CENTRO UNIVERSITÁRIO DA FEI

BIANCA CAMPOS LEMOS DE SOUZA

DESEMPENHO ESCOLAR

UMA ABORDAGEM VIA ANÁLISE DE REDES

São Bernardo do Campo

2016

BIANCA CAMPOS LEMOS DE SOUZA

DESEMPENHO ESCOLAR

UMA ABORDAGEM VIA ANÁLISE DE REDES

Projeto de Iniciação Científica submetido ao Centro Universitário FEI, como parte dos requisitos necessários para obtenção de bolsa de iniciação didático-científica e condução do projeto. Orientado pelo professor Tiago Estrela.

São Bernardo do Campo

2016

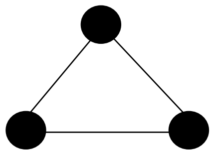
RESUMO

Objetos, de qualquer natureza, que estão interligados, podem ser considerados redes complexas. Ou seja, uma rede complexa pode ser um sistema biológico, a internet, sociedade, entre outros mil sistemas. Esse projeto visa uma espécie de rede complexa social: um grupo de alunos. Nesse grupo, os objetos são os alunos e os relacionamentos de amizade, as interligações. O projeto tem como finalidade a descoberta de uma relação entre as amizades e as notas dos alunos, isto é, se uma amizade influencia no desempenho de um aluno. Para tal, utilizar-se-ão diversas métricas definidas pela teoria do grafo.

Sumário

1 INTRODUÇÃO

Uma rede, assim como um grafo, é todo tipo de conjunto onde os objetos possuem ligações entre si. Nesse contexto, os objetos são chamados de nós ou vértices; enquanto os relacionamentos entre eles, de arestas. A figura a seguir é uma representação de rede simples.



Nessa rede, observa-se a existência de três nós e três arestas. Com isso, é possível determinar várias métricas.

A teoria do grafo pode ser aplicada em diversos temas. Os tipos de redes mais evidentes, na sociedade, são: grupo de pessoas (social), tecnologia e internet. A rede social , os cujos nós são as pessoas; os relacionamentos, as arestas, pode abranger um grupo pequeno de pessoas, ou a sociedade em si; ainda que a quantidade de elementos seja diferente, as características são idênticas. O mesmo acontece com tecnologia e internet.

A tecnologia e a internet estão ligadas uma à outra, porém, são redes extremamente divergentes. Os nós e as arestas da rede tecnológica são, respectivamente, computadores (ou celulares, notebooks etc.) e cabos (ou conexão wifi), à medida que a rede de internet se refere ao world wide web, cujos nós são os links e as arestas, os hiperlinks. É conveniente ressaltar que essas espécies de redes mudam a todo momento. Milhares de pessoas nascem em um dia, ao mesmo tempo que milhões de páginas novas são criadas e outros milhares de computadores são instalados. Portanto, a dinamicidade é um aspecto de redes complexas.

Diversas métricas foram criadas para facilitar o estudo de uma rede. Segundo os autores do livro Social Networks Analysis, Stanley Wasserman e Katherine Faust, o grau de um nó determina a quantidade de arestas ligadas a ele, e a variância indica o quanto este varia. Tais características são extremamente importantes para a definição de uma rede, uma vez que distingue determinada rede das outras. Outro aspecto significativo é o coeficiente de clusterização. Sobre ele, há um trecho muito interessante no artigo Complex Networks: Structure and dynamics (2005, p. 9): “Clusterização, também conhecido como transitividade, é uma propriedade típica de redes conhecidas, onde dois indivíduos com amigos em comum possuem uma probabilidade de se conhecerem. ”. Todos os conceitos relevantes para este projeto serão definidos na Teoria do Grafo.

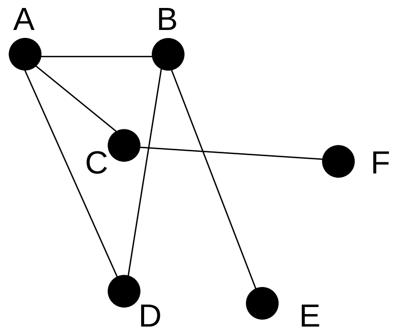
2 OBJETIVO

Este trabalho tem por objetivo identificar se existe relação, utilizando teoria de redes complexas, entre alunos que estudam juntos e suas notas.

3 TEORIA DO GRAFO

Uma rede complexa possui as mesmas características de um grafo, e, portanto, pode ser definida matematicamente através da Teoria dos Grafos. Definem-se os objetos da rede como nós, e os relacionamentos, como arestas. Por consequência, na representação a seguir, os nós são os pontos ou as bolinhas; e as arestas, os traços. É irrelevante como essas bolinhas e traços são desenhados; o ponto principal da representação é se um ponto está conectado a outro. Uma rede é intitulada não direcionada quando apresenta arestas não orientadas, ou seja, a ligação de A com B implica que haja uma ligação de B com A. Como exemplo, a rede não direcionada, demonstrada abaixo.

Figura 1- Representação da rede S



Dessa maneira, é possível ter uma visão completa da rede e, assim, os teoremas e as análises são aplicados mais facilmente, e, além disso, características difíceis de detectar tornam-se mais perceptíveis.

Grafos, que possuem apenas um ponto ou nó, são denominados triviais; todos os outros são considerados não triviais. Grafos que não apresentam nenhuma linha (aresta) são intitulados vazios. Esses grafos são extremamente raros em redes sociais, devido à grande aparição de relacionamentos entre as pessoas.

No cotidiano, observam-se diversos exemplos de redes não direcionadas, como exemplo, a rede social. A rede social, cujos nós são as pessoas e as arestas, os relacionamentos, não possui ligações direcionadas, ou seja, se a pessoa A se relaciona com a B, a pessoa B também se relaciona com a A. A representação acima apresenta tal característica.

Há a rede onde os relacionamentos não são simétricos, ou seja, a aresta possui um sentido; essa rede é denominada direcionada. Nesse caso, os traços da representação serão substituídos por setas. Um primeiro nó se relaciona com um segundo; porém esse segundo não se relaciona com o primeiro. Essa característica é típica de rede web, onde um site possui um hiperlink que leva ao outro; todavia, esse outro site não possui (necessariamente) um hiperlink que leve de volta ao primeiro.

3.1 Definição de rede

Para formalizar uma rede através da matemática, consideram-se n o número de nós, e m o número de arestas. O conjunto de todos os nós de uma rede será denotado por V, ou seja, $n=|V|$. Os objetos i e j pertencem a V e o par ordenado (i,j) representa a existência de aresta. O conjunto de arestas será denotado por E, e, assim, $m=|E|$. Desta forma, é possível determinar o conjunto E na linguagem matemática: $E=\{(i,j)|i,j \in V, \text{ i está relacionado com j}\}$. Por fim, uma rede é definida por estes dois conjuntos, $R= (V, E)$.

3.2 Matriz de adjacência

Outra representação possível é a matricial, que facilita a computação de dados da rede. Essa matriz é chamada de adjacência e indica se um nó está conectado a outro através de números. Indica-se tal matriz por A e esta é quadrada de ordem n (caso o par i,j esteja relacionado: $A(i,j)=1$; do contrário: $A(i,j)=0$). Utilizando o mesmo exemplo do gráfico, monta-se a tabela que dará origem à matriz A.

Tabela 1- Tabela com dados da rede S

	A	B	C	D	E	F
A	-	1	1	1	0	0
B	1	-	0	1	1	0
C	1	0	-	0	0	1
D	1	1	0	-	0	0
E	0	1	0	0	-	0
F	0	0	1	0	0	-

Os traços (-) presentes na tabela indicam que uma determinada pessoa não pode ter relação com ela mesma, por isso, na matriz, esse traço será representado pelo número zero. Com essa tabela, é possível construir a matriz A

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Nota-se que matrizes de redes não direcionadas são sempre simétricas, por consequência, a matriz A é igual a sua transposta. Já as matrizes de redes direcionadas podem ser ou não simétricas, dependendo da relação entre i e j.

3.3 Grau

O grau de um nó representa o número de arestas ligadas a ele e é denominado di, sendo i um determinado vértice. O valor do grau pode variar de zero a m-1. Em uma rede direcionada, há dois tipos de grau: *indegree* e *outdegree*. *Indegree* é o número de arestas que chegam até um determinado nó; *outdegree* é o número de arestas que saem de tal nó. Como a matriz e o grafo representam uma rede não direcionada, os dois tipos de grau serão iguais. Através do grafo, observa-se que três arestas estão ligadas ao nó A, por conseguinte o grau de A vale 3. Esse resultado pode ser obtido somando-se os elementos da respectiva linha de A na matriz de adjacência. É interessante observar que a soma do grau de todos os vértices é igual ao dobro de número de arestas, isto é: $\sum_{iEV} di = 2m$. Obtém-se o grau médio dividindo a soma do grau de todos os vértices pelo número de nós, ou seja, $\bar{d} = \frac{2m}{n}$. O valor do grau médio da rede S é 1,667. A variância dos graus, exemplificada como S_D^2, indica quanto variam os valores dos graus da rede; matematicamente, temos: $S_D^2 = \frac{\sum_{i-1}^{m}(di-\bar{d})}{m}$.

3.4 Centralidade

Um nó é considerado central quando possui o maior número de arestas conectadas ao mesmo, ou seja, possui o maior número de grau da rede. O valor da centralidade de cada nó é o seu próprio grau ($C_D(ni) = d(ni)$). Como o valor máximo da centralidade é m-1, a padronização é dada por: $C'_D = \frac{d(ni)}{m-1}$.

3.5 Densidade

O número máximo de arestas possível é dado pela expressão: $\frac{n(n-1)}{2}$. A densidade de uma rede é o número de arestas presente na rede, divido pelo número máximo de arestas possível na rede, ou seja: $\rho = \frac{2m}{n(n-1)}$. O valor mínimo da densidade é zero; isso ocorre quando a rede não possui arestas ($m = 0$). Já o valor máximo é um, nesse caso, o grafo é dito completo, pois o número de arestas é o máximo possível ($m = \frac{n(n-1)}{2}$). Juntando a equação de grau médio com a de densidade, tem-se a equação abreviada: $\rho = \frac{\bar{d}}{n-1}$. A densidade da rede S vale 0,333.

3.6 Distribuição de grau

Outro aspecto estrutural importante é a distribuição empírica de grau, cujo valor indica o número de vértices que possui certo grau k. O termo “empírico” diferencia esse conceito da distribuição de probabilidade. Sendo nk o número de nós com grau k; e fk, a distribuição de grau, temos: $fk = \frac{nk}{n}$. Entretanto, a distribuição complementar cumulativa do grau (CCDF) é mais utilizada para analisar uma rede. A CCDF é a fração de vértices que tem grau maior ou igual a k; esse valor é calculado através da expressão: $Fk = 1 - \sum_{i=0}^{k-1} fi$. É notável que a CCDF é sempre uma função decrescente.

3.7 Caminho, distância e diâmetro

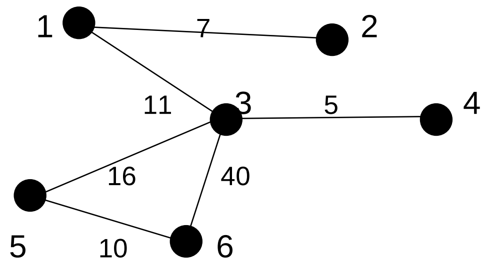
O caminho, a distância e o diâmetro de uma rede são características significativas. O caminho é definido por uma sequência de nós que possuem aresta ligando-os. Para a rede S, o caminho de A até E são as arestas entre A e B, e, B e E. O comprimento de um caminho é o número de arestas que o define. Deste modo, o comprimento de A até E são 2. No entanto, a distância é o caminho mais curto entre dois nós. Seja l(i,j) a distância entre os vértices i e j, define-se a distância média por: $\bar{l} = \frac{\sum_{i,j \in V} l(i,j)}{\binom{n}{2}}$. Essa propriedade é fundamental em redes sociais, visto que quantifica o quão longe estão um par de vértices. O diâmetro de uma rede é a maior distância entre qualquer par de vértices da rede, assim sendo: $L = \max l(i,j)$.

3.8 Coeficiente de clusterização

A ideia do coeficiente de clusterização, ou transitividade, é a probabilidade de dois vértices, que estão conectados a um terceiro, também apresentarem arestas entre si. Em termos de um grafo genérico G, transitividade representa um alto número de triângulos. Matematicamente, essa propriedade é a fração de arestas que os vizinhos de i possuem entre si e o máximo de arestas possível que poderiam ter, de outro modo: $ci = \frac{2Ei}{di(di-1)}$, sendo Ei o número efetivo de arestas entre os vizinhos do nó i e $\binom{\mathrm{di}}{2}$ o maior número de arestas entre os vizinhos. Constata-se que o coeficiente de clusterização não está definido para vértices com grau zero; em outras palavras, a definição só se aplica para grau maior ou igual a 1. O coeficiente de clusterização é a média aritmética destes, por outra forma: $\bar{c} = 1/n \sum_{i \in V} ci$. Caso exista algum coeficiente não definido ($di = 0$), a média deve ser apenas sobre os vértices que possuem coeficientes definidos. No exemplo da rede (nome da rede), os coeficientes de clusterização de A, B, C, D, E e F são 0, 0, 0, ½, 0, 0, respectivamente. Então, coeficiente de clusterização da rede vale 0,5.

3.9 Grafos com peso

Existem grafos cujas arestas estão associadas à um valor. Estes são chamados de grafos com peso, visto que cada ligação possui um peso. É importante ressaltar que grafos com peso podem ser direcionados ou não. A seguir, um exemplo de grafo com peso (direcionado):



Observa-se que a aresta que conecta os nós 1 e 2 possui peso 7. O caminho mais curto entre dois nós de um grafo com peso, é aquele que a soma dos pesos das arestas é menor. Por exemplo, é possível determinar dois caminhos divergentes entre os nós 3 e 6, o caminho mais curto, nesse caso, será as arestas de 3 a 5 e de 5 a 6, pois a soma do peso dessas arestas é 26, e a soma do peso da aresta de 3 a 6 (diretamente) é 40.

4 MATERIAIS E MÉTODOS

Esse projeto será constituído de duas fases. Na primeira, utilizar-se-á a turma da disciplina CA2211 a ser lecionada no primeiro semestre de 2017. Esta turma servirá de modelo para uma análise preliminar. Na segunda fase do projeto, serão aplicadas as técnicas obtidas na primeira fase em três turmas de uma disciplina do departamento de matemática na engenharia, gerando um aumento nos volumes de dados.

CA2211 foi escolhida por ser uma disciplina que pertence ao segundo semestre do currículo de ciência da computação. Sendo assim, os discentes já possuem um grau de amizade entre si. Além disto, ela possui uma única turma, viabilizando as análises preliminares visto a quantidade de dados.

A seguir, serão descritas algumas das técnicas utilizadas neste trabalho. Por simplicidade será considerado um exemplo de uma turma fictícia contendo 5 alunos, a saber, Pedro, Maria, Paulo, Ana e João. Monta-se uma tabela estudo, ver tabela 2, contendo os nomes dos estudantes nas linhas e nas colunas, cada um dos estudantes indicará com um X quais destes outros alunos ele estuda. Observe que Maria estuda com Pedro, Paulo e Ana, já Ana estuda com Pedro Maria e João.

Tabela 2: Exemplo de uma tabela estudo de uma turma fictícia contendo 5 alunos.

	Pedro	Maria	Paulo	Ana	João
Pedro		X	X	X	X
Maria	X		X	X	
Paulo	X	X			X
Ana	X	X			X
João	X		X	X	

Fonte: Autor.

Com a tabela 2 pode-se criar uma matriz de adjacência, $A = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix}$, utilizando 1 para o X e zero para a lacuna vazia. Munido da matriz de A podemos determinar várias métricas como o grau de cada nó, a centralidade e o coeficiente de clusterização.

Suponha que os alunos Pedro, Maria, Paulo, Ana e João tenham tirado na P1 3,4,5,7 e 6 respectivamente. Com estas notas, monta-se uma tabela, ver tabela 3, e

Tabela 3: Tabela contendo as notas dos alunos com os quais cada um dos discentes estudou.

	Pedro	Maria	Paulo	Ana	João
Pedro		4	5	7	6
Maria	3		5	7	
Paulo	3	4			6
Ana	3	4			6
João	3		5	7	

Fonte: Autor.

gera-se uma matriz $B = \begin{pmatrix} 0 & 4 & 5 & 7 & 6 \\ 3 & 0 & 5 & 7 & 0 \\ 3 & 4 & 0 & 0 & 6 \\ 3 & 4 & 0 & 0 & 6 \\ 3 & 0 & 5 & 7 & 0 \end{pmatrix}$ de adjacência associada a ela. Observa-se que esta matriz de adjacência está associada a um grafo cujas arestas tem pesos, sendo necessário explorar métricas para este tipo de grafos. Munido das métricas diferenciadas para este tipo de grafo espera-se obter algum tipo de relação entres as duas redes.

Visto o tamanho dos dados é impraticável fazer cálculos destas métricas de forma manual necessitado assim de um software específico. Para tal tarefa utilizaremos o software UCINET® (Borgatti 2002).

5 PLANO DE TRABALHO DO BOLSISTA

Candidata:

As atividades que serão desenvolvidas pela candidata serão as seguintes:

Atividade 1: Revisão bibliográfica;

Atividade 2: Aplicação do questionário sobre amizade dos alunos na disciplina CA2211;

Atividade 3: Treinamento e uso dos softwares UCINET®;

Atividade 4: Aplicação das técnicas de redes utilizando o UCINET® nos dados de amizade dos alunos na disciplina CA2211.

Atividade 5: Aplicação das técnicas de redes utilizando o UCINET® nas notas de P1 dos alunos na disciplina CA2211.

Atividade 6: Análise dos resultados obtidos nas atividades 4 e 5.

Atividade 7: Elaboração do Relatório Parcial.

Atividade 8 Aplicação das técnicas de redes utilizando o UCINET® nas notas de P2 dos alunos na disciplina CA2211 e comparação com os dados já obtidos anteriormente.

Atividade 9: Aplicação do questionário sobre amizade dos alunos na disciplina da engenharia;

Atividade 10: Aplicação das técnicas de redes utilizando o UCINET® nos dados de amizade dos alunos na disciplina CA2211.

Atividade 11: Aplicação das técnicas de redes utilizando o UCINET® nas notas de P1 dos alunos na disciplina CA2211.

Atividade 12: Análise dos resultados obtidos nas atividades 10 e 11.

Atividade 13: Elaboração do Relatório Final.

A Tabela 4, por seu turno, apresenta o cronograma de desenvolvimento dessas atividades

Tabela 4 – Cronograma de atividades a serem desenvolvidas pela candidata à bolsa de iniciação científica do Centro Universitário da FEI.

	Mês											
Atividade	1	2	3	4	5	6	7	8	9	10	11	12
1	X											
2		X										
3		X	X									
4			X	X								
5				X	X							
6					X	X						
7					X	X						
8							X					
9							X					
10								X	X			
11									X	X		
12										X	X	
13											X	X

Fonte: Autor.

6 BIBLIOGRAFIA

Borgatti, S. E. (2002). Ucinet 6 for Windows: Software for Social Network Analysis. Harvard.

Wang, X.; Chen, G. Complex Networks: small-word, scale-free and beyond, v.3, p. 6-20, 8 sep 2003.

S. Boccaletti; V. Latora; Y. Moreno; M. Chavez; D.-U. Hwang. Complex Networks: Methods and dynamics, p. 3-11, 27 ouc 2005.

Wasserman, S.; Faust, K. (1994). Social Network Analysis: Methods and aplications. Cambridge University Press.

Margarida Dias, Tópicos de Matemática Discreta. Disponível em: <http://www.mdias.uac.pt/MatemDiscreta/Cap4_cont2.pdf>. Acesso em 8 de dezembro de 2016.

Virginia Tech, Data Structures and Algorithms. Disponível em: <http://courses.cs.vt.edu/~cs3114/Fall10/Notes/T22.WeightedGraphs.pdf>. Acesso em 8 de dezembro de 2016.