Análise do Engajamento dos Alunos em Ambientes Virtuais de Aprendizagem para detecção de comunidade

Bernadete Aquino[1,2], Victor Ströele[1], Jairo Francisco de Souza[1,2]

[1]Programa de Pós Graduação em Ciência da Computação
Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora –MG – Brasil

[2]LApIC Research Group
Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora –MG – Brasil

{baquino, victor.stroele, jairo.souza}@ice.ufjf.br

Abstract. *There is an increasing number of courses taught using virtual learning environments. However, these environments face the challenge of keeping students motivated and engaged. A huge amount of data is generated by these environments that can be used to discover patterns regarding student engagement. In this work, a model based on undirected graphs was proposed, where the elements represent students and their connections are similarity in their behavior. The Label Propagation algorithm was used to group students based on three engagement metrics. A quantitative analysis was carried out to identify students who are not engaged who may need help. The results point to a significant difference concerning students' actions that represent the phenomenon of engagement between students with better and worse performance.*

Resumo. *Há um número crescente de cursos ministrados utilizando ambientes virtuais de aprendizagem. Contudo, esses ambientes enfrentam o desafio de manter alunos motivados e engajados. Uma grande quantidade de dados é gerada por esses ambientes e podem ser usados para análise de comportamento dos alunos. Neste trabalho, é proposto um modelo baseado em grafos não direcionados para identificar o engajamento de alunos. O algoritmo Label Propagation foi utilizado para agrupar os alunos com base em três métricas de engajamento. Uma análise quantitativa foi realizada para identificar os alunos não engajados que possam precisar de intervenção. Os resultados apontam para uma diferença significativa em relação às ações dos alunos que representam o fenômeno do engajamento entre os alunos de melhor e pior desempenho.*

1. Introdução

Os últimos anos testemunharam um aumento significativo nas matrículas nos cursos *on-line*. O aprendizado *on-line* acomoda melhor as diversas necessidades dos alunos, quebrando as barreiras geográficas e físicas. Apesar desse rápido desenvolvimento, o controle da retenção e de atitudes positivas dos alunos em relação ao aprendizado *on-line* se tornou um dos principais desafios [Zhu et al. 2020]. As pesquisas descritas em [Oliveira et al. 2019b, Vytasek et al. 2020] apontam uma lacuna na literatura e uma escassez de trabalhos com foco no entendimento do fenômeno de engajamento com o objetivo de gerar dados importantes para uma mudança de postura dos alunos.

DOI: 10.5753/cbie.sbie.2020.952

O engajamento é um conceito multifacetado que engloba aspectos comportamentais, emocionais e cognitivos [Muir et al. 2019]. O engajamento comportamental refere-se à participação e inclui o envolvimento em atividades acadêmicas, sociais ou extracurriculares. O engajamento emocional abrange reações afetivas a professores, colegas de classe e à instituição em que o aprendizado ocorre. Finalmente, o engajamento cognitivo incorpora consideração e disposição para exercer o esforço de compreender o assunto e as habilidades principais. Entender como o engajamento ocorre e gerar indicadores para avaliar os alunos possibilitam intervenções que evitam que eles percam a motivação e o engajamento [Oliveira et al. 2019b].

A mineração de dados educacionais baseada em grafos pode ser utilizada para analisar cenários educacionais passíveis de serem representados por redes complexas. A abordagem de rede complexa não é apenas útil para simplificar e visualizar essa quantidade enorme de dados interconectados, mas também é eficaz na identificação de elementos-chave do sistema e na descoberta de suas interações mais importantes, além de descobrir agrupamentos de características da rede [Oliveira et al. 2019a].

A análise de agrupamento é o processo de agrupar itens em grupos, chamados de *clusters*, para que os objetos dentro do *cluster* tenham uma similaridade muito alta e, ao mesmo tempo, sejam diferentes dos objetos dos demais *clusters* [Saraiya and Ganage 2018]. O princípio básico é garantir alta similaridade *intra-cluster* e manter uma dissimilaridade *inter-cluster*.

A análise de agrupamento pode ajudar os instrutores do curso a identificar os alunos que não estão engajados e melhorar as taxas de aprendizado e retenção. Assim, essa pesquisa aplica mineração de dados em grafos nas interações dos alunos referentes ao engajamento comportamental a fim de gerar indicadores que auxiliem gestores na tomada de decisões com base em evidências. Para isso, é apresentada uma modelagem de uma rede complexa com intuito de representar e armazenar informações detalhadas dos alunos e empregar a análise de agrupamento por meio da utilização do *Label Propagation Algorithm* (LPA). Esse modelo é utilizado com o objetivo de extrair informações pedagógicas que contribuirão para entender como o engajamento ocorre e gerar indicadores para avaliar os alunos.

Com finalidade de alcançar esse objetivo, as seguintes etapas foram realizadas: coleta dos dados brutos que são os registros de ações realizadas pelos alunos nos Ambientes Virtuais de Aprendizagem (AVA), o pré-processamento que é o processo de preparação, organização e estruturação desses dados, a construção do modelo de dados necessário para aplicar a técnica de mineração de dados, e, por fim, a geração do grafo considerado para a análise dos resultados.

O trabalho está organizado como se segue. Na seção 2 é descrito o referencial teórico para uma melhor compreensão do assunto, além de serem apresentadas exemplos de utilização de redes complexas no contexto educacional. Na seção 3, é apresentada a abordagem proposta para geração da rede, além disso faz uma análise da rede gerada. Nas seções 4 e 5 são apresentados os resultados e as considerações finais, respectivamente.

2. Trabalhos Relacionados

A Mineração de Dados Educacionais (MDE) é um campo de pesquisa que se concentra na aplicação de mineração de dados, aprendizado de máquina e méto-

dos estatísticos para detectar padrões em grandes coleções de dados educacionais [Hernández-Blanco et al. 2019]. A mineração de dados baseada em grafos e a análise de dados educacionais tornaram-se disciplinas emergentes na MDE [Lynch et al. 2017]. Nesta seção, são apresentadas aplicações da abordagem de redes complexas para extração de informações a partir de dados educacionais.

Um levantamento realizado em [Oliveira et al. 2019a] sobre aplicações no domínio educacional que implementam mineração de dados baseados em grafos identificou 30 trabalhos. Em aproximadamente 70% desses trabalhos, a pesquisa se concentra na análise de interações dos alunos e seus comportamentos em ambientes virtuais de aprendizagem. Nesse grupo, estão os artigos que analisam as trocas de mensagens em fóruns [Gitinabard et al. 2017, Brown et al. 2015, Kovanovic et al. 2014, Jiang et al. 2014], avaliação da rede de alunos matriculados nos mesmos cursos de uma instituição [Gardner and Brooks 2018], extração de caminhos de aprendizado frequentes que muitos alunos seguiram em uma plataforma de aprendizado *on-line* [Patel et al. 2017] e identificação de padrões de retenção em um programa de pós-graduação a fim de diminuir a taxa de evasão dos alunos [Costa et al. 2019].

A representação por meio de grafos tem se mostrado uma técnica eficiente de visualização e representação, principalmente quando envolve bancos de dados que apresentam claramente relações entre entidades. A relevância das mídias sociais e da educação *on-line*, onde a interação entre os atores é naturalmente representada por dados gráficos, leva a pesquisas que visam extrair informações pedagógicas que contribuirão para melhorar o sucesso acadêmico. Portanto, é esperado um aumento para pesquisas nesse domínio [Oliveira et al. 2019a].

Em [Oliveira et al. 2019b], os autores utilizaram técnicas de mineração para mostrar a relação do engajamento dos alunos no AVA com o desempenho dos mesmos. As ações dos alunos são armazenadas e essas ações são usadas para entender padrões e comportamentos. No mapeamento das melhores técnicas utilizadas para se visualizar engajamento, em especial nos fóruns, os cinco artigos levantados trabalharam com análise de redes sociais representadas por grafos. Já [Moubayed et al. 2018] utilizou o *Apriori algorithm* para gerar um conjunto de regras que relacionam o engajamento com o desempenho acadêmico. Os resultados experimentais mostraram que existe uma alta correlação positiva entre o engajamento e o desempenho acadêmico. Devido a esta correlação positiva, o engajamento pode ser usado como um preditor do desempenho acadêmico dos alunos. Isso, por sua vez, pode ser usado para identificar os alunos não engajados que podem precisar de ajuda com o curso e trabalhar com eles para melhorar seu engajamento e possivelmente seu desempenho.

Pode ser observado que a análise de redes no domínio da educação é bastante utilizada pela comunidade acadêmica. Isso se deve ao fato de permitir a extração de conhecimento muitas vezes implícitos nos ambientes virtuais de aprendizagem. Contudo, não foram encontrados trabalhos que utilizam mineração de dados em grafos para entender o engajamento dos alunos, gerando indicadores que possibilitem a avaliação e acompanhamento dos mesmos. O trabalho proposto busca explorar diferentes comportamentos dos alunos durante o curso *on-line* e identificar seus impactos sobre a performance acadêmica.

3. Modelagem e Análise da Rede

Este trabalho faz uso do estudo de [Chai et al. 2019], o qual levantou as principais variáveis utilizadas para entender padrões e comportamentos dos alunos que representam o fenômeno de engajamento. Para revelar tais variáveis, os autores utilizaram um algoritmo de classificação de variáveis baseado em *Decision Tree* para realizar a triagem preliminar que forma um conjunto de atributos candidatos. Em seguida, foi utilizada a *Recursive Feature Elimination* para classificar os atributos candidatos por sua importância. Os 3 principais fatores de influência descobertos foram:

1. *Submit Rate*: representa a porcentagem de avaliações realizadas pelos alunos em relação a quantidade de avaliações propostas no curso. Se o número de avaliações que o aluno enviou for $|z_i|$ e o o número de avaliações que o curso possui é Z, então:
$$SubmitRate = \frac{|z_i|}{Z} \tag{1}$$
2. *Submit Time*: retrata a pontualidade do aluno nas avaliações enviadas. Assumindo que S_{ik} é o momento em que o aluno submeteu a avaliação k, e L_k é o prazo final da avaliação k, então:
$$SubmitTime = \frac{1}{|z_i|}\sum_{k=1}^{|z_i|}(L_k - S_{ik}) \tag{2}$$
3. *Number Of Times Material Is Accessed*: corresponde a quantidade de vezes que o aluno acessou algum material do curso. Assumindo que M_i é o conjunto de todas os acessos aos materiais realizados pelo aluno i, então:
$$NumberOfTimesMaterialIsAccessed = |M_i| \tag{3}$$

Como base de dados, foi utilizado o conjunto de dados desenvolvido pela *Open University*[1] para apoiar pesquisas na área educacional. Esse conjunto de dados inclui informações dos alunos, os resultados das avaliações e registros de suas interações com os materiais em formato bruto e tabular. Foi utilizado os dados do curso AAA, o qual possui o maior número de dados: 357 alunos, 5 avaliações e 202 materiais. A Figura 1 apresenta a distribuição de alunos por *status* de conclusão no curso selecionado para o estudo.

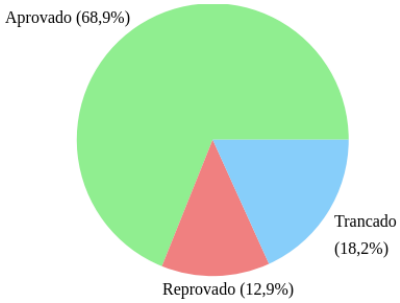

Figura 1. Distribuição de Alunos por Status no Curso

[1] https://www.kaggle.com/rocki37/open-university-learning-analytics-dataset

As 3 variáveis que definem o engajamento foram utilizadas para a criação da rede, onde os nós representam os alunos e as arestas possuem pesos que representam o nível de similaridade entre o engajamento dos alunos no curso. Essa rede pode ser extraída de qualquer ambiente virtual de aprendizagem e sua análise é útil no monitoramento dos alunos e na avaliação de sua participação no curso.

Para a construção da rede, primeiramente, foram coletadas as informações necessárias para o cálculo das variáveis relacionadas ao engajamento. Em seguida, essas informações foram utilizadas para o cálculo das variáveis. Como esses valores estão em escalas diferentes, foram normalizados. Os dados consolidados foram organizados em uma tabela que possui os atributos *submitRate, submitTime, numberOfTimesMaterialIsAccessed* para cada aluno do curso. A fim de criar relacionamentos entre os alunos, foi calculada a similaridade do cosseno (*Cosine Similarity*) para cada par de alunos utilizando os atributos dessa tabela. O algoritmo *Cosine Similarity* calcula a similaridade entre dois vetores numéricos, quanto maior o valor retornado mais parecidos são os objetos. Com o objetivo de reduzir a quantidade de relacionamentos do grafo, somente os alunos com similaridade maior que 0,9 foram conectados. Esse limite foi definido de maneira empírica. Experimentos com um limiar mais baixo (0,7 e 0,8) geraram um número muito alto de arestas, tiveram elevado custo computacional e não geraram diferenças significativas nos resultados.

Neste modelo, é considerado que a relação de similaridade entre os alunos é simétrica. Assim, as arestas que os ligam não possuem qualquer orientação. O grafo gerado possui 365 vértices e 11.361 arestas e pode ser descrito como $G(V, A)$, onde:

- V = {v | v é um nó no grafo que representa um aluno}
- A = {(v, w, s) | <há similaridade maior que 0,9 entre v e w, sendo s o valor da similaridade de cossenos entre as variáveis que definem o engajamento dos alunos no AVA>}

Uma análise quantitativa foi realizada nas métricas calculadas para fornecer uma melhor compreensão do envolvimento dos alunos. Neste trabalho, foi utilizado o algoritmo *Label Propagation Algorithm* (LPA) para agrupar os alunos em *clusters* com base em suas atividades e interações *on-line*. A Figura 2 apresenta uma parte da rede gerada com as comunidades identificadas pelo algoritmo LPA. Ao todo foram identificadas sete diferentes comunidades.

O algoritmo LPA é um dos mais comuns no campo da detecção de comunidades [Jokar and Mosleh 2019]. O algoritmo foi selecionado por possuir complexidade de tempo linear. A abordagem utilizado pelo LPA: na primeira etapa, cada nó da rede recebe um rótulo exclusivo e, em um processo de repetição, o algoritmo organiza os nós da rede em uma ordem aleatória. Em seguida, cada nó na lista ordenada recebe o rótulo transportado pelo maior número de nós vizinhos (se esse rótulo máximo não for exclusivo, o LPA selecionará aleatoriamente um deles para atualizar o rótulo do nó). Esse processo de repetição continua até chegar ao ponto em que o rótulo de todos os nós da rede é rotulado com o maior número de vizinhos e, portanto, nenhuma alteração é feita nos rótulos. No final do algoritmo, todos os nós com os mesmos rótulos são colocados em uma comunidade ou *cluster* [Jokar and Mosleh 2018].

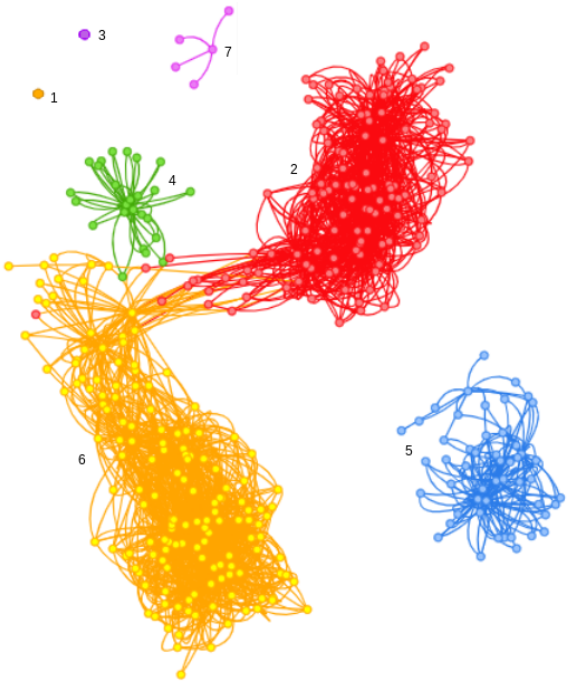

Figura 2. Comunidades Geradas com LPA

4. Resultados e Discussões

O gráfico da Figura 3 apresenta a distribuição de alunos por *status* de conclusão entre as sete comunidades identificadas. As comunidades 2 e 6 são compostas em sua grande maioria por alunos que obtiveram sucesso no curso. Já as comunidades 4 e 5 são formadas em sua maior parte por alunos que trancaram ou falharam no curso. As demais comunidades representam menos de 2% da quantidade total de alunos no curso. Isso mostra que os alunos que obtiveram o mesmo resultado no curso tiveram alta densidade de arestas entre si, com menor número de arestas para os demais grupos.

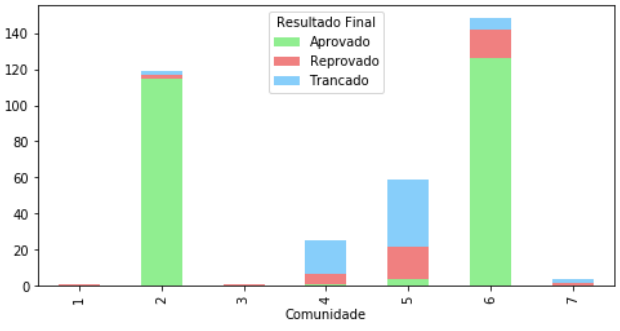

Figura 3. Distribuição dos Alunos entre as Comunidades

A Figura 4 apresenta a distribuição por grau dos alunos na rede. São apresentados os grupos de alunos que passaram, falharam ou trancaram o curso. Esse gráfico revela uma diferença significativa entre o grau (número de relacionamentos) de alunos de baixo e alto desempenho. O que indica que os alunos de alto desempenho (aprovados) desenvolvem redes maiores do que os de baixo desempenho (falha ou trancamento). Os

alunos engajados tendem a participar com mais frequência das atividades do curso para se manterem atualizados com os requisitos e evitarem possível falha.

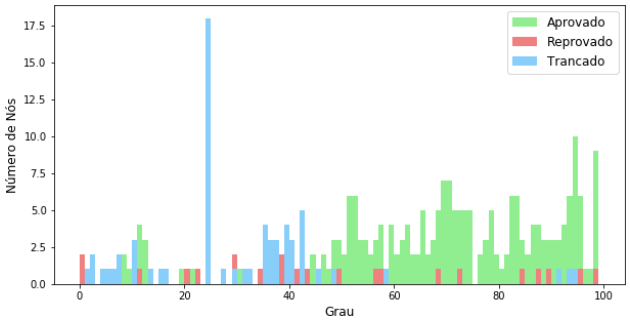

Figura 4. Distribuição dos Alunos por Grau

Na Figura 5, são apresentados dois gráficos com os valores das três variáveis utilizadas para medir a similaridade dos alunos quanto ao engajamento, um agrupado por *status* final no curso, outro pelas comunidades identificadas. É possível observar que a variável que representa o engajamento que mais contribuiu para o desempenho positivo dos alunos foi a *submitRate*. A grande maioria dos alunos com valores maiores que 0,2 para essa variável foram também aprovados.

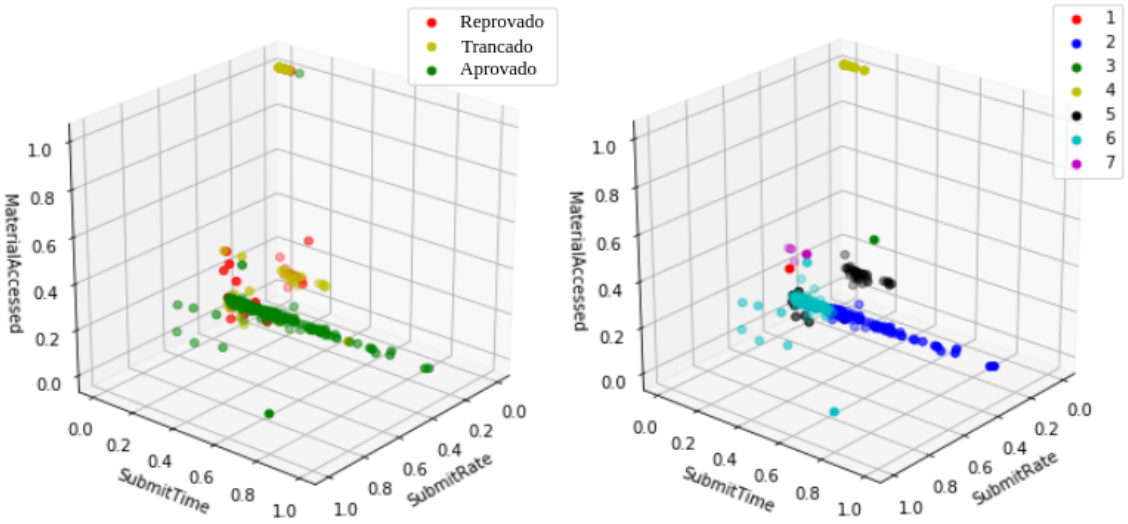

Figura 5. Comunidades Geradas com LPA

Foi realizada uma comparação entre os alunos das diferentes comunidades encontradas através das seguintes características: nota final, número de materiais acessados, grau dos nós e número de acessos ao ambiente virtual de aprendizagem. A Figura 6 apresenta essa comparação. As comunidades 1 e 3 são compostas por somente um aluno cada, ambos os alunos não possuem relacionamentos e falharam no curso. Estão em comunidades diferentes devido à quantidade de acessos aos materiais. Já as comunidades 2 e 6, em sua maioria, são compostas por alunos aprovados no curso, diferem entre si devido às notas obtidas, número de materiais acessados e número de acessos ao ambiente. Esses números são maiores na comunidade 2 que representa, assim, os alunos mais engajados

e com melhor desempenho. Contudo, os maiores valores para o grau estão na comunidade 6, o que nos mostra que, entre os alunos aprovados, a maioria possui desempenho médio. Por último, as comunidades 4, 5 e 7 são formadas, em sua grande maioria, por alunos que falharam ou trancaram o curso. As comunidades 4 e 7 representam os alunos com menor engajamento, menor número de acesso aos materiais e ao ambiente. Por fim, a comunidade 5 possui um engajamento semelhante ao da comunidade 6, a qual representa os alunos aprovados. Isso mostra que pequenas intervenções junto aos alunos da comunidade 5 poderia ajudá-los a obterem êxito no curso.

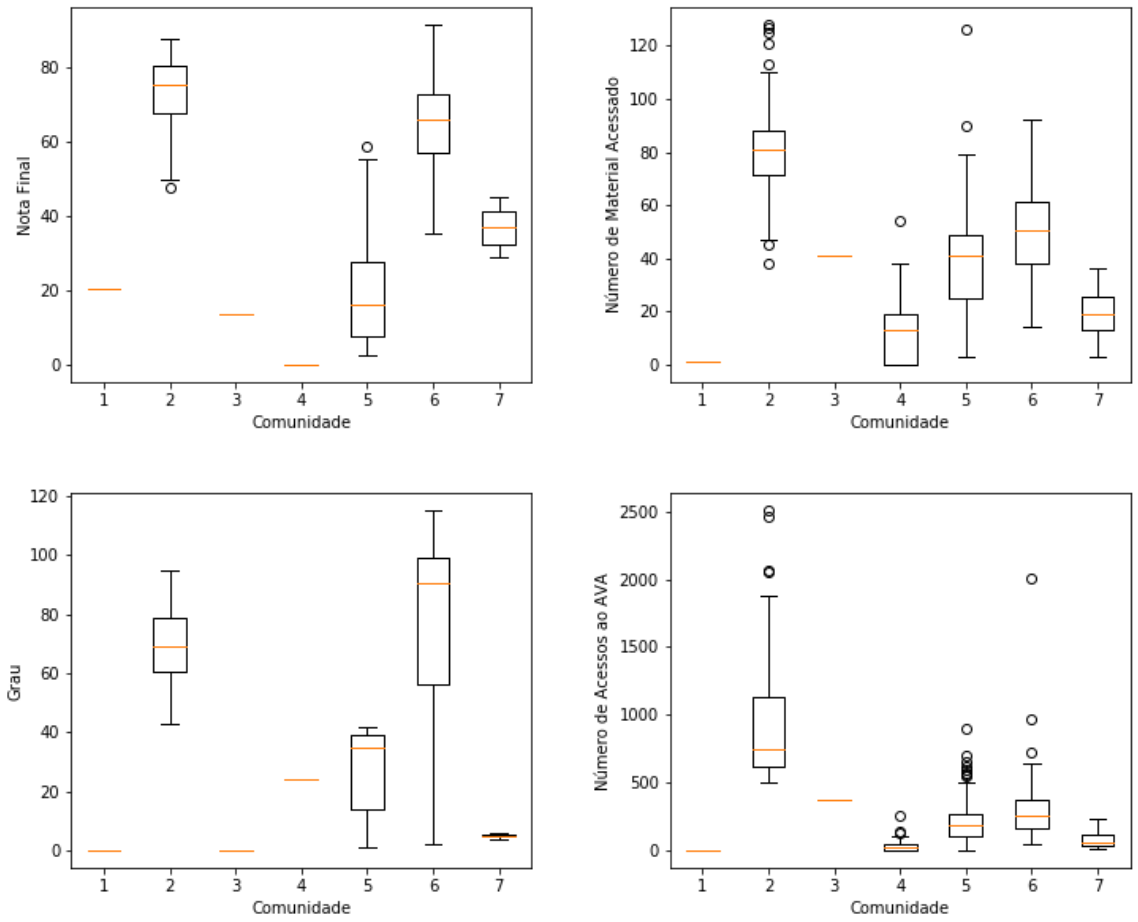

Figura 6. Características das Comunidades

5. Considerações Finais

Com a expansão da tecnologia, o *e-learning* tornou-se uma área de grande interesse. Por outro lado, manter os alunos engajados e motivados nesses ambientes, tornou-se um desafio. Esse trabalho propôs o uso do algoritmo LPA para agrupar os alunos em diferentes grupos de níveis de engajamento e forneceu uma base para identificar alunos não engajados por meio de suas interações com o AVA. O uso dessas interações como base para a identificação dos alunos não engajados proporciona aos instrutores a chance de se comunicar mais com esses alunos individualmente para assim, discutir e identificar as situações que possam estar prejudicando seu desempenho ou diminuindo sua motivação e engajamento.

Esse artigo possui limitações que precisam ser exploradas em estudos futuros. O nível de similaridade entre os alunos para a criação da rede foi definido de maneira empírica, considerando somente o número total de arestas geradas. Como trabalhos futuros, pretende-se analisar a rede gerada para verificar se pequenas alterações nesse valor gerariam comunidades mais homogêneas. Além disso, pretende-se verificar o modelo proposto em um curso diferente do investigado para analisar se o mesmo poderá ser generalizado. Ainda, para avaliar melhor o engajamento dos alunos, um número maior de variáveis que representam o engajamento poderão ser utilizadas.

Referências

[Brown et al. 2015] Brown, R., Lynch, C., Wang, Y., Eagle, M., Albert, J., Barnes, T., Baker, R., Bergner, Y., and McNamara, D. (2015). Communities of performance & communities of preference. In *CEUR Workshop Proceedings*, volume 1446. CEUR-WS. 8th International Conference on Educational Data Mining, EDM 2015 ; Conference date: 26-06-2015 Through 29-06-2015.

[Chai et al. 2019] Chai, Y., Lei, C., and Yin, C. (2019). Study on the influencing factors of online learning effect based on decision tree and recursive feature elimination. pages 52–57.

[Costa et al. 2019] Costa, J., Bernardini, F., Artigas, D., and Viterbo, J. (2019). Mining direct acyclic graphs to find frequent substructures — an experimental analysis on educational data. *Information Sciences*, 482.

[Gardner and Brooks 2018] Gardner, J. and Brooks, C. (2018). Coenrollment networks and their relationship to grades in undergraduate education. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, LAK '18, page 295–304, New York, NY, USA. Association for Computing Machinery.

[Gitinabard et al. 2017] Gitinabard, N., Xue, L., Lynch, C., Heckman, S., and Barnes, T. (2017). A social network analysis on blended courses.

[Hernández-Blanco et al. 2019] Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., and Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. *Complexity*, pages 1–22.

[Jiang et al. 2014] Jiang, S., Fitzhugh, S. M., and Warschauer, M. (2014). What is the source of social capital? *Workshop on Graph-Based Educational Data Mining*, 14.

[Jokar and Mosleh 2018] Jokar, E. and Mosleh, M. (2018). Community detection in social networks based on improved label propagation algorithm and balanced link density. *Physics Letters A*.

[Jokar and Mosleh 2019] Jokar, E. and Mosleh, M. (2019). Community detection in social networks based on improved label propagation algorithm and balanced link density. *Physics Letters A*, 383.

[Kovanovic et al. 2014] Kovanovic, V., Joksimovic, S., Gasevic, D., and Hatala, M. (2014). What is the source of social capital? *Workshop on Graph-Based Educational Data Mining*.

[Lynch et al. 2017] Lynch, C., Barnes, T., Xue, L., and Gitinabard, N. (2017). Graph-based educational data mining (g-edm 2017) proceedings.

[Moubayed et al. 2018] Moubayed, A., Injadat, M., Shami, A., and Lutfiyya, H. (2018). Relationship between student engagement and performance in e-learning environment using association rules. pages 1–6.

[Muir et al. 2019] Muir, T., Milthorpe, N., Stone, C., Dyment, J., Freeman, E., and Hopwood, B. (2019). Chronicling engagement: students' experience of online learning over time. *Distance Education*, 40(2):262–277.

[Oliveira et al. 2019a] Oliveira, J., Alexandrino, R., and Ambrósio, A. (2019a). A survey of applications that use graph-based educational data mining. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, 30(1):1401.

[Oliveira et al. 2019b] Oliveira, P., Souza, A., and Rodrigues, R. (2019b). Identificação de pesquisas referentes ao engajamento de alunos em plataformas de lms e suas relações com o desempenho acadêmico. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, 30(1):1631.

[Patel et al. 2017] Patel, N., Sellman, C., and Lomas, D. (2017). Mining frequent learning pathways from a large educational dataset.

[Saraiya and Ganage 2018] Saraiya, P. R. and Ganage, Y. (2018). Study of clustering techniques in the data mining domain. *International Journal of Computer Science and Mobile Computing*, 7.

[Vytasek et al. 2020] Vytasek, J. M., Patzak, A., and Winne, P. H. (2020). *Analytics for Student Engagement*, pages 23–48. Springer International Publishing, Cham.

[Zhu et al. 2020] Zhu, Y., Zhang, J. H., Au, W., and Yates, G. (2020). University students' online learning attitudes and continuous intention to undertake online courses: a self-regulated learning perspective. *Educational Technology Research and Development*, 68:1485–1519.