
**Programa de Iniciação Científica e de Iniciação em
Desenvolvimento Tecnológico e Inovação da USP**

Projeto de Pesquisa

**Predição de variáveis dinâmicas no modelo
de Sznajd em Redes Complexas**

Aluno: Vítor Amorim Fróis

Orientador: Francisco Aparecido Rodrigues

Instituto de Ciências Matemáticas e Computação (ICMC)

Universidade de São Paulo (USP)

1 Introdução

A interação entre componentes de um sistema complexo resulta em propriedades não presentes em partes isoladas, conhecidas como propriedades emergentes. Um exemplo bem conhecido de emergência pode ser visto em colônias de formigas, onde animais exibem comportamento complexo a partir de regras simples. Este fenômeno não se limita à fauna e se manifesta na sociedade através da interação entre pessoas, grupos e instituições [10].

Uma consequência substancial da emergência é a polarização social, na qual agentes de um sistema rapidamente afunilam suas opiniões, diminuindo a tolerância para opiniões distintas e causando uma divisão na sociedade. Diversos estudos [8, 11] associam a polarização com os eventos violentos que sucederam as últimas eleições brasileiras. Assim, as causas e efeitos da polarização em redes sociais devem ser compreendidas a fim de desenvolver estratégias efetivas de comunicação e intervenções sociais para mitigar o seu impacto.

Dessa forma, no contexto da compreensão da dinâmica do consenso, modelos matemáticos têm sido adaptados para investigar não apenas a polarização, mas também o processo de consenso em redes complexas [5]. Esses modelos levam em consideração a relação entre agentes e a estrutura subjacente da rede, como o modelo de Sznajd [15], que explora como opiniões similares podem influenciar outros, resultando na formação de *clusters* de opiniões. No entanto, para uma compreensão mais realista do fenômeno do consenso, é crucial simular esses modelos em diferentes topologias de rede, uma vez que a topologia da rede desempenha um papel fundamental na dinâmica do consenso e na polarização resultante. Estudos recentes destacam a influência significativa da topologia da rede nos resultados de consenso e polarização [12].

Dada a significativa influência da topologia da rede na formação de consenso, surge a necessidade de explorar a viabilidade de um modelo de Aprendizado de Máquina para prever variáveis dinâmicas de sistemas com base nas propriedades de rede subjacente. Essa abordagem, amplamente aplicada em campos como sincronização e disseminação de epidemias [13], levanta a questão sobre sua aplicabilidade no estudo do modelo de Sznajd. Este projeto visa investigar essa possibilidade, focalizando na capacidade do aprendizado de máquina de antecipar variáveis dinâmicas do modelo de Sznajd, com base na topologia da rede. Destaca-se assim,

o potencial dessas análises de rede para a compreensão de sistemas dinâmicos, fornecendo *insights* valiosos sobre a emergência e evolução da polarização na sociedade.

Portanto, pretendemos prever o tempo para consenso e a frequência de troca de opinião no modelo Sznajd se baseando em propriedades da rede complexa. Esse trabalho é uma continuação das pesquisas apresentada em [12], onde os autores predizeram essas propriedades para o modelo *Q-Voter*. Assim sendo, objetivamos verificar se a mesma acurácia na predição pode ser obtida para o modelo de Sznajd, que é uma generalização do *Q-Voter*.

2 Objetivos

Os principais objetivos do projeto envolvem:

- Desenvolver um estudo com modelos de Aprendizado de Máquina Supervisionado para prever (i) o tempo para consenso e (ii) a frequência de troca de opinião no modelo Sznajd baseado na topologia da rede;
- Determinar as propriedades da rede que mais influenciam na emergência do consenso.

Após a validação dos resultados, diversas análises listadas abaixo como objetivos secundários serão efetuadas, enriquecendo o trabalho:

- Determinar quais *features*, ou seja, as medidas topológicas da rede, são essenciais para realizar predições i e ii;
- Analisar como diversos métodos de inicialização de opinião positiva nos nós influenciam a importância das *features*;
- Comparar os resultados obtidos para os modelos de opinião (Sznajd e *Q-Voter*).

3 Materiais e métodos

3.1 Ferramentas em Python

Todo o código referente ao projeto será desenvolvido na linguagem de programação *Python*, com auxílio das bibliotecas *igraph*, *NetworkX*, *Scikit-Learn* e *Matplotlib*.

3.2 Modelo de Sznajd

O modelo de *spin* de Ising é um dos modelos mais utilizados na mecânica estatística [5]. No artigo [15] é proposto o modelo de Sznajd, uma adaptação de Ising para descrever dinâmicas de opinião em uma comunidade.

O modelo original segue uma simulação estocástica implementando o fenômeno de validação social nos agentes $S_i, i = 1, 2, \dots, N$ com opiniões $O = \{-1, +1\}$. A cada passo, dois vizinhos são selecionados e o sistema é atualizado de acordo com as seguintes regras dinâmicas:

- Se $S_i S_{i+1} = 1$, então os vizinhos S_{i-1} e S_{i+2} recebem a opinião do par S_i, S_{i+1}
- Se $S_i S_{i+1} = -1$, então $S_{i-1} = S_{i+1}$ e $S_{i+2} = S_i$

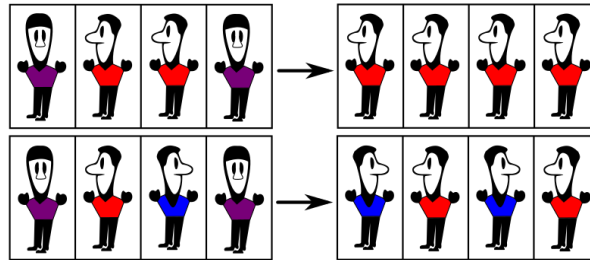


Figura 1: Regras de mudança de opinião no modelo de Sznajd. Azul representa a opinião positiva e vermelho, a negativa. Fonte: [1]

O modelo original foi proposto para um sistema unidimensional. No entanto, a dinâmica foi modificada de forma incluir uma rede complexa [14]. Nesse trabalho será utilizada a adaptação apresentada em [2] para implementação do modelo de Sznajd em redes com duas opiniões. Considere uma rede de N pessoas, com opiniões $O = \{-1, +1\}$ inicialmente distribuídas de

forma aleatória. Cada indivíduo é uma variável dinâmica binária $s(x, t) = O$ de grau k_x , em que $x = 1, \dots, N$. Uma iteração t de uma sequência de iterações até o consenso é descrita abaixo:

- Uma dupla de nós vizinhos i e j é escolhida aleatoriamente
- Se $s(i, t) \neq s(j, t)$ a iteração termina
- Se $s(i, t) = s(j, t)$, cada vizinho de i pode mudar a sua opinião para $s(i, t)$ com probabilidade $1/k_i$. De forma semelhante, cada vizinho de j pode mudar sua opinião para $s(j, t)$ com probabilidade $1/k_j$.

O tempo de consenso, definido como o período necessário para que o sistema alcance um estado estacionário, é uma métrica crucial na análise da dinâmica de consenso. Durante a simulação, registramos tanto o tempo de consenso quanto a frequência de troca de opinião como indicadores-chave do comportamento do sistema.

Os parâmetros para as redes e o modelo serão fixados para proporcionar um patamar conciso durante os testes com os algoritmos de aprendizado de máquina. Ao fixar esses parâmetros é possível focar no impacto de outras variáveis na análise. Dessa forma, as simulações contarão com as redes com um número de nós fixo, a saber, $N = 1000$, além de uma porcentagem de nós com opiniões positivas $p = 0,2$.

Além disso, adotaremos três abordagens distintas de inicialização para os nós com opiniões positivas nas simulações. Primeiramente, implementaremos a inicialização aleatória, atribuindo aleatoriamente opiniões positivas aos nós. Em seguida, adotaremos a estratégia de inicialização de baixo grau, na qual os nós com menor grau receberão opiniões positivas. Por fim, aplicaremos a inicialização de alto grau, na qual os nós mais influentes na rede receberão opiniões positivas. É de suma importância simular o sistema com diferentes inicializações, possibilitando analisar como a importância das *features* são influenciadas em cada caso e compreender melhor como situações de consenso podem ser favorecidas.

3.3 Redes

Oito diferentes topologias das redes serão examinadas. As redes Erdős–Rényi, Barabási–Albert linear, Barabási–Albert não linear com $\alpha = 0.5$ and $\alpha = 1.5$, Lancichinetti-Fortunato-Radicchi (LFR), Watts–Strogatz, Waxman e grafo de caminho [3, 6]. Essas topologias buscam abordar diferentes estruturas que sociedades reais possam admitir, considerando a presença de hubs, comunidades e *small-world*. Ou seja, como as redes geradas por esses modelos apresentam diferentes propriedades que podem ser controladas através de seus parâmetros, poderemos gerar um banco de dados com muitos exemplos de topologias diferentes. Assim, poderemos verificar os efeitos de propriedades topológicas no processo dinâmico, visto que muitas propriedades, como distância entre os vértices ou nível de centralidade, sofrerão variações nas bases geradas. Essa variação é importante para oferecermos exemplos diferentes aos modelos de aprendizado que usaremos na fase de predição das variáveis dinâmicas. Para cada uma dessas redes, 100 instâncias serão criadas visando diminuir efeitos da aleatoriedade na construção do modelo.

3.4 Métricas de rede

Nesse projeto as hipóteses do tempo para alcançar consenso Y_i e a frequência de mudança de opinião C_i serão inferidos a partir de um vetor de *features* derivado da estrutura da rede e denotado por $X_i = \{X_{i1}, X_{i2}, \dots, X_{ik}\}$, em que X_{ik} é a k -ésima métrica da rede i . A explicação foca na predição de Y_i mas também é válida para C_i .

$$Y_i = f(X_i) + \delta$$

O objetivo é encontrar uma função f que relaciona Y_i às métricas da rede. Trataremos predição de Y_i como um problema de regressão em que δ é um termo que representa uma distribuição normal com média zero e desvio padrão σ . Esse termo representa a incerteza nos dados, que incluem as medidas que não foram incluídas no modelo e as flutuações aleatórias na simulação das redes e modelos.

Para caracterizar as redes, serão utilizadas diversas medidas, incluindo o coeficiente de *clus-*

tering, *closeness centrality*, *betweenness centrality*, *average shortest path length*, coeficiente de correlação de Pearson do grau, *information centrality*, *approximate current flow betweenness centrality* e *eigenvector centrality*, dentre outras. Tais medidas, usadas coletivamente aqui, fornecem *insights* valiosos sobre a topologia, conectividade, eficiência, influência e organização em redes complexas [6].

3.5 Aprendizado de Máquina

Para a predição das variáveis dinâmicas será utilizado o algoritmo *Random Forest* (RF) [4]. O RF se baseia na criação de um conjunto de árvores de decisão com aleatorização. São amplamente utilizadas na predição de dados tabulares com alta acurácia. Apesar de sua simplicidade, oferece alta robustez, não é sensível a outliers e possibilita a análise das importância das *features* utilizadas no modelo.

A técnica *grid search* [7] será utilizada para aumentar a acurácia do modelo ao otimizar seus hiperparâmetros. De forma conjunta, a fim de garantir a acurácia e prevenir *overfitting*, a validação cruzada será utilizada. Por fim, um conjunto adicional de dados teste será providenciado com as técnicas descritas para assegurar a qualidade do modelo.

4 Métricas de Avaliação

O coeficiente de determinação, R^2 , é uma métrica usada para medir o quão bem um modelo de regressão se ajusta aos dados [9]. A fim de analisar os resultados utilizaremos o R^2 no modelo de aprendizado de máquina, juntamente com técnicas descritas acima, como a validação cruzada e etapa de teste em um conjunto oculto de dados. Essa etapa busca garantir que o modelo foi capaz de generalizar com base nos dados de treinamento e consegue realizar boas previsões em dados novos.

5 Cronograma

O projeto será desenvolvido considerando as atividades e o cronograma descritos na Tabela 1.

Tabela 1: Cronograma das atividades a serem desenvolvidas durante o projeto.

Etapa	Mês											
	1	2	3	4	5	6	7	8	9	10	11	12
1												
2												
3												
4												
5												
6												
7												
8												

1. Estudo do papel da topologia da rede na dinâmica de opiniões e aprofundamento dos modelos Sznajd e Q-Votação através do artigo base [5];
2. Geração das diversas topologias de rede, cálculo das métricas e implementação do modelo Sznajd utilizando as três diferentes formas de inicialização;
3. Criação do conjunto de dados contendo informações sobre todas redes geradas: medidas de topologia, método de inicialização, tempo de consenso e frequência de mudança de opinião;
4. Aplicação e avaliação do algoritmo de aprendizado de máquina para predição das variáveis resposta;
5. Análise de importância das *features* utilizando as ferramentas da biblioteca *Scikit-Learn* e determinação de características importantes para predições;
6. Extensão do item 5 para análise de como os diferentes métodos de inicialização influenciam a importância das *features*.
7. Comparação dos resultados obtidos com o artigo [12] a fim de entender como diferentes dinâmicas de opinião (*Q-Voter* e Sznajd) influenciam o tempo de consenso, a frequência de mudança de opinião e a importância de *features*;
8. Escrita de artigo científico reunindo e apresentando os resultados obtidos.

Referências

- [1] Sznajd model. https://en.wikipedia.org/wiki/Sznajd_model.
- [2] A.T. Bernardes, D. Stauffer, and J. Kertész. Election results and the sznajd model on barabasi network. *The European Physical Journal B - Condensed Matter*, 25(1):123–127, January 2002. ISSN 1434-6036. doi: 10.1140/e10051-002-0013-y. URL <http://dx.doi.org/10.1140/e10051-002-0013-y>.
- [3] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.
- [4] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
- [5] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, May 2009. ISSN 1539-0756. doi: 10.1103/revmodphys.81.591. URL <http://dx.doi.org/10.1103/RevModPhys.81.591>.
- [6] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, January 2007. ISSN 1460-6976. doi: 10.1080/00018730601170527. URL <http://dx.doi.org/10.1080/00018730601170527>.
- [7] Gavin Hackeling. *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd, 2017.
- [8] Ruben Interian and Francisco A Rodrigues. Group polarization, influence, and domination in online interaction networks: a case study of the 2022 brazilian elections. *Journal of Physics: Complexity*, 4(3):035008, September 2023. ISSN 2632-072X. doi: 10.1088/2632-072x/acf6a4. URL <http://dx.doi.org/10.1088/2632-072X/acf6a4>.
- [9] Paul Johnson and Holger Schielzeth. The coefficient of determination r^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expan-

- ded. *Journal of The Royal Society Interface*, 14:20170213, 09 2017. doi: 10.1098/rsif.2017.0213.
- [10] Steven Johnson. *Emergence: The connected lives of ants, brains, cities, and software*. Simon and Schuster, 2002.
- [11] Matthew L. Layton, Amy Erica Smith, Mason W. Moseley, and Mollie J. Cohen. Demographic polarization and the rise of the far right: Brazil’s 2018 presidential election. *Research & Politics*, 8(1):2053168021990204, 2021. doi: 10.1177/2053168021990204. URL <https://doi.org/10.1177/2053168021990204>.
- [12] Aruane M. Pineda, Paul Kent, Colm Connaughton, and Francisco A. Rodrigues. Machine learning-based prediction of q-voter model in complex networks, 2023.
- [13] Francisco A. Rodrigues, Thomas Peron, Colm Connaughton, Jurgen Kurths, and Yamir Moreno. A machine learning approach to predicting dynamical observables from network structure, 2019.
- [14] Juan R. Sanchez. A modified one-dimensional sznajd model, 2004.
- [15] KATARZYNA SZNAJD-WERON and JÓZEF SZNAJD. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11(06):1157–1165, September 2000. ISSN 1793-6586. doi: 10.1142/s0129183100000936. URL <http://dx.doi.org/10.1142/S0129183100000936>.