Predição de Variáveis Dinâmicas no Modelo de Sznajd em Redes Complexas

Vítor Amorim Fróis

Resumo

O presente trabalho utiliza Aprendizado de Máquina para prever variáveis complexas no modelo de Sznajd em Redes Complexas: o Tempo de Consenso e a Frequência de Troca de Opinião. Ao utilizar medidas topológicas para caracterização de redes e consequentemente como features, podemos prever as variáveis com alta acurácia. Ao explorar a convergência entre estrutura e dinâmica de redes, esse projeto responde dúvidas relacionadas aos mecanismos de polarização em interações sociais e abre caminho para novos questionamentos. O código do projeto pode ser encontrado no repositório github.com/vitorfrois/SznajdNetworks.

Conteúdo

1 Introdução

A interação entre componentes de um sistema que possuem regras simples leva a formação de padrões complexos e características como emergência, livre de escala e heterogeneidade. Fenômenos emergentes são presentes em sistemas complexos e caracterizados pelo resultado espontâneo da interação entre os milhares de componentes que constituem o sistema. Um grande exemplo de emergência ocorre durante a noite do sudeste asiático, quando vagalumes da região piscam de acordo ajustam a frequência do piscar de suas luzes de acordo com os vizinhos mais próximos, até que o efeito seja extendido por todo o sistema, de forma que os indivíduos pisquem em sincronia (S. Johnson 2002).

No contexto de dinâmicas sociais, isto é, modelos matemáticos que buscam reproduzir o comportamento humano em redes, a emergência pode ser caracterizada como um fenômeno relacionado a polarização (Maia, Ferreira, e Martins 2021). Aqui e no restante do relatório, definimos polarização como a fragmentação de opiniões, um estado contrário ao consenso. Diversos estudos mostram que a polarização pode ter profunda influência no âmbito político, como visto nas manifestações anti-democráticas e violentas ocorridas em Brasília no dia 8 de Janeiro de 2023 (Interian e Rodrigues 2023; Layton et al. 2021). Dessa forma, é de suma importância estudar a polarização para evitar que cenários de discórdia se repitam.

A física estatística desenvolveu ferramentas para o estudo de sistemas de muitas partículas interagentes, os quais são adaptados com facilidade para o estudo de dinâmicas sociais. Ersnt Ising encontrou a solução exata para um modelo de paramagneto, representando materiais que podem alcançar dois estados conflitantes e buscam um estado de mínima energia. O modelo recebeu o nome de Ising e pode ser considerado como um modelo para simples opiniões, onde há uma transição de fase entre os estados de polarização e consenso. O modelo de Sznajd foi inspirado pelo primeiro modelo e busca explorar como opiniões semelhantes são necessárias para influenciar outros. Já o modelo votante ilustra como a maioria pode influenciar vizinhos, explorando por sua vez como a ordem emerge a partir da opinião maioria.

Para uma compreensão mais realista do fenômeno do consenso, é crucial simular esses modelos em diferentes topologias de rede, uma vez que ela 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 (Pineda et al. 2023). 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 (Rodrigues et al. 2019), levanta a questão sobre sua aplicabilidade no estudo do modelo de Sznajd. Este trabalho investiga 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.

Esse trabalho apresenta valiosos *insights* na relação entre topologia de rede e dinâmicas sociais, destacando o potencial do uso de métricas de rede para análise de sistemas dinâmicos. Visto a alta colinearidade nas métricas de caracterização (ver figura 4) e comportamento das variáveis resposta, uma metodologia baseada em *Forward Selection* e Regressão não Linear foi proposta, garantindo alta acurácia, robustez e maior explicabilidade em relação a improtância de *features* das *Random Forests*.

2 Materiais e Métodos

2.1 Geração de Redes Aleatórias

Seis diferentes topologias das redes foram examinadas. As redes Erdös–Rényi, Barabási–Albert linear, Barabási–Albert não linear com $\alpha = 0.5$ and $\alpha = 1.5$, Watts–Strogatz e Waxman (Boccaletti et al. 2006; Costa et al. 2007). 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 podems ser controladas através de seus parâmetros, poderemos gerar um banco de dados com exemplos de topologias diferentes. Assim, os efeitos de propriedades topológicas no processo dinâmico podem ser verificados, 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 foram criadas visando diminuir efeitos da aleatoriedade na construção do modelo.

2.1.1 Erdos-Renyi (ER)

O modelo de Erdos-Renyi (ER) é um dos mais estudados e detalhados na teoria dos grafos. É formado ao ligar N nós entre as possíveis arestas com probabilidade p. Apesar de não representar com fidelidade cenários do mundo real, possui apelo matemático por possuir características bem definidas.

2.1.2 Small-World de Watts e Strogatz

Diversas redes do mundo real exibem a propriedade *small-world*, isto é, a maioria dos vértices podem ser alcançados pelo restante a partir de um pequeno número de arestas. Essa propriedade é muito comum em redes sociais.

Outra propriedade muito relevante em redes é a presença de *loops* de tamanho três: se i está conectado a j e k, há uma grande probabilidade que j e k estejam conectados por sua vez. As redes ER possuem característica de pequeno mundo, porém não apresentam muitos triângulos. De forma contrária, é fácil construir redes com abundância de loops, mas é difícil garantir a presença de características de pequeno mundo.

O modelo mais popular que uniu as duas características foi desenvolvido por Watts e Strogatz e recebeu o nome de modelo *small-world* de Watts-Strogatz (WS). Para construí-lo, comece com uma grade triangular e realize a reconexão de cada aresta presente com probabilidade p. Para $p \approx 0$, a rede original é mantida, enquanto que para $p \approx 1$ há uma rede aleatória.

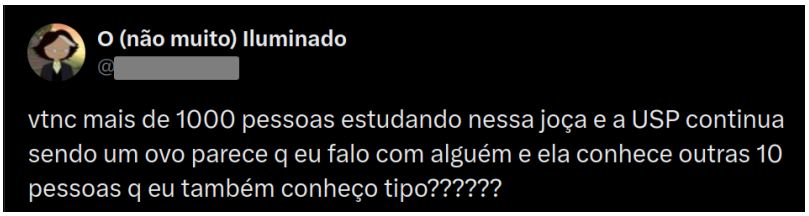

Figura 1: Post na plataforma X (Twitter) discorre sobre a rede de interações da Universidade de São Paulo, a qual possui características estudadas por Watts e Strogatz. Acessado em 03/11/2024 em https://x.com/LuzMadLED/status/1852872862903771258

2.1.3 Redes Livre de Escala de Barabási e Albert

Barabási e Albert demonstraram que a distribuição do grau de inúmeros sistemas do mundo real é caracterizada por uma distribuição assimétrica. Nessas redes, alguns vértices são altamente conectados enquanto outros possuem poucas conexões. Uma característica muito importnate dessa rede é a existência de *hubs*, vértices que são conectados a uma fração significativa do total da rede. A construção das redes Barabási-Albert inicia com

um conjunto de vértices e iterativamente adiciona arestas de forma que os vértices mais conectados possuam maior chance de formar novas arestas. Esse processo leva a formação de uma rede onde a distribuição dos nós segue uma lei de potência.

2.1.4 Redes Geográficas

A maioria das redes complexas mora em um espaço abstrato, onde a posição dos vértices não tem um sentido particular. Em algumas redes, porém, a posição dos vértices pode ter importante impacto, como por exemplo, no caso de redes de transporte rodoviário, aéreo e redes neuronais. Esses exemplos recebem o nome de redes geográficas. Uma maneira simples de gerar redes geográficas é distribuir N vértices em um espaço abstrato e conectá-los com uma probabilidade que decai de acordo com a distância entre eles.

2.2 Simulação de Monte Carlo do modelo de Sznajd

O modelo de *spin* de Ising é um dos modelos mais utilizados na mecânica estatística~(Castellano, Fortunato, e Loreto 2009). No artigo (SZNAJD-WERON e SZNAJD 2000) 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 $S_i, i = 1, 2, ..., N$ 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 $SiS_{i+1} = 1$, então os vizinhos S_{i-1} e S_{i+2} recebem a opinião do par S_i, S_{i+1}
- Se $SiS_{i+1} = -1$, então $S_{i-1} = S_{i+1}$ e $S_{i+2} = S_i$

O modelo original foi proposto para um sistema unidimensional. No entanto, a dinâmica foi modificada de forma incluir uma rede complexa (Sanchez 2004). Nesse trabalho será utilizada a adaptação apresentada em (Bernardes, Stauffer, e Kertész 2002) para implementação do modelo de Sznajd em redes com duas opiniões. Considere uma rede de N pessoas, com opiniões $O = \{-1, +1\}$ inicialmente distribuidas de forma aleatória. Cada indivíduo é uma variável dinâmica binária $s(x, t) = O$ de grau k_x, em que $x = 1, ..., N$. 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 $s(i, t) = s(j, t)$, a união dos vizinhos de i e j recebe a opinião de i.

2.2.1 Variáveis dinâmicas de interesse

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, bem como a **frequência da troca de opinião**. Durante a simulação, registramos tanto o tempo de consenso quanto a frequência de troca de opinião como indicadores-chave do comportamento do sistema. O histograma de ambas variáveis aleatórias são exibidos abaixo com a estimativa de densidade correspondente nas figuras 2 e 3. Se faz necessária a utilização da escala logarítmica para visualização devido ao aspecto de cauda pesada das distribuições.

2.2.2 Inicialização dos nós

Os parâmetros para as redes e o modelo foram 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, adotamos três abordagens distintas de inicialização para os nós com opiniões positivas nas simulações. Primeiramente, a inicialização aleatória, atribuindo aleatoriamente opiniões positivas aos nós. Em seguida, adotamos a estratégia de inicialização inversa, na qual os nós com menor grau receberão opiniões positivas. Por fim, aplicaremos a inicialização direta, na qual os nós mais influentes na rede receberão opiniões positivas. É de suma importância simular o sistema com diferentes inicializações, possibilitando analisar

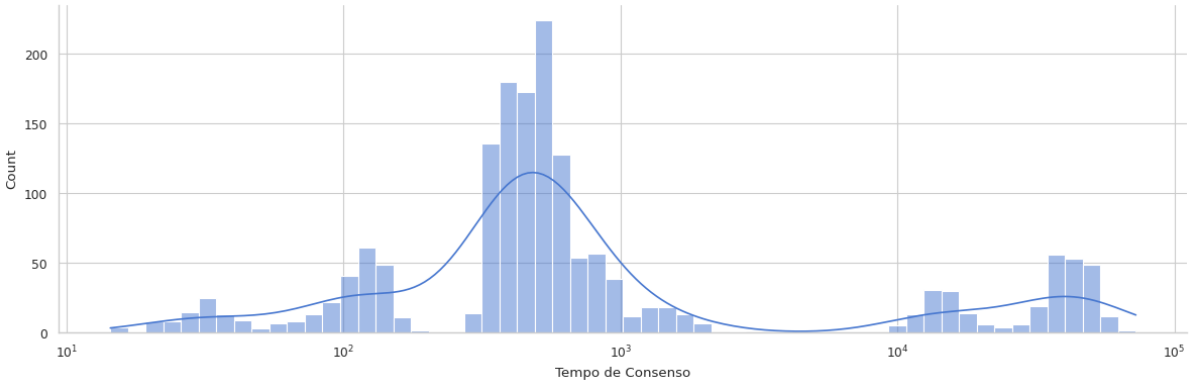

Figura 2: Histograma do Tempo de Consenso na escala logarítmica

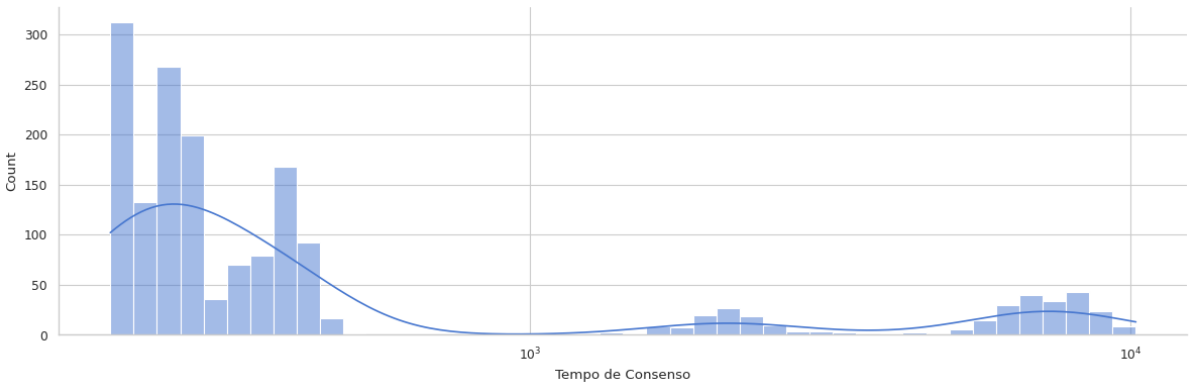

Figura 3: Histograma da Frequência de Troca de Opinião na escala logarítmica

como a importância das *features* são influenciadas em cada caso e compreender melhor como situações de consenso podem ser favorecidas.

2.3 Caracterização de Redes

Buscamos caracterizar cada rede i utilizando um vetor de features derivado de sua estrutura e denotado por $X_i = \{X_{i1}, X_{i2}, ..., X_{ik}\}$, em que X_{ik} é a k-ésima métrica da rede i. Assim, foram utilizadas diversas medidas, incluindo o coeficiente de *clustering*, *closeness centrality*, *betweenness centrality*, *average shortest path lenght*, coeficiente de correlação de Pearson do grau, *information centrality*, *approximate current flow betweenness centrality* e *eigenvector centrality*, Entropia de Shannon e segundo momento do grau. Tais medidas, usadas coletivamente aqui, fornecem *insights* valiosos sobre a topologia, conectividade, eficiência, influência e organização em redes complexas (Costa et al. 2007).

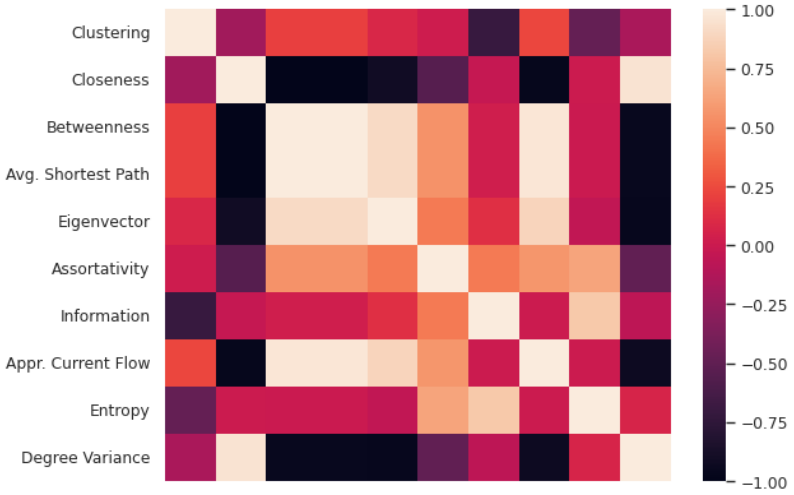

Figura 4: *Heatmap* utilizando correlação de Spearman entre as features. As colunas estão ordenadas como as linhas. A cor vermelha indica coeficiente de Spearman 0 e, portanto, baixa colinearidade. De forma contrária, valores muito claros ou muito escuros indicam colinearidade entre as métricas para as redes geradas. Assim, é possível observar alta colinearidade entre diversas features.

As métricas descritas acima são dividas entre três grandes grupos, sendo eles medidas de centralidade (*closeness centrality*, *betweenness centrality*, *average shortest path lenght*, *information centrality*, *approximate current flow betweenness centrality* e *eigenvector centrality*), de transitividade (*clustering*) e de conectividade (Assortatividade, Entropia de Shannon e segundo momento do grau). Podemos obter um *heatmap* entre as features obtidas para as redes geradas utilizando a correlação de Spearman, uma medida que quantifica a colinearidade entre duas variáveis. Ao analisar o *heatmap*, vemos que há grande correlação linear entre diversas feautres, principalmente aquelas que pertencem aos mesmos grupos. Esse resultado é importante pois quando há informação mútua entre variáveis, o grau de influência no resultado de modelos de Aprendizado de Máquina é diluído.

A seguir, realizamos uma revisão das métricas de rede mais importantes para compreensão desse trabalho.

2.3.1 *Closeness Centrality*

Em redes, quanto mais próximo a outros um vértice está, maior a sua importância na rede. Assumindo que as interações entre nós seguem o caminho mais curto, a *Closeness Centrality* de um nó u é definida como o recíproco da distância do caminho mais curto entre u e os outros $n-1$ nós da rede $v = 1, ..., n$.

$$CC(u) = \frac{n-1}{\sum_v d(u,v)}$$

Onde $d(u, v)$ é a distância do caminho mais curto entre v e u. Quanto maior o valor de *Closeness*, maior a importância do vértice na rede. Para caracterizar a rede foi utilizada o *Closeness* médio dos nós.

2.3.2 Coeficiente de *Clustering*

Uma maneira simples de caracterizar a presença de *loops* de tamanho três é através do coeficiente de *Clustering*.

$$C = 3\frac{\#\text{triângulos}}{\#\text{tríades}}$$

O fator 3 leva em conta que cada triangulo pode ser parte de três triplas diferentes, cada uma com um vértice sendo o principal e garante que $C \in [0, 1]$.

2.3.3 Entropia de Shannon

Entropia é um conceito chave em termodinâmica, mecânica estatística e teoria da informação e está relacionada fisicamente com a quantidade de disordem e informação presentes em um sistema. Na teoria ade informação, entropia descreve quanta aleatoriedade está presente em um evento aleatório. Esse conceito pode ser aplicao para o estudo de redes complexas ao calcular a entropia da distribuição do grau. Essa medida provê uma média de heterogeneidade da rede e pode ser definida como

$$H = -\sum_k P(k) \log P(k)$$

O valor máximo de entropia é obtido para uma distribuição uniforme quando todos vértices possuem o mesmo grau e está relacionada com a robustez e resiliência da rede.

2.3.4 Assortatividade

Uma característica muito importante em redes é a presença de conexões homogeneas. Podemos nos perguntar, por exemplo, quão provável é a conexão entre nós similares. A assortatividade mede a similaridade de conexões no grafo com respeito ao grau do nó. Quando vértices de alto grau tendem a se conectar com vértices de alto grau, a rede é assortativa. Por outro lado, se os vértices de alto grau se conectam vértices de baixo grau a rede é disassortativa.

O cálculo da assortatividade é feito através do Coeficiente de Correlação de Pearson r. Caso $r > 0$, a rede é assortativa; se $r < 0$, a rede é disassortativa; para $r = 0$ não existe relação entre o grau dos vértices.

2.4 Aprendizado de Máquina

Nesse trabalho assumimos que o tempo para alcançar consenso Y_i e a frequência de mudança de opinião C_i podem ser inferidos a partir do vetor de *features* X_i. A explicação abaixo foca na predição de Y_i mas também é válida para C_i.

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

Nosso objetivo é encontrar a 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.

2.4.1 Coeficiente de Determinação (R2)

O coeficiente de determinação, R^2, é uma métrica usada para medir o quão bem um modelo de regressão se ajusta aos dados (P. Johnson e Schielzeth 2017).

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

~~A fórmula é mostrada acima~~. Para cada amostra i, y_i é o valor real, $\hat{y}_i$ é o valor predito e $\bar{y}$ é a média dos valores reais. Um valor de 1 significa que o modelo realiza predições perfeitas. De forma contrária, um valor igual ou menor a 0 indica que o modelo não possui habilidade de predição.

No entanto, quando adicionamos mais preditores ao modelo, o R^2 pode aumentar mesmo que esses novos preditores não ajudem realmente a explicar a variação na variável dependente (Bishop 2006; Murphy 2012). Para lidar com isso, utilizaremos o R^2 ajustado, que leva em consideração o número de preditores p e penaliza a inclusão daqueles que são irrelevantes. Esse ajuste fornece uma avaliação mais precisa de quão bem o modelo prevê o resultado. Isso garante uma avaliação mais confiável do desempenho do modelo. Na fórmula abaixo, n indica o número de amostras no conjunto.

$$R^2_{\text{adj}} = 1 - \frac{(1 - R^2)(n - 1)}{(n - p - 1)}$$

2.4.2 *Forward Selection* (FS)

Forward Stepwise Selection é uma maneira eficiente para selecionar *features*, que começa com um modelo sem preditores e adiciona variáveis uma a uma, até que os preditores exigidos estejam no modelo. De modo particular, em cada passo é adicionado o melhor preditor ao modelo. Considerando a alta colinearidade entre as variáveis explicativas, o FS desempenha um papel muito eficiente ao selecionar a melhor variável em cada passo sem descartar suas correlações (James et al. 2014).

```
1. Considere o modelo nulo M0, sem variáveis preditoras.

2. Para k = 0,..., p - 1:
   a) Considere todos p-k modelos que adicionem uma variável ao modelo anterior Mk
   b) Escolha Mk+1 como o melhor entre os p-k modelos

3. Escolha o melhor entre todos modelos M0,...,Mp do passo 2 utilizando uma métrica como R2
```

2.4.3 Validação Cruzada

A fim de analisar os resultados, foi utilizado 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.

A validação cruzada divide o conjunto de treinamento em *k-folds* de tamanho semelhante. O primeiro *fold* é tratado como conjunto de validação, e o modelo é treinado nos *k-1 folds* restantes. A métrica de avaliação é então computada com as observações de validação e o valor é armazenado. Ao final das k iterações, o valor da métrica é a média de cada iteração (James et al. 2014).

A validação cruzada foi utilizada durante FS e para otimização de hiperparâmetros no treinamento das Random Forests. Para o caso de seleção, as duas *features* mais importantes são aquelas com maior frequência entre todos os k folds.

2.4.4 Regressão não Linear

A Regressão não Linear é apropriada em casos onde a variável resposta não é linear de acordo com as *features* e pode ser realizada através de uma transformação não-linear apropriada (James et al. 2014). No nosso caso é utilizado o logaritmo e buscamos estimar os coeficientes $\beta_1, ..., \beta_p$ tal que

$$\log(Y_i) = X_i\beta + \delta$$

também pode ser escrito como

$$Y_i = e^{X_i\beta+\delta}$$

A imagem da função exponencial é $(0, \infty)$, garantindo que o valor estimado Y_i sempre será positivo.

2.4.5 Normalização dos Dados

Para análise de regressão, é essencial que os dados estejam normalizados, a fim de impedir o cálculo de coeficientes imprecisos e ajudar a determinar quais variáveis possuem maior importância. Para tanto, basta subtrair os valores pela média e dividir pelo desvio padrão do conjunto de treino. A normalização é efetuada após a transformação logarítmica (James et al. 2014).

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

2.4.6 Random Forests

Modelos de aprendizado de máquina robustos para dados tabulares envolvem o *ensemble* de árvores de decisão, em que a resposta final é uma média de cada uma das árvores. Dentre esses modelos, as *Random Forests* se popularizaram ao propor uma construção de árvores através de *bootstrap aggregation* ou *bagging*. Em cada passo, uma árvore é treinada a partir de um conjunto obtido a partir de amostragem com reposição do conjunto de treinamento. Após um grande número de árvores ser gerado, uma nova amostra é predita a partir das médias dos valores de todas outras árvores (Breiman 2001).

3 Resultados

3.1 Predição de Variáveis Dinâmicas

A figura 5 apresenta um boxplot com os modelos Random Forest e Regressão não Linear para predição de variáveis dinâmicas. É possível observar que ambos modelos aprendem os dados do conjunto de treinamento, mas o modelo de Regressão não Linear alcança uma generalização levemente melhor. Notavelmente, o segundo modelo possui treinamento mais simples e apresenta maior explicabilidade. Dessa forma, em conjunto com o Forward Selection, se caracteriza como um método para análise topológica da rede, aprofundado na seção 3.3.

3.2 Importância de Features em Random Forests

Análise das features mais importantes foi feita para ambas variáveis resposta e diferentes métodos de inicialização, demarcados de acordo com as cores. As features estão ordenadas de forma descendente no gráfico por importância média. É importante notar a diluição da importância das features no Tempo de Consenso, em que há uma grande distinção entre importância para cada inicialização diferente 6. Já na Frequência de Troca de Opinião, a Variância do Grau é dominante para todas inicializações. De forma contrária, *Eigenvector* e Assortatividade não demonstram nenhuma capacidade preditiva 6. Em todos os casos, não fica claro quais métricas são determinantes para predição das variáveis resposta, reforçando a importância da seleção de *features* nesse cenário.

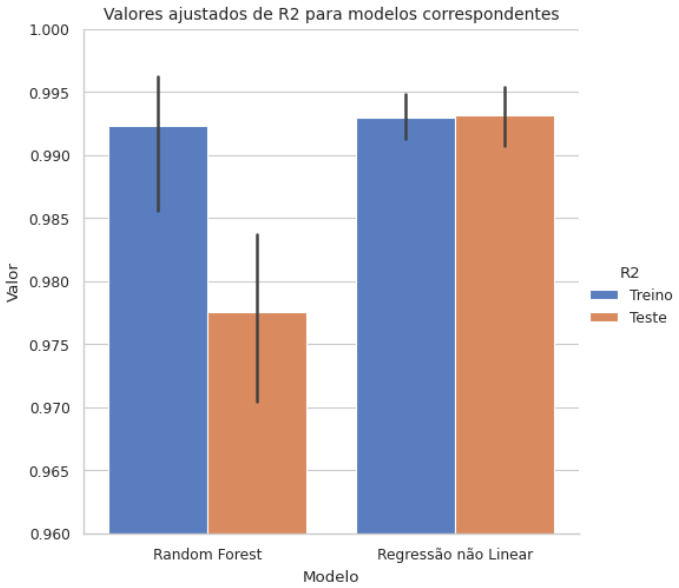

Figura 5: Cada coluna representa a distribuição dos valores ajustados de um modelo e conjunto (treino ou teste) para predição das variáveis dinâmicas. É possível observar que ambos modelos aprendem os dados do conjunto de treinamento, mas o modelo de Regressão não Linear (direita) alcança uma generalização levemente melhor em relação as Random Forests (esquerda).

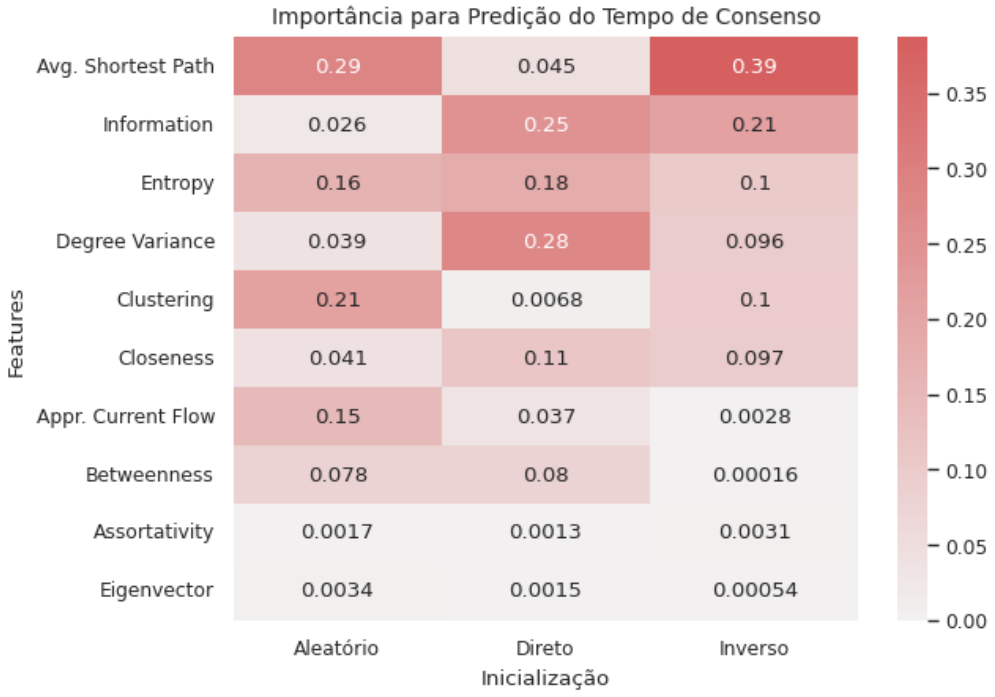

Figura 6: Importância de Features para o Tempo de Consenso: há uma diluição de importância entre as *features*. Apesar da Entropia apresentar um grande impacto, outras características topológicas também demonstram ter um impacto significativo, tornando a análise dessas features mais difícil.

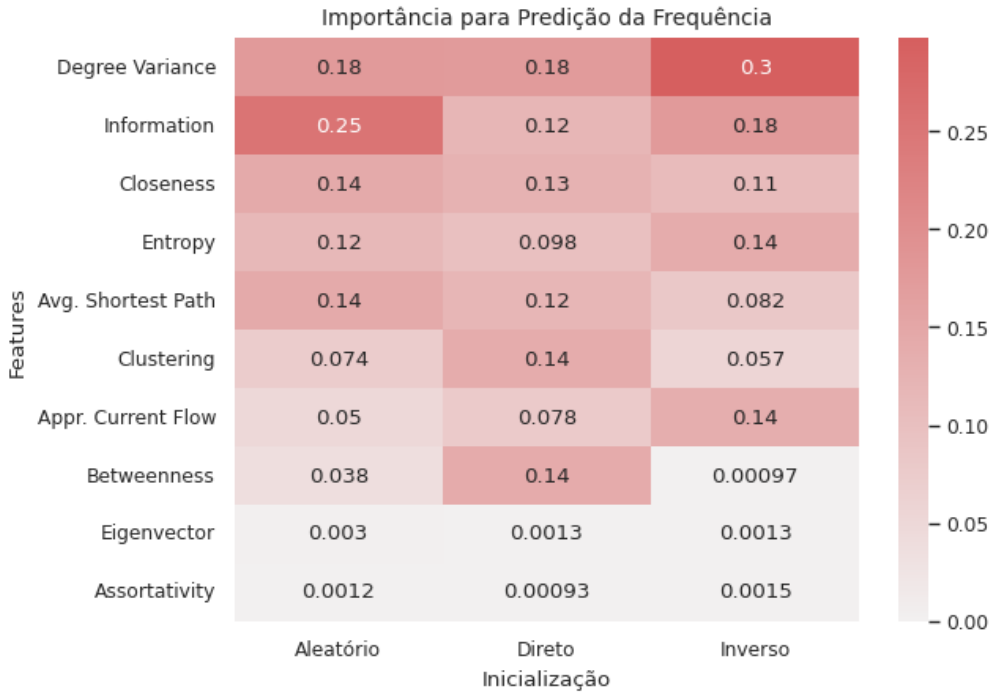

Figura 7: Importância de Features para a Frequência de Troca de Opinião: a Variância do Grau tem grande importância nesse cenário, mas ainda é difícil analisar quais *features* podem ser decisivas na determinação da Frequência de Troca de Opinião.

3.3 Análise das Features utilizando Regressão não Linear e Forward Selection

Os métodos de Regressão com mínimos quadrados nos permitem aprofundar nos resultados para maior interpretabilidade das variáveis resposta através dos coeficientes de regressão, p-valores e outras informações (Murphy 2012). Aqui, realizamos uma seleção empírica das variáveis das seção 2.3 prezando pela diversidade e explicabilidade. Assim, os próximos resultados advém do mesmo cenário da subseção anterior considerando apenas as variáveis descritas na seçao 2.3: Entropia de Shannon, Assortatividade, *Closeness Centrality* e Coeficiente de *Clustering*.

Através das tabelas obtidas para cada uma das variáveis resposta e cada um dos métodos de inicialização, exibidas no apêndice A.1, percebemos através dos baixos valores de *p-value* e *standard error* que há uma grande significância das *features* selecionadas. Além disso, com apenas duas métricas selecionadas, é ~~p~~possível alcançar um coeficiente de determinação superior a 0.98, como exibido na figura 5. No caso da Frequência de Troca de Opinião (figura 8), o *Clustering* se apresenta para as três inicializações em uma relação direta: quanto maior o coeficiente, maior a frequência de troca de opinão. É uma medida que aumenta de acordo com o número de triângulos do total presentes na rede. Dessa forma, é possível elaborar uma tese que a presença de triângulos na rede incita a troca de opiniões entre os indivíduos. Para o Tempo de Consenso (figura 9), pode se observar o *Closeness Centrality* em uma relação inversa com a variável resposta. A *feature* em questão aumenta a medida que a distância média entre os pares de nós na rede diminui. No nosso caso, isso pode indicar que quando, em média, os nós da rede estão mais próximos uns aos outros, menor o tempo necessário para alcançar um estado estacionário.

Finalmente, ao comparar o método proposto com a análise promovida pela importância nas *Random Forests*, percebe-se uma diminuição no espaço de exploração, facilitando a análise e interpretação dos resultados. Além disso, a regressão não linear também apresenta coeficientes explicáveis, permitindo interpretação direta dos resultados. Por exemplo, utilizando as equações da seção 2.4.4 e os resultados de A.1, é possível escrever a Frequência de Troca de Opinião F como a exponencial da combinação linear das *features*.

$$F = \exp(6.48 + \text{clustering} \times 1.222 + \text{closeness} \times 0.082)$$

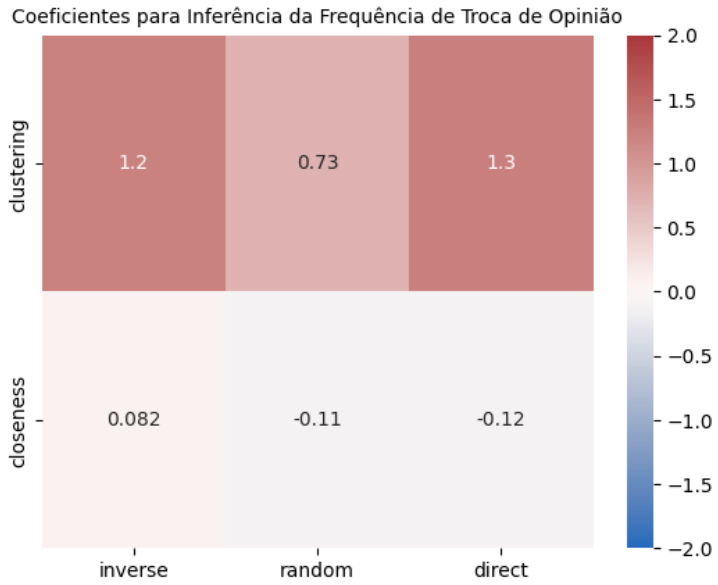

Figura 8: Coeficientes de Regressão obtidos para para Inferência da Frequência de Troca de Opinião de acordo com diferentes inicializações.

4 Conclusão

Nesse trabalho, conseguimos predizer variáveis dinâmicas associadas com o modelo de Sznajd utilizando métricas de topologia de rede. Verificamos que a predição obteve grande acurácia e propusemos um método para obter maior explicabilidade e semelhante acurácia quando comparado a Random Forests. Assim, conseguimos verificar não apenas quais *features* são mais importantes na emergência de polarização, mas também qual o nível de influência. Principalmente, mostramos que o Coeficiente de *Clustering* e *Closeness Centrality* podem ser utilizado para predizer as variáveis dinâmicas associadas as simulações. Além disso, três mudanças nos métodos de inicialização dos nós foram considerados, buscando entender como as medidas topológicas podem ser influenciadas nesse caso. Inicialmente, os nós foram escolhidos de forma aleatória, seguindo o modelo original de Sznajd. Após, nós com maior grau foram selecionados para investigar como seu grande número de conexões pode influenciar a dinâmica da rede. Por fim, os nós na periferia da rede foram selecionados para entender o impacto de agentes menos influentes. Apesar dessas modificações impactarem o resultado das simulações, conseguimos observar que o impacto é pequeno e as métricas selecionadas se conservam ao longo dos experimentos.

Ao analisar a importância de features obtida a partir de experimento com alta acurácia a partir de Random Forests, foi encontrada uma diluição de importância, isto é, duas features tem comportamento semelhate e impactam a regressão de maneira semelhante , o que torna a análise dos dados mais difícil. Tal observação se confirma ao analisar o Heatmap na figura 9. Assim, o método proposto busca encontrar um subconjunto de *features* que obtenha alta acurácia através de Regressão não Linear, possibilitando análise dos coeficientes. Nesse cenário, os coeficientes de Regressão indicam que a presença de triângulos na rede incita a troca de opiniões entre os indivíduos e que, em média, quão menor a distância entre os nós da rede, menor o tempo necessário para alcançar um estado estacionário. Notavelmente, a magnitude de influência de *Closeness Centrality* na predição do Tempo de Consenso aumenta de acordo com o grau dos nós que recebem a opinião predominante, mostrando que a proximidade entre nós é especialmente influente quando os nós dominantes são mais conectados.

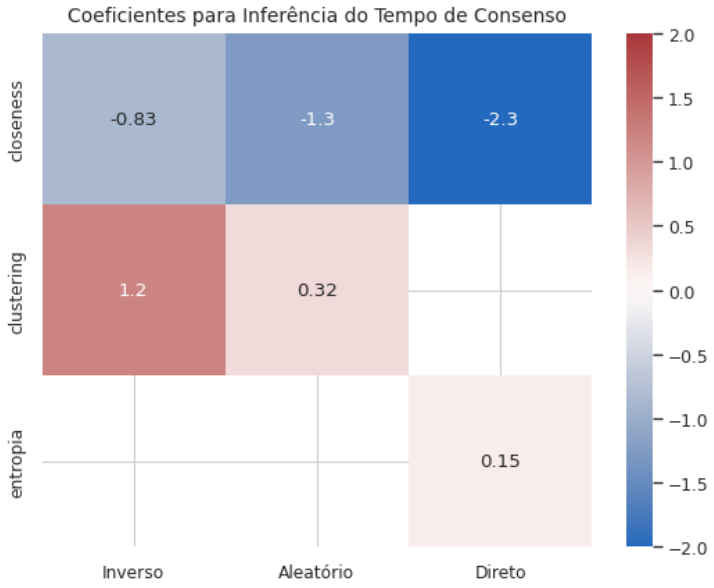

Figura 9: Coeficientes de Regressão obtidos para para Inferência do Tempo de Consenso de acordo com diferentes inicializações.

A expansão da metodologia proposta para predição e análise de variáveis topológicas além do modelo de Sznajd pode promover novos *insights* relativos a diversos cenários e estudos em dinâmicas sociais. No apêndice são apresentados resultados que motivaram o projeto de intercâmbio e estudos em realização. Esperamos que trabalhos futuros nessa direção contribuam para um melhor entendimento de dinâmicas complexas para a polarização e suas implicações. A combinação de aprendizado de máquina com redes complexas tem um grande potencial para revolucionar nossa compreensão de sistemas sociais, levando a um maior entendimento do comportamento e desenvolvimento de estratégias para alcançar resultado social positivo.

5 Referências

Bernardes, A. T., D. Stauffer, e J. Kertész. 2002. «Election results and the Sznajd model on Barabasi network». *The European Physical Journal B - Condensed Matter* 25 (1): 123–27. https://doi.org/10.1140/e10051-002-0013-y.

Bishop, Christopher M. 2006. «Pattern recognition and machine learning». *Springer Google Scholar* 2: 645–78.

Boccaletti, Stefano, Vito Latora, Yamir Moreno, Martin Chavez, e D-U Hwang. 2006. «Complex networks: Structure and dynamics». *Physics Reports* 424 (4-5): 175–308.

Breiman, Leo. 2001. «Random Forests». *Machine Learning* 45: 5–32.

Castellano, Claudio, Santo Fortunato, e Vittorio Loreto. 2009. «Statistical physics of social dynamics». *Reviews of Modern Physics* 81 (2): 591–646. https://doi.org/10.1103/revmodphys.81.591.

Costa, L da F, Francisco A Rodrigues, Gonzalo Travieso, e Paulino Ribeiro Villas Boas. 2007. «Characterization of complex networks: A survey of measurements». *Advances in Physics* 56 (1): 167–242.

Interian, Ruben, e Francisco A Rodrigues. 2023. «Group polarization, influence, and domination in online interaction networks: a case study of the 2022 Brazilian elections». *Journal of Physics: Complexity* 4 (3): 035008. https://doi.org/10.1088/2632-072x/acf6a4.

James, Gareth, Daniela Witten, Trevor Hastie, e Robert Tibshirani. 2014. *An Introduction to Statistical Learning: with Applications in R*. Springer Publishing Company, Incorporated.

Johnson, Paul, e Holger Schielzeth. 2017. «The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded». *Journal of The Royal Society Interface* 14 (setembro): 20170213. https://doi.org/10.1098/rsif.2017.0213.

Johnson, Steven. 2002. *Emergence: The connected lives of ants, brains, cities, and software*. Simon; Schuster.

Layton, Matthew L., Amy Erica Smith, Mason W. Moseley, e Mollie J. Cohen. 2021. «Demographic polarization and the rise of the far right: Brazil's 2018 presidential election». *Research & Politics* 8 (1): 2053168021990204. https://doi.org/10.1177/2053168021990204.

Maia, H. P., S. C. Ferreira, e M. L. Martins. 2021. «Adaptive network approach for emergence of societal bubbles». *Physica A: Statistical Mechanics and its Applications* 572: 125588. https://doi.org/https://doi.org/10.1016/j.physa.2020.125588.

Murphy, Kevin P. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Pineda, Aruane M., Paul Kent, Colm Connaughton, e Francisco A. Rodrigues. 2023. «Machine learning-based prediction of Q-voter model in complex networks». https://arxiv.org/abs/2310.09131.

Rodrigues, Francisco A., Thomas Peron, Colm Connaughton, Jurgen Kurths, e Yamir Moreno. 2019. «A machine learning approach to predicting dynamical observables from network structure». https://arxiv.org/abs/1910.00544.

Sanchez, Juan R. 2004. «A modified one-dimensional Sznajd model». https://arxiv.org/abs/cond-mat/0408518.

SZNAJD-WERON, KATARZYNA, e JÓZEF SZNAJD. 2000. «OPINION EVOLUTION IN CLOSED COMMUNITY». *International Journal of Modern Physics C* 11 (06): 1157–65. https://doi.org/10.1142/s0129183100000936.

A Apêndice

A.1 Tabelas de Resultados para Regressão não Linear

A.1.1 Tempo de Consenso

Inicialização Aleatória: Adj. R2: 0.988

	Coef	p-valor	Std. error
const	5.937	0	0.004
clustering	0.73	0	0.009
closeness	-0.111	0	0.009

Inicialização Direta: Adj. R2: 0.993

	Coef	p-valor	Std. error
const	6.023	0	0.006
clustering	1.265	0	0.011
closeness	-0.121	0	0.011

Inicialização Inversa: Adj. R2: 0.993

	Coef	p-valor	Std. error
const	6.48	0	0.004
clustering	1.222	0	0.009
closeness	0.082	0	0.009

A.1.2 Frequência de Troca de Opinião

Inicialização Aleatória: Adj. R2: 0.993

	Coef	p-valor	Std. error
const	6.626	0	0.006
clustering	0.316	0	0.012
closeness	-1.265	0	0.012

Inicialização Direta: Adj. R2: 0.991

	Coef	p-valor	Std. error
const	6.793	0	0.01
closeness	-2.271	0	0.013
shannon_entropy	0.146	0	0.013

Inicialização Inversa: Adj. R2: 0.997

	Coef	p-valor	Std. error
const	6.606	0	0.005
clustering	1.208	0	0.01
closeness	-0.835	0	0.01

A.2 Análise do Tempo de Consenso em Redes Erdos-Renyi com p variável

O estudo demonstrou relativa facilidade de predição do Tempo de Consenso para casos gerais. Assim, os testes foram expandidos para analisar o Tempo de Consenso em modelos de rede com parâmetros variados. Nesse sentido, foi feito um experimento considerando redes Erdos-Renyi de tamanho $N = 1000$ e $p\,Unif(0,1)$. Note que para essas redes, o p coincide com o coeficiente de *Clustering*. A análise dos dados de simulação gerou o gráfico abaixo.

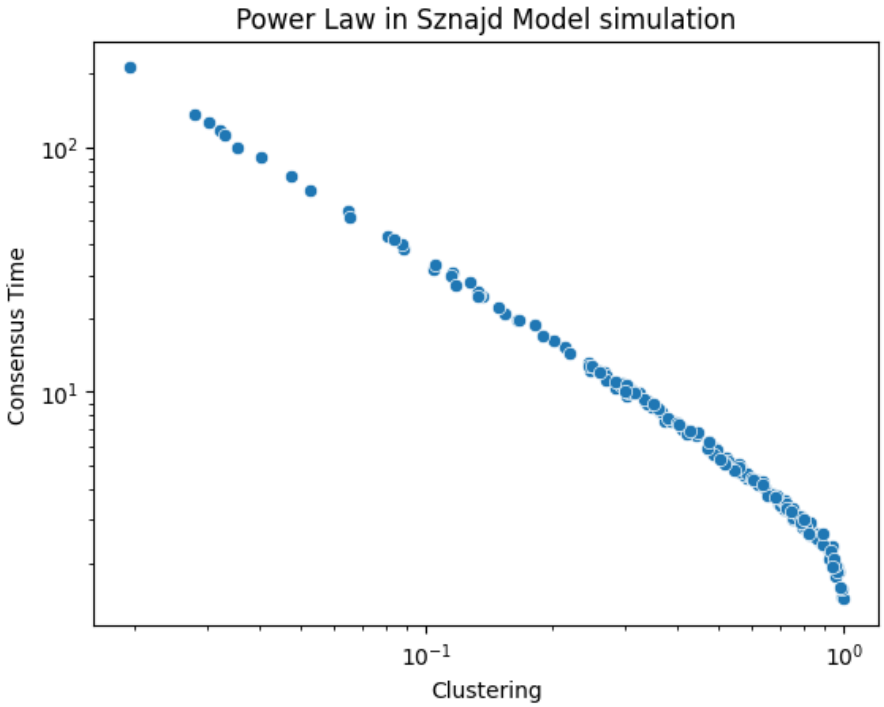

Figura 10: Simulações sugerem uma lei de potência entre o Tempo de Consenso e o coeficiente de *Clustering* nas simulações do modelo de Sznajd em redes Erdos-Renyi com $N = 1000$ e $p\ Unif(0,1)$

É possível observar uma lei de potência entre as métricas de rede e o Tempo de Consenso. Assim, o projeto de Intercâmbio proposto visa se aprofundar no tratamento analítico de dinâmicas de opinião buscando derivar expressões fechadas para os modelos estudados.