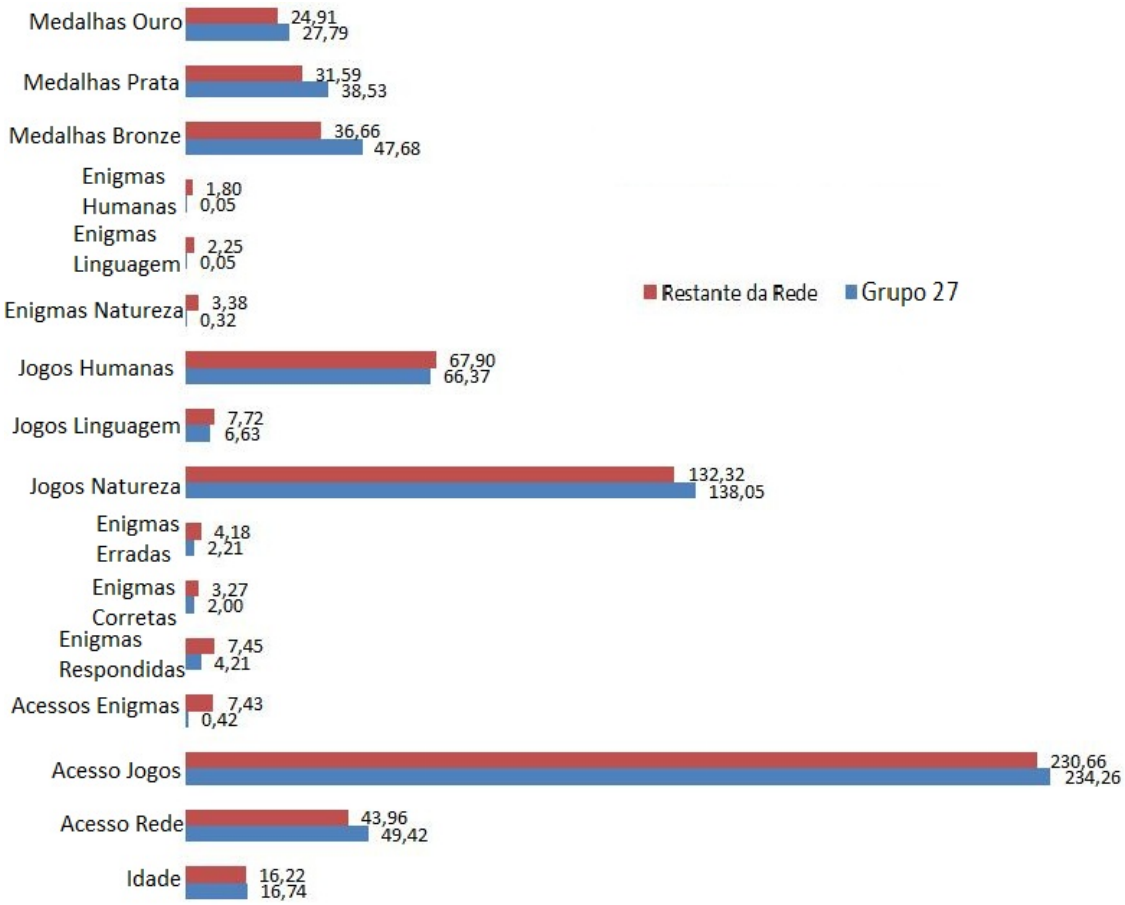

Figura 4.11: Gráfico de barras com as médias de cada atributo do grupo 27 em comparação ao restante da rede.

Uma limitação identificada no método de caracterização de grupos proposto, está no caso onde as diferenças entre todos os atributos de um grupo em relação ao restante da rede são mínimas (não significante). Em tal situação, o teste de Wilcoxon não revelou nenhum atributo caracterizador para o grupo, não o caracterizando. Todavia, essas comunidades podem ser utilizadas como amostras da rede OJE-PE, uma vez que os usuários apresentam comportamentos muito semelhantes (indiferentes) ao restante da rede. Na Figura 4.11, pode-se ver como exemplo o grupo 27. A seguir, na Seção 4.5.3,

Tabela 4.8: *P* Valores retornados pelo teste WRS para a rede OJE-AC

Atributos/Grupos	0	2	5	8	12	13	19	20	24	30
Idade	0.551	0.29	0.76	0.08	0.4	0.012	0.23	0.48	0.74	0.047
Acessos Rede	0.06e-5	0.92	0.17	0.30	0.7	0.009	0.046	0.77	0.40	0.025
Acessos Jogos	0.001	0.88	0.69	0.057	0.7	0.041	0.154	0.31	0.97	0.28
Acessos Enigmas	0.036	0.15	0.41	0.12	0.28	0.010	0.18	0.005	0.19	0.79
Enigmas Respondidos	0.054	0.14	0.78	0.12	0.3	0.032	0.26	0.009	0.27	0.94
Enigmas Corretas	0.018	0.8	0.60	0.15	0.94	0.090	0.82	0.055	0.57	0.38
Enigmas Erradas	0.074	0.045	0.93	0.32	0.16	0.024	0.23	0.016	0.22	0.8
Acessos Jogos Natureza	0.015e-2	0.74	0.88	0.12	0.40	0.014	0.14	0.285	0.96	0.133
Acessos Jogos Linguagem	0.016	0.67	0.97	0.4	0.23	0.058	0.61	0.044	0.48	0.68
Acessos Jogos Humanas	0.028	0.92	0.87	0.033	0.78	0.23	0.15	0.28	0.78	0.3
Acessos Enigma Natureza	0.025	0.12	0.51	0.09	0.39	0.024	0.21	0.005	0.10	0.84
Acessos Enigma Linguagem	0.093	0.50	0.86	0.058	0.28	0.048	0.36	0.029	0.58	0.83
Acessos Enigma Humanas	0.006	0.18	0.29	0.24	0.38	0.004	0.26	0.020	0.056	0.8
Medalhas de Bronze	0.013e-2	0.63	0.90	0.14	0.48	0.022	0.048	0.10	0.52	0.1
Medalhas de Prata	0.07e-3	0.7	0.92	0.14	0.40	0.003	0.087	0.23	0.71	0.19
Medalhas de Ouro	0.082e-3	0.93	0.94	0.26	0.34	0.016	0.0316	0.059	0.54	0.2

são apresentados os resultados obtidos nos experimentos realizados na rede OJE-AC.

4.5.3 Resultados Rede OJE-AC

Igualmente aos experimento realizados na base da OJE-PE, dividiu-se a rede em duas amostras, grupo e restante da rede, e para cada atributo selecionado no módulo de representação dos usuários o teste WRS foi executado. Na Tabela 4.8, são apresentados os p valores retornados pelo teste WRS, para cada atributo dos grupos.

Indiferente do que foi realizado para OJE-PE, os atributos com p valores marcados de verde ou vermelho indicam que foram selecionados como caracterizador para o grupo. Onde os verdes, apresentarem uma diferença estatisticamente relevantes para a medida de significância considerada ($\alpha = 0.05$), já os vermelhos são os que apresentaram maiores diferenças estatísticas, possuindo p valor inferior ao α_{max} (igual 0.01). Os demais são os atributos que não apresentaram diferença estatisticamente relevante, ou seja, não foram selecionados como caracterizadores para os grupos.

Para a rede OJE-AC também foram analisadas as médias dos valores dos atributos em um grupo comparados ao restante da rede. Na Figura 4.12, são apresentados os atributos caracterizadores selecionados para cada grupo, de acordo com os resultados obtidos pelo

teste de WRS. Com a mesma caracterização realizada na OJE-PE, atributos marcados de azul indicam que seu valor médio no grupo é superior ao seu valor no restante da rede, já os de vermelho indicam o contrário.

Grupos	Atributos Caracterizadores
5, 12 e 24	Nenhum Atributo Selecionado
2	Enigmas Erradas
8	Jogos Humanas
19	Acessos a Rede, Medalhas de Bronze
20	Acessos a Enigmas, Enigmas Respondidos, Enigmas Erradas, Jogo Linguagem, Enigmas Humanas, Enigmas Linguagem, Enigmas Natureza
30	Acessos a Rede, Idade
0	Acessos a Rede, Acessos a Jogos, Acessos a Enigmas, Enigmas Corretos, Jogos Natureza, Jogo Linguagem, Jogo Humanas, Enigmas Natureza, Enigmas Humanas, Medalhas Ouro, Medalhas Prata, Medalhas Bronze
13	Idade, Acessos a Jogos, Acessos a Enigmas, Acessos a Rede, Acessos a Jogos, Acessos a Enigmas, Enigmas Respondidos, Enigmas Erradas, Jogo Natureza, Enigmas Natureza, Medalhas Ouro, Medalhas Prata, Medalhas Bronze

Figura 4.12: Atributos caracterizadores de cada grupo - rede OJE-AC

Analisando-se os dados, observa-se que as combinações dos caracterizadores identificados para cada grupo na rede OJE-AC é totalmente distinta, ou seja, nenhum grupo apresentou os mesmo atributos de um outro. Isso valida, mais uma vez, o potencial do método proposto. Para obter-se uma melhor compreensão dos resultados na OJE-AC, a seguir serão discutidos quatro exemplos concretos: os grupos 0, 12, 13 e 20.

Conforme apontado na Figura 4.12, descobriu-se que o grupo 0 é o mais proeminentes da rede. A Figura 4.13 apresenta um gráfico de barras contendo o valor médio de cada atributo selecionado para caracterizar o grupo 0 em comparação com o valor médio para o restante da rede. Analisando-se o grupo 0, concluiu-se que esse é o grupo mais

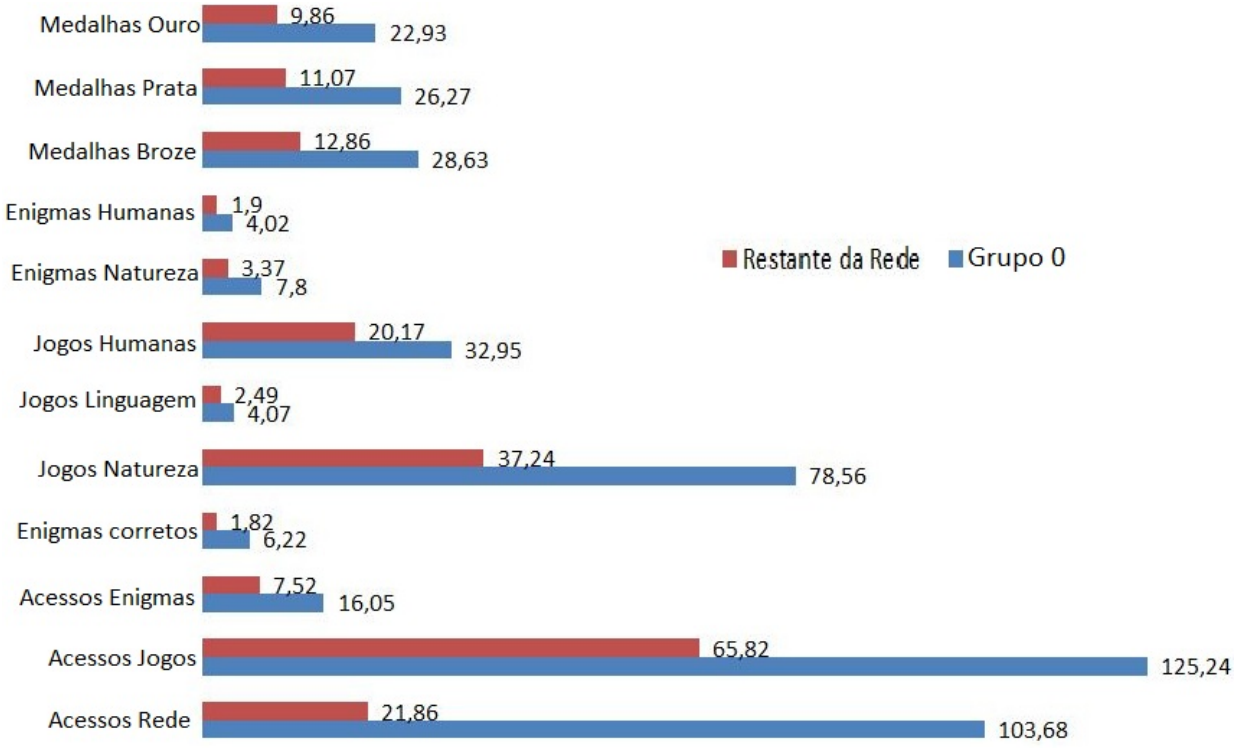

Figura 4.13: Gráfico de barras com as médias de cada atributo caracterizador selecionado para o grupo 0

proeminente da rede; constituído por pessoas que são muito envolvidas na OJE, desde do uso da rede social, até um elevado nível de participação em jogos e enigmas. Refletindo, ainda, em um número bastante elevado de premiações (medalhas) obtidas pelos alunos desse grupo.

Em relação aos grupos 13 e 20, pode-se ver o comportamento oposto, como pode ser visto nas Figura 4.14 e 4.15. Exceto para os atributos idade e medalhas de ouro no grupo 13, todos os demais atributos selecionados como caracterizadores para descreverem os dois grupos apresentam um valor médio que é inferior a média observada na amostra do restante da rede. Esses grupos de alunos estão unidos em comunidade, não só pela estrutura identificada pelo método de detecção da comunidade, mas também por um comportamento comum.

O grupo 13 apresenta um idade mais elevada considerada ao restante da rede(média de 17,47 anos). Entretanto apesar dos atributos selecionados como caracterizadores para esse grupo, possuírem médias inferiores ao restante da rede, o grupo apresenta uma média de premiações em medalhas de ouro superior aos demais. Já para o grupos 20, todos os atributos apresentam médias inferiores comparadas ao restante da rede.

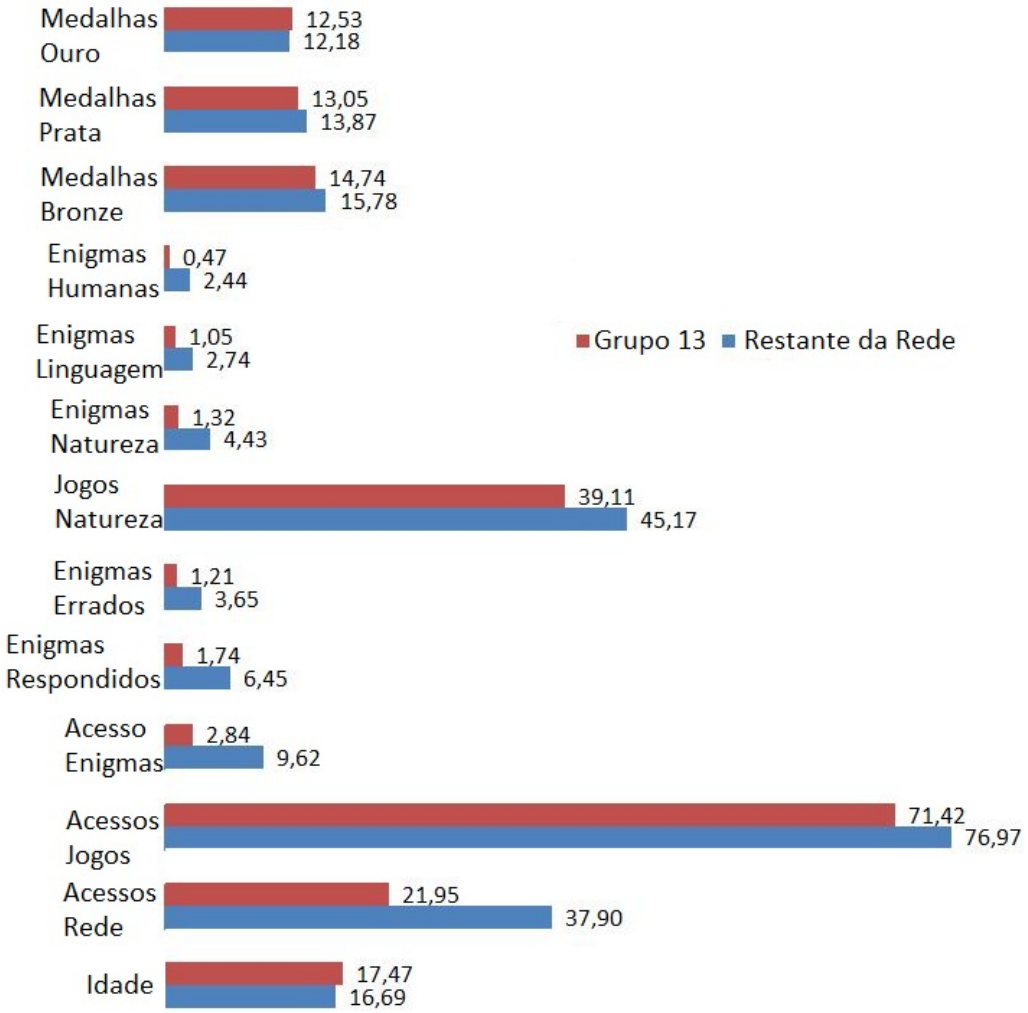

Figura 4.14: Gráfico de barras com as médias de cada atributo caracterizador selecionado para o grupo 13

Outras investigações poderiam ser aplicadas sobre este conjunto de usuários para explicar esses comportamentos. Com exemplo, justificar o porquê de tais estruturas terem se formados, ou mais além, como o grupo 13 que apresenta baixíssimos acessos a rede, jogos e enigmas apresenta um número considerável em premiação de medalhas de ouro.

Igualmente como ocorreu para rede OJE-PE, na rede OJE-AC o método proposto apresentou a mesma limitação, não caracterizando grupos onde as diferenças entre seus atributos e o restante da rede são mínimas (não significantes estatisticamente). Como já mencionado em tais casos, o teste de Wilcoxon não revela nenhum atributo caracterizador para o grupo. Contudo, esses grupos podem ser utilizados como amostras da rede OJE-AC, uma vez que os usuários apresentam comportamentos muito semelhantes (indiferentes) ao restante da rede. Na Figura 4.16, pode-se ver como exemplo o grupo 12.

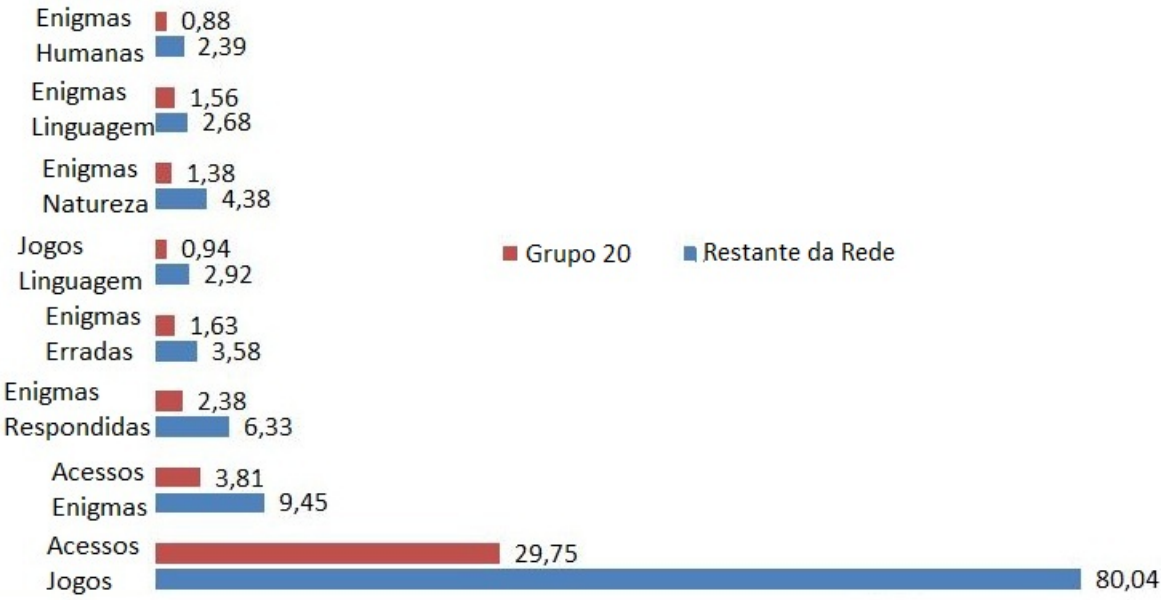

Figura 4.15: Gráfico de barras com as médias de cada atributo caracterizador selecionado para o grupo 20

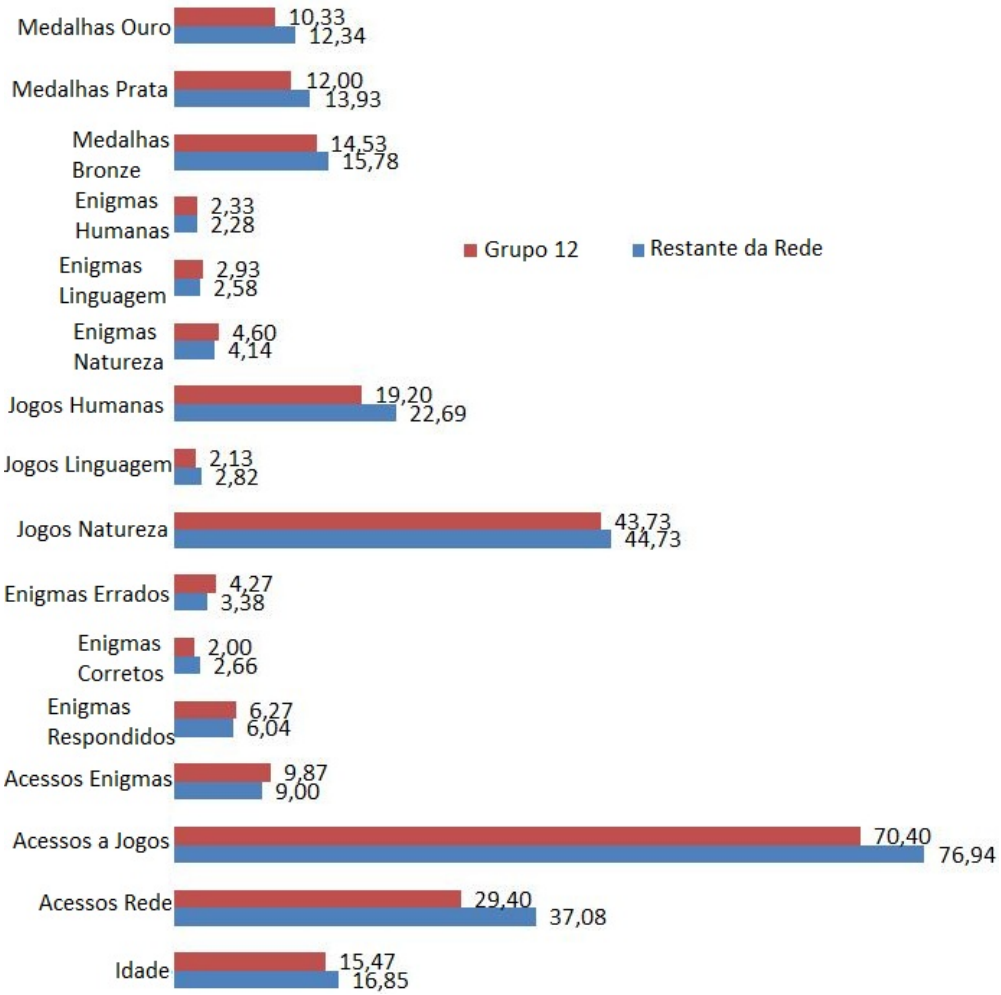

Figura 4.16: Gráfico de barras com as médias de cada atributo caracterizador selecionado para o grupo 12

4.6 Considerações Finais

Este capítulo procurou descrever o método proposto, fruto da pesquisa realizada neste trabalho, assim como, apresentar a viabilidade do método através da realização de experimentos em duas redes sociais educacionais. Os resultados levantados demonstraram a eficácia da abordagem proposta para geração de caracterizadores de grupos, conforme visto nos experimentos.

Analisando-se as redes OJE-PE e OJE-AC, verificou-se comportamentos distintos, apesar das redes apresentarem as mesmas funcionalidades. Por exemplo, na rede OJE-PE, formou-se uma comunidade (grupo 34) de alunos com grande destaque na utilização dos enigmas e baixo acesso aos jogos. Verifica-se que alunos estão interagindo e formando comunidades para uma atividade a fim, no caso, para resolução de enigmas. Enquanto que na OJE-AC, uma comunidade (grupo 32) de alunos com elevados acessos a todas as funcionalidades da OJE foi formada. Tais dados demostram a influência da OJE na interação dos alunos e formação de comunidades, possibilitando, ainda, uma maior comunicação e continuação do aprendizado fora das salas de aula.

Também foi verificado comportamentos comuns as duas redes. Por exemplo, em ambas as redes, formou-se comunidades de alunos com baixos acessos as funcionalidades da OJE. Isso destaca a necessidade da realização de estudos mais aprofundados sobre essas comunidades, na busca por um entendimento real para formação desses grupos.

Dessa forma, concluí-se que apesar do método proposto não conseguir caracterizar as comunidades, quando as diferenças entre o grupo e o restante da rede não são relevantes estatisticamente, o método foi efetivo na identificação de atributos para caracterização dos grupos. De fato, através dos atributos individuais de cada aluno, características descritivas foram identificadas para 80% dos grupos da rede OJE-PE e 70% para OJE-AC (TOP-10). Esses melhores resultados para OJE-PE eram de certa forma esperados, uma vez que essa rede é mais conectada que a OJE-AC (verificar Tabela 4.4). Como visto na análise dos resultados, os atributos identificados pelo método proposto tornaram-se bons caracterizadores para grupos, apontando os prováveis motivos para a formação das comunidades. Outro ponto positivo do método, deve-se ao fato da possibilidade de visualização dos grupos, facilitando o entendimento e reflexão dos resultados. No Capítulo 5, a seguir, são apresentadas as considerações finais do trabalho.

5
Considerações Finais

Perder a paciência
é perder a batalha.
—MAHATMA GANDHI

Este capítulo apresenta as considerações finais deste trabalho. Inicialmente são expostas as conclusões e contribuições obtidas pela pesquisa, traçando um paralelo entre os objetivos do trabalho e os resultados alcançados por ele. Em seguida, são discutidas algumas limitações do estudo e, por fim, direcionamentos para trabalhos futuros.

5.1 Conclusões e Contribuições

A revisão bibliográfica permitiu o entendimento de várias técnicas relacionadas a MDE, mais precisamente as voltadas para análise de grupos e redes sociais. A partir desse embasamento, realizou-se um levantamento bibliográfico reunindo as principais características das redes complexas, voltadas para a detecção de comunidades. Com posse dessa informações, iniciou-se um estudo sobre uma das subáreas mais inovadores da detecção de comunidade: a caracterização de grupos. Não só os aspectos gerais, como também as dificuldades e deficiências atuais do problema foram primordiais para a definição dos objetivos almejados e contribuições deste trabalho para a comunidade acadêmica. Em suma, os fundamentos aprendidos ajudaram na definição do escopo de um trabalho prático, sob o qual foi possível a aplicação de conceitos, técnicas e metodologias da área de pesquisa.

A partir do levantamento feito, verificou-se que diferentes abordagens de caracterização de grupos podem ser adotadas. Uma abordagem natural seria agregar os atributos individuais compartilhados com uma maior frequência dentro de um grupo. Observando-

se os atuais sistemas de etiquetagem, é verificado que essa abordagem é amplamente utilizada, na forma de nuvens de *tags*. No entanto, conforme o estudo comparativo apontado por Tang et al. (2011), essa abordagem só é viável em um ambiente relativamente livre de ruído, em tais situações melhores perfis caracterizadores de grupos são obtidos com métodos baseados na diferenciação. Outro ponto importante, deve-se ao fato dos dados coletados das redes sociais dificilmente apresentarem uma distribuição normal, ou seja, os valores dos atributos dos usuários dificilmente são equilibrados (diferentes níveis de acesso entre os usuários). Neste contexto, através da realização de um estudo estatístico diferenças estatísticas entre amostras podem ser identificas. Ressaltando que para dados não normalizados um teste não paramétrico é frequentemente a escolha mais segura. Dessa forma, todos esses fatores estimularam para que o presente estudo seguisse por esta vertente.

A primeira fase do trabalho consistiu na modelagem da rede, mais precisamente na representação dos atributos individuais dos usuários e identificação das comunidades da rede. A partir dos atributos individuais dos usuários e comunidades identificadas, é proposto um método baseado em diferenciação para caracterização de grupos em redes sociais, aplicando-se o teste estatístico não-paramétrico *Wilcoxon Rank Sum* (Gomes et al., 2013). O método proposto tem como objetivo caracterizar os grupos verificando interesses/características comuns a um grupo (primeira amostra), que diferenciam esse grupo do restante da rede (segunda amostra).

A partir dos experimentos realizados, expostos no Capítulo 4, foi possível concluir que apesar do método proposto não conseguir caracterizar as comunidades, quando as diferenças entre o grupo e o restante da rede não são relevantes estatisticamente; o método foi efetivo na caracterização dos grupos. De fato, através dos atributos individuais de cada aluno, características descritivas foram identificadas para 80% dos grupos da rede OJE-PE e 70% para OJE-AC. Como visto na análise dos resultados, os atributos identificados pelo método proposto tornaram-se bons caracterizadores para grupos, retratando os possíveis motivos que levaram a formação das comunidades. Outro ponto positivo do método, deve-se ao fato da possibilidade de visualização dos grupos, facilitando o entendimento e reflexão dos resultados.

Baseado nos comentários acima, as principais contribuições deste trabalho foram:

- Proposição da aplicação do teste não paramétrico *Wilcoxon Rank Sum* como método baseado em diferenciação para caracterização de grupos em redes sociais. Através desse, outros pesquisadores podem replicar o método proposto, bem como realizar estudos comparativos com outros métodos;

- Estudo de caracterização de grupos em um rede social educacional. Não foi identificado nenhum trabalho voltado para a caracterização de grupos em um ambiente de RSE. Ou seja, o crescimento dessas plataformas educacionais gera a necessidade de estudos mais aprofundados sobre suas estruturas sociais, não se limitando a analise de dados (MDE);

- Os resultados obtidos podem ser de grande importância para modelagem e avaliação do processo de aprendizado, auxiliando, a tomada de decisão por parte de gestores e professores das RSEs (OJE-PE e OJE-AC). Possibilitando, ainda, o fornecimento de conteúdos adaptáveis e um aprendizado direcionado aos grupos;

- O método conseguiu caracterizar uma boa porcentagem dos grupos nos experimentos realizados. Na rede OJE-PE 80% dos grupos foram caracterizados e na rede OJE-AC, 70%. Conforme foi visto nos estudo os perfis gerados apontaram bons caracterizadores para os grupos, tornando possível um entendimento e facilitanto bastante a análise visual e comprensão das comunidades caracterizadas;

- O método torna possível a visualização e caracterização das comunidades. Dessa forma, um usuário que não se relaciona com ninguém (*singletons*), pode se identificar com um grupo de pessoas similares a ele, iniciando novos relacionamentos. Assim como, comunidades isoladas com interesses semelhantes podem passar a interagir, possibilitando o desenvolvimento das comunidades e formação de uma única comunidade maior;

- O método proposto pode ser utilizado no acompanhamento das mudanças de características dos grupos. Analisando-se os grupos em momentos diferentes, pode-se verificar se as comunidades ainda apresentam as mesmas características ou não;

- Produção de um artigo aceito e publicado em uma conferência (Gomes et al., 2013).

5.2 Limitações do Estudo

Não foi objetivo deste trabalho abordar todos os aspectos da caracterização de grupos, focando-se apenas na análise da viabilidade da aplicação do teste WRS como método baseado em diferenciação para caracterização de grupos em redes sociais. Dessa forma o estudo apresentou algumas limitações, dentre as quais:

- O método proposto não conseguiu caracterizar grupos com comportamento poucos diferentes do restante da rede. Como o método proposto se baseia na diferença estatística entre os atributos de um dado grupo comparados ao restante da rede, grupos que não revelaram essa diferença não recebem nenhum atributo para caracterização de seus perfis;
- O método proposto não foi comparado a outros métodos de caracterização de grupos já aplicados anteriormente. Os principais métodos de caracterização de grupos, se baseiam em textos coletados da rede (postagens dos usuários), e através desses tentam caracterizar os grupos. No presente trabalho, considerou-se um conjunto de atributos representativos para os usuários das redes, ou seja, a caracterização dos grupos não foi realizada baseando-se nos conteúdos coletados das postagens, dificultando assim a comparação dos métodos. A aplicação de técnicas de aprendizagem de máquina para classificação dos grupos, seria uma possível forma de comparação com o método proposto.

5.3 Trabalhos Futuros

As limitações deste trabalho dão uma visão do que ainda precisa ser feito para ampliar o campo de análise. Além disso, a estratégia aqui adotada pode ser empregada para outros propósitos dentro da ARS. Nosso objetivo é seguir com o trabalho realizado empregando as seguintes melhorias:

- Espera-se pesquisar e aplicar soluções ao problema relacionado a não geração de perfis caracterizadores para alguns grupos;
- Realizar estudos mais aprofundados para justificar a formação de comunidades entre usuários que praticamente não interagem com a rede (usuários que apresentam baixos acessos as funcionalidades da rede, por exemplo Figura 4.14);
- Pretende-se estudar a aplicação de técnicas de aprendizado de máquina para a geração de perfis caracterizadores de grupo em redes sociais. Realizando posteriormente um estudo comparativo como o método proposto;
- Evolução do método de caracterização de grupos para investigação da evolução das comunidades. Em um ambiente dinâmico com as redes sociais, as comunidades evoluem: podendo crescer, fundir-se, dividir-se, ou até mesmo dissolver-se; Estudo sobre a evolução das comunidades incentivam bastante a continuação da pesquisa.

Referências Bibliográficas

Almack, J. (1922). The school childs choice of companions. *School and Society 16*, 529–530.

Almeida, L. J. (2009). *Detecção de comunidades em rede complexas utilizando estratégia multinível*. MSc em Ciência da Computação e Matemática Computacional, Universidade de São Paulo.

Ayers, E., R. Nugent, and N. Dean (2009). A comparison of student skill knowledge estimates. In *EDM2009: 2nd International Conference on Educational Data Mining*.

Backstrom, L., D. Huttenlocher, J. Kleinberg, and X. Lan (2006). Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 44–54.

Baker, R., S. Isotani, and A. Carvalho (2011). Mineração de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação 19*(02).

Baker, R., A. Merceron, and P. Pavlik (2010). Educational data mining 2010, the 3rd international conference on educational data mining, pittsburgh, pa, usa, june 11-13, 2010. proceedings. In *EDM*.

Barabási, A. L. (2009). *Linked: a nova ciência dos networks*. São Paulo, Brasil: Leopardo.

Baradwaj, B. K. and S. Pal (2012). Mining educational data to analyze students' performance. *CoRR abs/1201.3417*.

Baumes, J., M. Goldberg, M. Magdon-Ismail, and A. Wallace (2004). Discovering hidden groups in communication networks. In *IN PROCEEDINGS OF THE 2ND NSF/NIJ SYMPOSIUM ON INTELLIGENCE AND SECURITY INFORMATICS*.

Beck, J. and B. Woolf (2000). High-level student modeling with machine learning. In *In Proceedings of Fifth International Conference on Intelligent Tutoring Systems*, pp. 584–593.

Biggs, N. (1976). *Graph theory, 1736-1936*. New York, USA: Oxford University Press.

Blondel, V., J. Guillaume, R. Lambiotte, and E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment 10*, 8.

Bollobas, B. (1998). *Modern Graph Theory*. Springer.

Bott, H. (1928). Observation of play activities in a nursery school. *Genetic Psychology Monographs 4*, 44–88.

Boyd, D. and N. Ellison (2010). Social network sites: definition, history, and scholarship. *Engineering Management Review, IEEE*, 16–31.

Brandão, M. F. R., C. R. S. Ramos, and B. T. Tróccoli (2006). Análise de agrupamento de escolas e núcleos de tecnologia educacional: mineração na base de dados de avaliação do programa nacional de informática na educação. pp. 366–374.

Castro, F., A. Vellido, A. Nebot, and F. Mugica (2007). Applying Data Mining Techniques to e-Learning Problems. In *Evolution of Teaching and Learning Paradigms in Intelligent Environment*, Studies in Computational Intelligence, pp. 183–221.

Chang, Y., W. Kao, C. Chu, and C. Chiu (2009). A learning style classification mechanism for e-learning. *Computers and Education 53*, 273–285.

Chen, C., C. Hong, and C. C. (2008). Mining interactive social network for recommending appropriate learning partners in a web-based cooperative learning environment. In *Cybernetics and Intelligent Systems, 2008 IEEE Conference on*, pp. 642–647.

Chen, W.-Y., J.-C. Chu, J. Luan, H. Bai, Y. Wang, and E. Y. Chang (2009). Collaborative filtering for orkut communities: discovery of user latent behavior. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pp. 681–690.

Clauset, A., C. Moore, and M. E. J. Newman (2008). Hierarchical structure and the prediction of missing links in networks. *Nature 453*(7191), 98–101.

Costa, L. F., F. A. Rodrigues, G. Travieso, and P. R. V. Boas (2007). Characterization of complex networks: A survey of measurements. In *Advances in Physics*.

Deo, N. (1974). *Graph theory with applications to engineering and computer science*. Englewood Cliffs, New Jersey, USA: Prentice Hall.

Drachsler, H., H. G. K. Hummel, and R. Koper (2008). Personal recommender systems for learners in lifelong learning networks: the requirements, techniques and model. *Int. J. Learn. Technol.*, 404–423.

Farzan, R. and P. Brusilovsky (2006). P.: Social navigation support in a course recommendation system. In *In proceedings of 4th International Conference on Adaptive Hypermedia and Adaptive Web-based Systems*, pp. 91–100.

Fortunato, S. and A. Lancichinetti (2009). Community detection algorithms: a comparative analysis: invited presentation, extended abstract. In *Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools*, VALUETOOLS '09, pp. 27:1–27:2.

Freeman, L. (2006). *The Development of Social Network Analysis*. Vancouver, Canada: Empirical Press.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks 1*(3), 215–239.

Freitas, C. M. D. S., L. C. Nedel, L. P.and Lamb, A. S. Spritzer, S. Fujiti, J. P. M. Oliveira, R. M. Araújo, and M. M. Moro (2008). Extração de conhecimento e análise visual de redes sociais. *Anais do XXVIII Congresso da SBC*.

Getoor, L. and C. P. Diehl (2005). Link mining: a survey. *SIGKDD Exploration Newsletter 7*(2), 3–12.

Girvan, M. and M. E. J. Newman (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences 99*(12), 7821–7826.

Gomes, J. E. A., R. B. C. Prudêncio, L. Meira, A. Azevedo Filho, A. C. A. Nascimento, and H. Oliveira (2013). Profiling for understanding educational social networking. *Software Engineering and Knowledge Engineering (SEKE 2013)*.

Goulden, C. (1956). *Methods of Statistical Analysis*. New York, USA: John Wiley I& Sons Inc.

Grob, H., F. Bensberg, and F. Kaderali (2004). Controlling open source intermediaries - a web log mining approach. In *Information Technology Interfaces, 2004. 26th International Conference on*, pp. 233–242.

Ha, S., S. Bae, and S. Park (2000). Web mining for distance education. *Management of Innovation and Technology, 2000. ICMIT 2000. Proceedings of the 2000 IEEE International Conference on 2*(1), 715–719.

Heraud, j., L. France, and A. Mille (2004). Pixed: an its that guides students with the help of learners interaction log. In *In International*, pp. 57–64.

Hämäläinen, W., T. H. Laine, and E. Sutinen (2004). Data mining in personalizing distance education courses. In *In World*.

Khribi, M., M. Jemni, and O. Nasraoui (2008). Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. In *Advanced Learning Technologies, 2008. ICALT '08. Eighth IEEE International Conference on*, pp. 241–245.

Klösgen, W. and J. Zytkow (2002). *Handbook of data mining and knowledge discovery*. New York, NY, USA: Oxford University Press, Inc.

Kotsiantis, S., C. Pierrakeas, and P. Pintelas (2003). Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems*, pp. 267–274.

Kumar, R., J. Novak, and A. Tomkins (2006). Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 611–617.

Lee, M., S. Chen, K. Chrysostomou, and X. Liu (2009). Mining students behavior in web-based learning programs. *Expert Systems with Applications*, 3459–3464.

Lehmann, E. (1975). *Non-parametrics: Statistical Methods Based on Ranks*. San Francisco, USA: Springer edition.

Lemire, D., H. Boley, S. McGrath, and M. Ball (2005). Collaborative filtering and inference rules for context-aware learning object recommendation. *International Journal of Interactive Technology and Smart Education 2*.

Leskovec, J., K. J. Lang, A. Dasgupta, and M. W. Mahoney (2008). Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pp. 695–704.

Liu, H. and H. Motoda (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Norwell, MA, USA: Kluwer Academic Publishers.

Lu, J. (2004). Personalized e-learning material recommender system. In *In: Proc. of the Int. Conf. on Information Technology for Application*, pp. 374–379.

Markham, S., J. Ceddia, J. Sheard, C. Burvill, J. Weir, B. Field, L. Sterling, and L. Stern (2003). Applying agent technology to evaluation tasks in e-learning environments. *Proceedings of the Exploring Educational Technologies Conference*.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology 27*, 415–444.

Mei, Q., D. Cai, D. Zhang, and C. Zhai (2008). Topic modeling with network regularization. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pp. 101–110.

Meira, L., A. M. M. Neves, and G. Ramalho (2009). Lan house na escola: uma olimpíada de jogos digitais e educação. *Brazilian Symposium on Games and Digital Entertainment 8*(8), 150–157.

Mills, N. (2011). Situated learning through social networking communities: The development of joint enterprise, mutual engagement, and a shared repertoire. *Journal CALICO*, 345–368.

Mostow, J. and J. Beck (2006). Some useful tactics to modify, map and mine data from intelligent tutors. *Nat. Lang. Eng. 12*, 195–208.

Muehlenbrock, M. (2005). Automatic Action Analysis in an Interactive Learning Environment. International Conference on Artificial Intelligence in Education. Disponível em http://outlier-0902326273/outlier.pdf. Último acesso em 14/03/2013.

Mui, Y. Q. and P. Whoriskey (2010). Facebook passes google as most popular site on the internet, two measures show. *The Washington Post*.

Newman, M. E. J. (2003). The structure and function of complex networks. *Society for Industrial and Applied Mathematics Review 45*(2), 167–256.

Newman, M. E. J. (2006). From the cover: Modularity and community structure in networks. *Proceedings of the National Academy of Science 103*, 8577–8582.

NIELSEN (2009). Social networks & blogs now 4th most popular online activity. Technical report. NIELSEN ONLINE REPORT. Disponível em http://www.nielsen.com/us/en/press-room/2009/social_networks__.html. Último acesso em 10/06/2013.

Pimentel, E. and N. Omar (2006). Descobrindo conhecimentos em dados de avaliacao aprendizagem com tecnicas de mineracao de dado. *Workshop sobre Informática na Escola. Anais do Congresso da Sociedade Brasileira de Computação*, 147–155.

Purcell, M. A. (2010). *The Networked Library: A Guide for the Educational Use of Social Networking Sites*. New York, USA: Tech Tools for Learning.

Reffay, C. and T. Chanier (2003). How social network analysis can help to measure cohesion in collaborative distance-learning. In *international conference on computer support for collaborative learning*, pp. 343–352.

Regueiro, R. (2009). *Redes Sociais na Internet*. Porto Alegre, BRA: Editora Meridional.

Romero, C. and S. Ventura (2007). Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl. 33*(1), 135–146.

Romero, C., S. Ventura, and P. Bra (2004). Knowledge discovery with genetic programming for providing feedback to courseware author. user modeling and user-adapted interaction: The. *Journal of Personalization Research*, 425–464.

Senot, C., D. Kostadinov, M. Bouzid, J. Picault, A. Aghasaryan, and C. Bernier (2010). Analysis of strategies for building group profiles. In *User Modeling, Adaptation, and Personalization*, Volume 6075, pp. 40–51. Springer Berlin Heidelberg.

Shapiro, S. S. and M. B. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika 3*(52).

Silva, A. B. O., F. S. Parreiras, R. F. Matheus, and W. C. Brandão (2007). Redes de coautoria dos professores da ciência da informação: um retrato da colaboração científica dessa disciplina no brasil. *Encontros Bibli 7*, 441–452.

Silva, S. R. P. and R. Pereira (2008). Aspectos da interação humano-computador na web social. In *Proceedings of the VIII Brazilian Symposium on Human Factors in Computing Systems*, pp. 350–351.

Solomonoff, R. and A. Rapoport (1951). Connectivity of random nets. *Bulletin of Mathematical Biology 13*, 107–117.

Sun, J., C. Faloutsos, S. Papadimitriou, and P. S. Yu (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*, pp. 687–696.

Superby, J., J. Vandamme, and N. Meskens (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. In *inProc. Int. Conf. Intell. Tutoring Syst. Workshop Educ. Data Mining*, pp. 1–8.

Talavera, L. and E. Gaudioso (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. In *Workshop on AI in CSCL*, pp. 17–23.

Tang, C., R. W. H. Lau, Q. Li, H. Yin, T. Li, and D. Kilis (2000). Personalized courseware construction based on web data mining. In *Web Information Systems Engineering, 2000. Proceedings of the First International Conference on*, pp. 204–211 vol.2.

Tang, L. and H. Liu (2010a). *Community Detection and Mining in Social Media*. New York, USA: Morgan & Claypool.

Tang, L. and H. Liu (2010b). Understanding group structures and properties in social media. In *Link Mining: Models, Algorithms, and Applications*, pp. 163–185. Springer New York.

Tang, L., X. Wang, and H. Liu (2011). Group profiling for understanding social structures. *ACM Trans. Intell. Syst. Technol. 3*, 15:1–15:25.

Tang, T. Y. and G. McCalla (2002). Student modeling for a web-based learning environment: a data mining approach. In *Eighteenth national conference on Artificial intelligence*, pp. 967–968.

TERRA (2010). Brasil: 90 Technical report. TERRA. Disponível em http://tecnologia.terra.com.br/internet/brasil-90-das-empresas-usam-redes-sociais-para-negocios,d81aeeb4bddea310VgnCLD200000bbcceb0aRCRD.html. Último acesso em 11/06/2013.

Wang, X., L. Tang, H. Gao, and H. Liu (2010). Discovering overlapping groups in social media. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, pp. 569–578.

Wasserman, S. and K. Faust (1994). *Social Network Analysis: methods and applications (structural analysis in the social sciences)*. New York, USA: Cambridge University Press.

Watts, D. J. and S. H. Strogatz (1998). Collective dynamics of small-world networks. *Nature 393*(6684), 440–442.

Wellman, B. (1926). The influence of intelligence on the selection of associates. *Journal of Educational Research 14*, 126–132.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 80–83.

Wild, C. J. (2000). *Chance encounters: A first course in data analysis and inference*. New York, USA: John Wiley & Sons.

Yacef, K., O. Zaïane, A. Hershkovitz, M. Yudelson, and J. Stamper (2012). Proceedings of the 5th international conference on educational data mining. In *EDM*.

Yuruk, N., X. Xu, and T. A. J. Schweiger (2007). Structure-connected as functional groups in metabolic networks. In *Proceedings of the International Workshop and Conference on Network Science*, NetSci '07.

Zaïane, O. R. (2001). Web usage mining for a better web-based learning environment.

Zorrilla, M. E., E. Menasalvas, D. Marín, E. Mora, and J. Segovia (2005). Web usage mining project for improving web-based learning sites. In *Proceedings of the 10th international conference on Computer Aided Systems Theory*, pp. 205–210.

Anexos

A

Testes Estatísticos Utilizados

A.1 Wilcoxon Rank Sum

O *Wilcoxon Rank Sum* (WRS) é um método não-paramétrico para o estabelecimento de diferença significante entre dois grupos de amostras não pareadas, utilizando magnitude baseado em ordenações (*ranks*). O Teste é um alternativa ao teste *t-student*, para comparar as médias de duas amostras independentes quando os dados não encontram-se em uma distribuição normal (Mei et al., 2008). Uma extensão deste método é o teste *Kruskal-Wallis*, que extende o WRS para mais de duas amostras.

1. Ordenar todas as observações em ordem crescente de magnitude, ignorando qual grupo eles vêm. Se as duas observações têm a mesma magnitude, independente do grupo, a estas são atribuídas a classificação média;

2. Adicione todas as fileiras associadas com as observações de menor grupo;

3. Finalmente, o P valor associado com a estatística de Wilcoxon é identificado.

Seguindo a implementação do WRS da ferramenta MATLAB, mais especificamente o método *ranksum*, utilizada nos experimentos deste trabalho; a aplicação do método WRS para a identificação de caracterizadores de grupos, é definido como: Sejam n_x o tamanho da amostra do grupo e n_y do restante da rede, a estatística do teste retornada pela função *ranksum* é a soma dos rankings da primeira amostra (grupo). O método *ranksum*, assume que as duas amostras são independentes. Onde X e Y podem ter tamanhos diferentes.

A estatística do teste WRS, U, é o número de vezes que uma amostra Y precede uma X em um arranjo ordenado de elementos nas duas amostras independentes X e Y. Estando relacionado com a estatística do teste WRS, da seguinte forma: Se X é uma amostra de tamanho n_x, então,

$$U = W - \frac{n_x(n_x - 1)}{2} \tag{A.1}$$

Todavia, para amostras grandes, onde n_x e n_y têm tamanhos iguais ou superior a 10 observações, a função *ranksum* utiliza uma estatística z para calcular o p valor aproximado do teste. Dessa forma, sendo X (amostra do grupo) e Y (amostras do restante da rede), duas amostras independentes de tamanho n_x e n_y, onde n_x<ny a estatística z é:

$$z = \frac{W - E(W)}{\sqrt{V(W)}} = \frac{W - \left[\frac{n_x n_y + n_x(n_x+1)}{2}\right] - 0.5 * sing(W - E(W))}{\sqrt{\frac{n_x n_y (n_x + n_y + 1) - tiescor}{12}}}, \tag{A.2}$$

com correção de continuidade e ajuste de "empate". Dessa forma, *Tiescor* é dada pela Função A.3:

$$tiescor = \frac{2 * tiead\,j}{(n_x + n_y)(n_x + n_y - 1)}, \tag{A.3}$$

Dessa forma, a função *ranksum* usa [ordenações,tieadj] = tiedrank(x,y) para obter os ajustes dos "empates". Onde a função *tiedrank* calcula as classificações dos valores nos vetores X e Y. Se os valores da amostra X ou Y são "empatados", a função *tiedrank* calcula sua posição média. O valor retornado pela função é um ajuste para os "empates", exigidos por testes não paraméticos, como o WRS. Assim como, para se calcular a correlação de Spearman.

A partir do p valor identificado é considerando uma medida de significância ou intervalo de confiança (α), o teste WRS revela se uma amostra (grupo) é significantemente diferente de outra (restante da rede), ou se não há diferença relevante. Em estatística, intervalos de confiança são usados para indicar a confiabilidade de uma estimativa, logo quanto menor o α considerado mais confiável será o resultado obtido. Em nosso estudo o nível de significância mede a probabilidade de rejeição da hipótese nula (HO). Assim, todas as vez que o p valor é superior ao α estabelecido, a HO não pode ser rejeitada; caso contrário, as amostras são significantemente diferentes.

A.2 Shapiro-Wilk

Na estatística, o Teste *Shapiro-Wilk* testa a hipótese nula de que a amostra vem de uma população com distribuição normal. Foi publicado em 1965 por Samuel Shapiro e Martin Wilk (Shapiro and Wilk, 1965). Em estatística, o usuário deve rejeitar a hipótese nula caso o p-valor seja menor que o nível de significância.

O teste *Shapiro-Wilk*, calcula uma variável estatística (W) que investiga se uma amostra aleatória provém de uma distribuição normal. A variável W é calculada da seguinte forma:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{A.4}$$

sendo,

- - x_i os valores ordenados de amostras (x_1 é o menor).
- a_i constantes geradas a partir de meio, variâncias e covariâncias da ordem estatística de uma amostra de tamanho n e uma distribuição normal.

Sendo X uma característica em estudo, então formula-se as hipóteses:

- H_0: X tem distribuição Normal;
- H_1: X não tem distribuição Normal.