

Análise de clusterização de jogadores de Counter-Strike 2

Luiz Felipe De Oliveira Barbosa Nunes (RA 255403)

1 Introdução e Materiais

Counter-Strike é um popular jogo de tiro em primeira pessoa desenvolvido pela Valve Corporation. Lançado inicialmente em 1999 como uma modificação (mod) para o jogo Half-Life, rapidamente se tornou um sucesso, destacando-se pelo foco em táticas de equipe e precisão de tiro. Em Counter-Strike, equipes de terroristas e contra-terroristas competem em diversos modos de jogo, incluindo desativação de bombas e resgate de reféns. Agora, em 2024, estamos na versão mais recente, Counter-Strike 2.

Neste trabalho, analisamos se os 20 melhores jogadores de Counter-Strike 2, eleitos no final de 2023, se seguem mutuamente na plataforma Twitter. Investigamos se fatores como política, nacionalidade, parceria de time ou eventos do jogo influenciam nas conexões entre eles.

Realizamos uma análise de clusterização de dados em redes utilizando o modelo de blocos estocásticos para estudar as relações entre os jogadores. Adicionalmente, foi ajustado um modelo de clusterização de Louvain para comparação com o modelo de blocos estocásticos.

Os dados foram coletados manualmente no dia 15 de junho de 2024 a partir dos perfis do Twitter de cada jogador. Os jogadores analisados, juntamente com suas nacionalidades, estão relacionados na Tabela 1.

O banco de dados e os scripts em R utilizados neste trabalho estão disponíveis no GitHub: <https://github.com/LuizNunes2020/Atividade3.ME921>.

| Rank | Jogador | País |
|------|--------------------------------------|----------------------|
| 1 | Mathieu " ZywOo " Herbaut | França |
| 2 | Nikola " NiKo " Kovač | Bósnia e Herzegovina |
| 3 | Robin " ropz " Kool | Estônia |
| 4 | Ilya " m0NESY " Osipov | Rússia |
| 5 | Lotan " Spinx " Giladi | Israel |
| 6 | Álvaro " SunPayus " García | Espanha |
| 7 | Oleksandr " s1mple " Kostylev | Ucrânia |
| 8 | Dmitry " sh1ro " Sokolov | Rússia |
| 9 | Martin " stavn " Lund | Dinamarca |
| 10 | Helvijs " broky " Saukants | Letônia |
| 11 | Nicolai " device " Reedtz | Dinamarca |
| 12 | David " frozen " Čerňanský | Eslováquia |
| 13 | Nemanja " huNter " Kovač | Bósnia e Herzegovina |
| 14 | Guy " NertZ " Iluz | Israel |
| 15 | Jakob " jabbi " Nygaard | Dinamarca |
| 16 | Benjamin " blameF " Bremer | Dinamarca |
| 17 | Emil " Magisk " Reif | Dinamarca |
| 18 | Casper " cadiaN " Møller | Dinamarca |
| 19 | Kaike " KSCERATO " Cerato | Brasil |
| 20 | Russel " Twistzz " Van Dulken | Canadá |

Tabela 1: Os 20 melhores jogadores de CS-2

2 Métodos

Com os dados em mãos, realizou-se a clusterização em redes, onde uma rede é composta por nós (ou atores) e as conexões entre eles. Cada nó representa um jogador, e as conexões indicam se um jogador segue outro no Twitter. A rede é representada por uma matriz de adjacência 20x20, refletindo os relacionamentos entre os 20 jogadores analisados, Y , onde y_{ij} indica o relacionamento entre o nó i e o nó j . Os valores de y_{ij} podem ser binários, de contagem ou contínuos. Aqui, usamos dados binários: $y_{ij} = 1$ se o jogador i segue o jogador j , e $y_{ij} = 0$ se o jogador i não segue o jogador j .

Em seguida, foi realizada uma análise exploratória utilizando um mapa de calor da matriz de adjacência e um gráfico de redes, facilitando a visualização dos relacionamentos. Além disso, foram gerados dois dendogramas, um para as linhas e outro para as colunas da matriz.

Utilizando uma técnica de agrupamento baseada no modelo de blocos estocásticos, foi atribuída uma probabilidade a cada observação de pertencer a um dos clusters, sendo então designado ao cluster com a maior probabilidade. O número ideal de clusters foi determinado utilizando o critério da verossimilhança completamente integrada (ICL).

Posteriormente, analisou-se o gráfico de rede do modelo SBM com o número ideal de clusters. Por fim, foram examinados os gráficos de rede gerados pelo modelo de Louvain e realizada uma comparação entre os dois modelos.

2.1 Modelo de Blocos Estocásticos (SBM)

Consideramos que existem G blocos, onde G , representa o número de clusters dos jogadores. Supõe-se que a probabilidade de um jogador pertencer ao bloco g é dada por τ_g . Define-se $z_i = (z_{i1}, z_{i2}, \dots, z_{iG})^T$ como um vetor coluna indicador de pertencimento ao bloco para o jogador i , onde $z_{ig} = 1$ se o jogador i pertence ao bloco g e $z_{ig} = 0$ caso contrário. A matriz de interação bloco-a-bloco $G \times G$, Θ , onde θ_{gh} é a probabilidade de que um jogador no bloco g siga um jogador no bloco h , é utilizada para modelar as interações entre os blocos. Supõe-se ainda que todas as interações par-a-par são independentes.

Dessa forma,

$$Py_{ij} = 1 | \Theta, z_{ig} = 1, z_{jh} = 1 = \theta_{gh} \quad (1)$$

$$= z_i^T \Theta z_j, \quad (2)$$

e

$$Py_{ij} = 0 | \Theta, z_{ig} = 1, z_{jh} = 1 = 1 - \theta_{gh} = 1 - z_i^T \Theta z_j. \quad (3)$$

A verossimilhança dos dados completos para o modelo de blocos estocásticos pode ser expressa como:

$$L_c(\tau, \Theta) = PY, Z | \tau, \Theta = PY | Z, \Theta PZ | \tau = \prod_{i,j} (z_i^T \Theta z_j)^{y_{ij}} (1 - z_i^T \Theta z_j)^{1-y_{ij}} \times \prod_{i=1}^n \prod_{g=1}^G \tau_g^{z_{ig}}, \quad (4)$$

onde $Z = (z_1, z_2, \dots, z_n)$ representa as pertencas aos blocos e o produto é tomado sobre todos os pares $(i, j) : i \neq j$ se a rede for direcionada e sobre os pares $(i, j) : i < j$ se a rede for não direcionada.

Além disso, implementamos um modelo Bayesiano de blocos estocásticos, com inferência realizada por meio de Markov Chain Monte Carlo (MCMC). Inicialmente, define-se o modelo utilizando `nimbleCode`, onde a probabilidade de conexão entre dois nós é modelada por uma distribuição Bernoulli, cujas probabilidades são transformadas pela função logit com parâmetros θ específicos para cada par de clusters. As pertencas aos clusters, Z , são modeladas por uma distribuição categórica com priors Dirichlet para as proporções dos clusters τ , e os parâmetros de ligação θ entre clusters seguem uma distribuição beta. Configuram-se os dados e parâmetros iniciais, incluindo a matriz de adjacência dos dados e uma amostra inicial de pertencas aos clusters. O modelo NIMBLE é então criado com `nimbleModel`, seguido pela configuração do amostrador MCMC usando `configureMCMC`. Compilam-se o modelo e o MCMC com `compileNimble`. Finalmente, o MCMC é executado com `runMCMC` para gerar amostras da distribuição posterior dos parâmetros. Selecionamos alguns parâmetros específicos e plotamos as distribuições marginais das amostras da distribuição a posteriori.

2.2 Modelo de Louvain

O algoritmo Louvain é uma técnica heurística de clusterização que busca maximizar a medida de modularidade Q . Essa métrica avalia a diferença entre a densidade de conexões dentro dos clusters e a densidade esperada caso as conexões fossem aleatórias, respeitando a conectividade de cada vértice. O algoritmo é composto por duas etapas principais e pode ser utilizado em redes com n nós.

Na etapa inicial, cada ponto de interseção é inicialmente tratado como uma comunidade separada. Para cada ponto de interseção i , são verificados os benefícios de modularidade que ocorreriam se i fosse retirado de sua comunidade atual e inserido na comunidade de cada um de seus vizinhos j . Depois de avaliar todos os vizinhos, o ponto de interseção i é transferido para a comunidade que oferece o maior benefício de modularidade, desde que esse benefício seja positivo. Caso contrário, o ponto de interseção i permanece em sua comunidade original. Esse processo continua de forma iterativa até que nenhum ponto de interseção possa aumentar a modularidade ao mudar de comunidade.

Na etapa seguinte, é feita a construção de uma rede renovada na qual os nós representam as comunidades identificadas na fase anterior. Os valores das conexões entre os novos nós são estabelecidos pela adição dos valores das arestas entre os nós das comunidades correspondentes na rede original. As conexões entre nós dentro da mesma comunidade são transformadas em laços na rede atualizada.

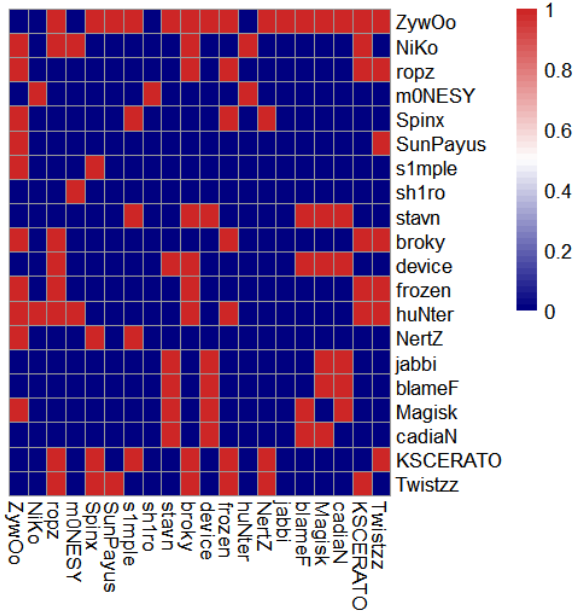
Depois da finalização do segundo estágio, a etapa inicial é novamente utilizada na rede atualizada. Cada repetição das duas fases é chamada de passo. Esses passos são executados várias vezes até que não haja mais alterações e a máxima modularidade seja atingida.

3 Resultados

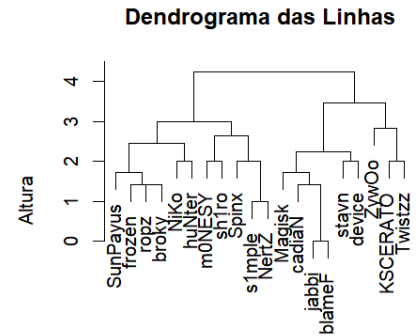
A Figura 1a representa o gráfico de calor no qual cada quadrado pintado na linha i e coluna j indica que a pessoa da linha i segue a pessoa da coluna j . Por convenção, a diagonal (pessoa i com i) não é pintada. Podemos observar que ninguém segue, nem é seguido por, todas as pessoas.

As Figuras 1b e 1c mostram os dendogramas. Na Figura 1b, jogadores que estão mais próximos uns dos outros são mais similares em termos de suas conexões. Na Figura 1c, jogadores que estão próximos uns dos outros são seguidos por grupos semelhantes de jogadores.

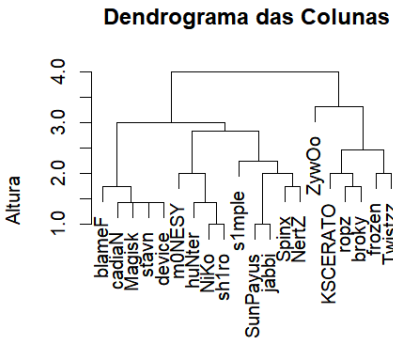
A Figura 1d ilustra as relações entre os jogadores, onde as setas que apontam de um nó para outro indicam uma relação de seguidor. Observa-se a formação de, pelo menos, três clusters distintos: um à esquerda, um central e outro à direita.



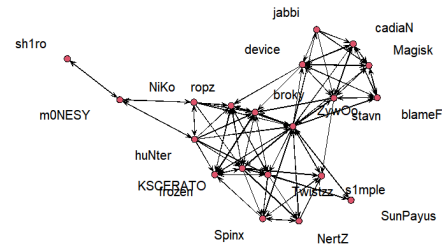
(a) Heatmap da matriz de adjacência dos dados.



(b) Dendrograma das linhas



(c) Dendrograma das colunas



(d) Rede de relações entre os jogadores

Figura 1: Visualizações da rede de relacionamento dos jogadores.

Para determinar o número ideal de clusters, comparamos os valores do critério ICL para modelos ajustados com diferentes números de clusters. A Tabela 2 apresenta esses valores. Observamos que o modelo com 5 clusters possui o maior valor de ICL (-282.3721), indicando que este é o número ideal.

| Número de Clusters | Critério ICL |
|--------------------|--------------|
| 1 | -445.0741 |
| 2 | -376.3608 |
| 3 | -334.6792 |
| 4 | -314.1105 |
| 5 | -282.3721 |
| 6 | -294.6102 |
| 7 | -321.7358 |
| 8 | -348.5904 |
| 9 | -396.7208 |
| 10 | -455.4032 |

Tabela 2

A seguir, seguem os resultados detalhados do modelo selecionado com 5 clusters:

- **Alphas:**

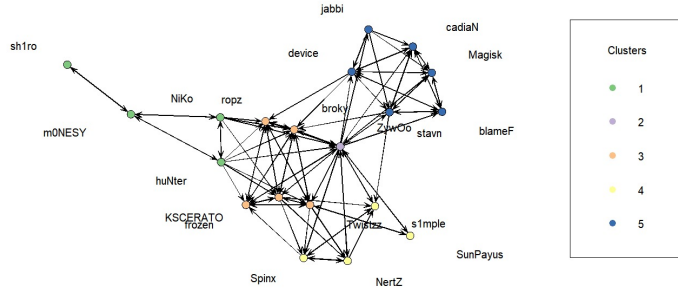
$$(0.20 \quad 0.05 \quad 0.25 \quad 0.20 \quad 0.30)$$

- **Matriz Θ :**

$$\begin{pmatrix} 0.6666667 & 0.0000010 & 0.000001 & 0.0000010 & 0.00000100 \\ 0.5000000 & 0.6578947 & 0.600000 & 0.9999990 & 0.16666667 \\ 0.4000000 & 0.9999990 & 0.999999 & 0.1000000 & 0.10000000 \\ 0.0000010 & 0.9999990 & 0.300000 & 0.4166667 & 0.04166667 \\ 0.0000010 & 0.9999990 & 0.000001 & 0.0000010 & 0.80000000 \end{pmatrix}$$

Os valores dos alphas representam as proporções dos diferentes clusters no modelo. Os alphas indicam que o quinto cluster é o mais populoso, enquanto o segundo é o menor.

A matriz Θ representa as probabilidades de ligação entre os diferentes clusters. Cada entrada π_{ij} na matriz indica a probabilidade de uma ligação entre um nó no cluster i e um nó no cluster j . Por exemplo, a probabilidade de ligação entre os clusters 1 e 2 é 0.0000010.



| Cluster | Jogadores |
|---------|--|
| 1 | NiKo, m0NESY, sh1ro, huNter |
| 2 | ZywOo |
| 3 | ropz, broky, frozen, KSCERATO, Twistzz |
| 4 | Spinx, SunPayus, s1mple, NertZ |
| 5 | stavn, device, jabbi, blameF, Magisk, cadiaN |

Tabela 3: Jogadores de cada cluster

Figura 2: Rede de clusters

A Figura 2, que é uma versão aprimorada da Figura 1d com a classificação dos jogadores sobreposta, revela que os três grandes grupos identificados anteriormente se subdividiram em cinco clusters. A Tabela 3d detalha os jogadores de cada cluster.

As Figuras 3a, 3b, 3c e a Tabela 3d apresentam os resultados da aplicação dos métodos de clusterização baseados no modelo de bloco estocástico e no algoritmo de Louvain.

A Figura 3a ilustra a distribuição a posterior dos parâmetros θ e τ para as ligações dentro dos clusters 1, 2 e 3, respectivamente.

As Figuras 3b e 3c mostram o gráfico de rede de clusters identificado pelo modelo de Louvain, que definiu três clusters ideais, e a tabela 3d que lista os jogadores de cada cluster, respectivamente.

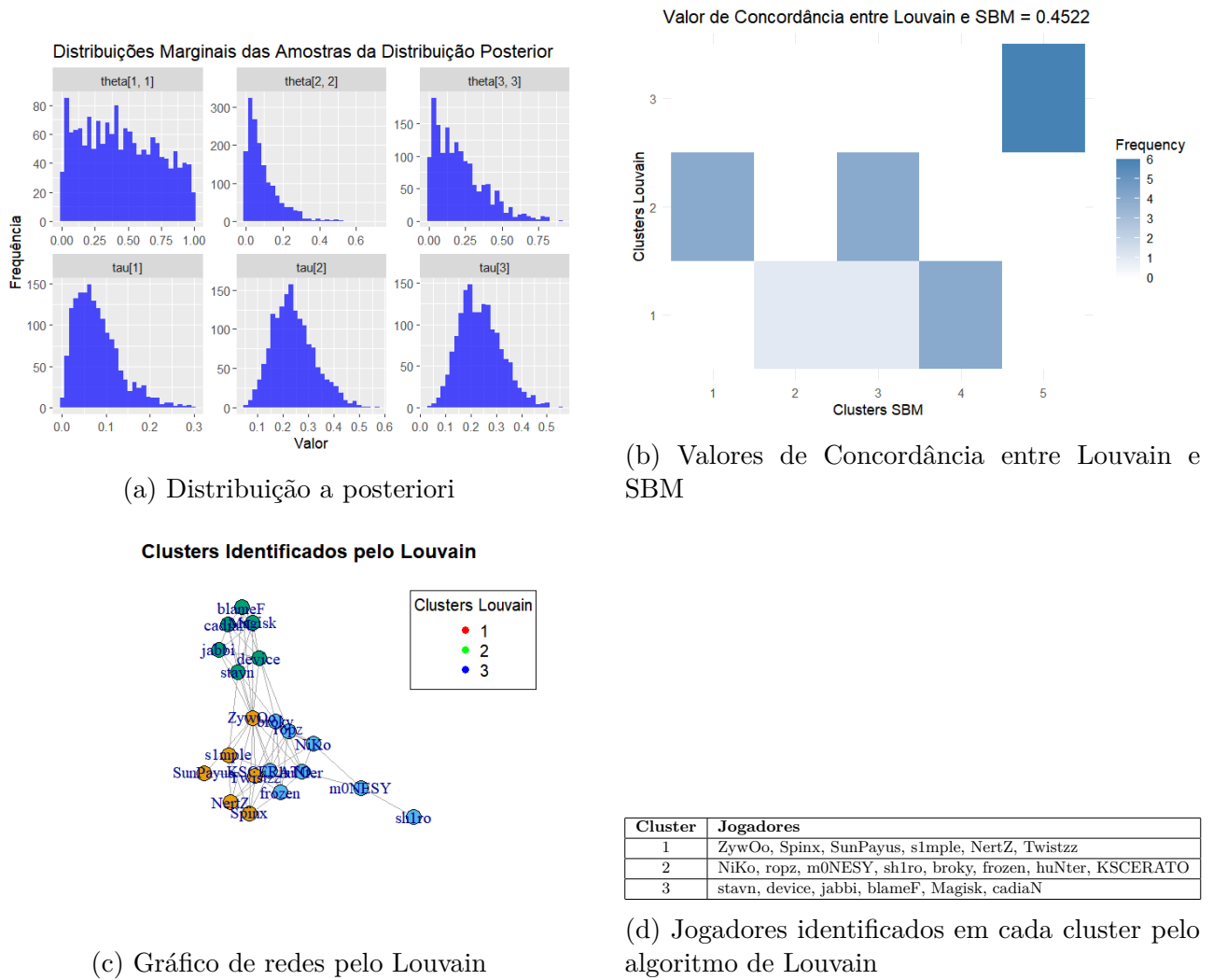


Figura 3: Análises dos Modelos de Clusterização

A Figura 3b apresenta um heatmap que compara a correspondência entre os clusters identificados pelo modelo de blocos estocásticos (SBM) e pelo algoritmo Louvain. Cada célula do heatmap mostra a frequência de jogadores pertencentes aos clusters correspondentes de ambos os métodos. A intensidade da cor nas células indica a frequência de jogadores que são categorizados no mesmo cluster.

O valor de concordância entre os clusters Louvain e SBM é 0.4522. Uma concordância perfeita seria 1, mas esse valor indica que, embora haja alguma correspondência entre os clusters identificados pelos dois métodos, também há diferenças significativas.

4 Conclusão

Com base nos resultados, as Figuras 1a, 1b e 1c mostram que "Zyw0o" é seguido pela maioria dos jogadores, provavelmente por ter sido considerado o melhor jogador de 2023. Ele, também, se destaca nos dendogramas, aparecendo isolado.

Os resultados do modelo SBM indicaram que os clusters mais populosos são o cluster 5 e o cluster 3, com 6 e 5 jogadores, respectivamente. Esses clusters não apresentam um padrão claro, com jogadores de diferentes nacionalidades (França, Estônia, Brasil, Espanha, etc.) e algumas conexões anteriores em times, mas sem nenhuma tendência clara.

Analisando os clusters 1, 2 e 4, podemos observar o seguinte:

- **Cluster 1:** Contém dois jogadores da Rússia ("m0NESY" e "sh1ro") e dois da Bósnia e Herzegovina ("huNter" e "NiKo"). Um fato interessante é que "m0NESY", "huNter" e "NiKo" jogam no mesmo time (G2 Sports) e se seguem mutuamente. O "sh1ro" segue apenas o "m0NESY", possivelmente por ser também russo.
- **Cluster 2:** Este cluster contém apenas "Zyw0o", como já era esperado pelos dendogramas.

- **Cluster 4:** Inclui quatro jogadores: "s1mple", "NertZ", "Spinx" e "SunPayus". O "s1mple" é ucraniano, "NertZ" e "Spinx" são israelenses e o "SunPayus" é espanhol. Nenhum deles jogou junto anteriormente. A Figura 1a revela que eles não seguem muitos jogadores, especialmente os russos.

Por fim, pelo modelo de Louvain, obtivemos 3 clusters. Podemos observar pela Tabela 3d e pelas Figuras 3b e 3c que os Clusters Louvain 1 e 2 têm sobreposições diferentes com os clusters SBM. No entanto, o Cluster 5 do SBM parece ter uma correspondência relativamente alta com o Cluster 3 do Louvain.

Pode-se concluir que a nacionalidade e a experiência de jogar no mesmo time influenciam quem os jogadores seguem no Twitter, especialmente no caso de jogadores russos, ucranianos e israelenses, considerando o momento atual que nos encontramos.

Porém, a análise mostrou que a matriz Θ não é diagonalmente dominante ($\theta_{gg} < \theta_{gh}$ para $h \neq g$). Isso significa que os jogadores têm maior probabilidade de se conectar com jogadores de outros blocos do que com jogadores dentro do mesmo bloco. Embora os blocos não correspondam a comunidades tradicionais, eles ainda são clusters válidos porque os nós dentro de cada bloco são estocasticamente equivalentes.

Referências

- [1] Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R* (Vol. 50). Cambridge University Press.
- [2] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2022). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.4 — For new features, see the 'Changelog' file (in the package source).
- [3] Handcock, M. S., Raftery, A. E., & Tantrum, J. (2007). *Model-based clustering for social networks*. Journal of the Royal Statistical Society: Series A (Statistics in Society), 170(2), 301-354.
- [4] Aires, V., & Nakamura, F. (2020). *Aplicação de Medidas de Centralidade ao Método Louvain para Detecção de Comunidades em Redes Sociais*. Revista de Sistemas e Computação, 10(2), 45-60.