Usando análises sociais na identificação de nós relevantes em um cenário multirredes: Operação Licitante Fantasma, um estudo de caso

Bruno Figueiredo[1], Fabiola Nakamura[1], Gardenya Felix[2], Eduardo Nakamura[1]

[1] Instituto de Computação, Universidade Federal do Amazonas (UFAM) – Manaus, AM – Brazil

[2] Programa de Pós-Graduação em Educação, Universidade Federal de Roraima (UFRR) – Boa Vista, RR – Brazil

bruno_cbf@uerr.edu.br, fabiola@icomp.ufam.edu.br, gardenyafelix2009@hotmail.com, nakamura@icomp.ufam.edu.br

Abstract. *This paper proposes the NDNS (Nodes Detection using Network Science) model that, using complex networks theory, aims to find the most relevant nodes in a multi-network scenario in a more efficient way than well-known and established centrality metrics. The paper uses, as a case study, a Brazillian corruption investigation of public tenders known as Ghost Bidder operation. Considering four years of investigation, the NDNS model, when compared to four metrics of centrality (betweenness, eigenvector, weighted degree, page rank, and its normalized geometric mean), achieved 93% precision and 94% recall in detecting fraudulent values against 38% and 51%, respectively, of the second-best positioned measures.*

Resumo. *Este artigo propõe o modelo NDNS (Nodes Detection using Network Science) que, usando redes complexas, busca encontrar os nós mais relevantes, em um cenário multi-redes, de forma mais eficiente do que medidas de centralidade estabelecidas. O artigo utiliza, como estudo de caso, uma investigação de corrupção em licitações públicas no Brasil – Operação de Licitante Fantasma. Considerando um período de quatro anos de investigações, o NDNS, quando comparado a quatro medidas de centralidade (betweenness, eigenvector, weighted degree, page rank e sua média geométrica normalizada), alcançou uma precisão de 93% e uma revocação de 94% na detecção de valores fraudulentos contra 38% e 51%, respectivamente, das segundas medidas mais bem posicionadas.*

1 Introdução

O termo "licitação pública" refere-se a uma concorrência deflagrada pelo setor público para convidar fornecedores a concorrer pelo fornecimento de bens e serviços [Mankiw and Taylor 2011, 223–26]. Um dos tipos mais populares de fraude em licitações públicas é o conluio com a formação de carteis, onde empresas, supostamente concorrentes, simulavam uma disputa onde uma, previamente escolhida, seria a ganhadora. Essa prática visava a majoração de preços e eliminação da concorrência. A grande frequência e diversidade de esquemas de fraude inviabiliza os processos manuais de detecção e combate [Koh et al. 2003].

A operação do Licitante Fantasma [Diário Digital 2017; Federal 2017] teve como objetivo investigar a formação de um cartel envolvendo um grupo de empresas de suprimentos médicos e hospitalares, entre 2013 e 2016, no Mato Grosso do Sul.

Neste artigo, é proposto o modelo NDNS (*Nodes Detection using Network Science*) que, usando redes complexas, busca encontrar os nós mais relevantes, em um cenário multirredes, de forma mais eficiente do que medidas de centralidade estabelecidas. Como estudo de caso são usados os dados da operação do Licitante Fantasma mapeados em quatro redes complexas, sendo cada rede correspondente a um ano de licitações suspeitas, entre 2013 e 2016.

O modelo NDNS considera que em cenários descritos por mais de uma rede complexa, os nós relevantes advêm da interseção dessas redes e não da junção delas, como comumente é feito na aplicação de medidas de centralidade. Essa estratégia elimina distorções do tipo: em um cenário composto por muitas redes, se um determinado nó tem uma grande relevância em poucas redes, quando da junção, ele poderá continuar tendo uma grande relevância total. Esse tipo de distorção é eliminado, uma vez que a relevância do nó é diretamente proporcional ao número de redes em que ele ocorre. Esse tipo de abordagem levou o modelo NDNS a obter resultados superiores a medidas de centralidade quanto às métricas: precisão, revocação e F_1; na detecção de valores fraudulentos.

A organização do trabalho é a seguinte: a seção 2 discute os trabalhos relacionados à detecção de entidades mais relevantes num contexto, com foco especial à detecção de fraudes e de seus participantes, a seção 3 apresenta o modelo NDNS, a seção 4 descreve a operação do licitante fantasma e aplica no modelo NDNS ao estudo de caso e a seção 5 descreve as contribuições da pesquisa e aponta as direções futuras.

2 Trabalhos Relacionados

Estratégias para determinar as entidades mais relevantes em um cenário são temas recorrentes em estudos científicos. A lista vai de técnicas de mineração de dados, inteligência artificial, redes complexas, entre outros.

A seguir serão discutidos trabalhos relacionados à detecção de fraudes por meio da detecção das entidades mais relevantes, podendo ser essas entidades empresas ou indivíduos suspeitos de fraude. Complementarmente há uma discussão do modelo NDNS frente a essas tecnologias.

2.1 Mineração de Dados e Inteligência Artificial

O uso de heurísticas e agentes inteligentes, associados a técnicas de mineração de dados, foi testado com sucesso em bancos de dados reais de auditoria como uma ferramenta contra conluios em licitações públicas [Cunha, Rodrigues, and Bugarin 2014; C. V. S. Silva and Ralha 2010]. Encontra-se uma abordagem semelhante no desenvolvimento de uma ferramenta de mineração de agentes (AGMI) que, usando dados reais da Controladoria Geral da União (CGU), propõe ser uma ferramenta de previsão de detecção e prevenção de corrupção na identificação de formação de cartel, alcançando 90% de precisão [Ghedini Ralha and Sarmento Silva 2012].

Uma abordagem bem-sucedida é o uso de mineração de dados e regras de auditoria na detecção de cartéis. O uso de padrões de preços pode indicar a existência dessas associações em licitações públicas [C. V. S. Silva and Ralha 2010; Costa and Aparicio 2011; C. V. S. Silva and Ralha 2011].

A Controladoria Geral da União (CGU) utiliza ontologias e redes bayesianas como ferramentas para evitar fraudes no setor público [Hu et al. 2013], para tanto as ontologias consideram as informações semânticas e.g., relações sociais ou comerciais entre indivíduos ou empresas [Hu et al. 2013; Balaniuk et al. 2013].

No entanto, todos esses trabalhos restringem-se à detecção de fraudes em licitações públicas, já o NDNS se propõe a ser um modelo generalista de detecção nós relevantes, podendo ser utilizado, como no estudo de caso, para a detecção de fraudes.

2.2 Redes Complexas

Uma rede complexa é um grafo em que os nós representam entidades e as arestas seus relacionamentos. Sua estrutura topológica é irregular, não trivial e evolui com o tempo [Boccaletti et al. 2006]. Com base nas características das conexões entre entidades, utilizando as chamadas “medidas de centralidade”, é possível identificar as principais entidades (nós) do cenário considerado.

As medidas de centralidade quantificam a relevância dos nós em redes complexas sob aspectos distintos. Como exemplos de centralidades tem-se: a *betweenness centrality* representa o número de caminhos mínimos que passam por um nó [Otte and Rousseau 2002]. As centralidades *eigenvector* e *page rank* consideram o número de links de um nó para classificá-lo [Bonacich 2007]. A distribuição do grau ponderado (*weighted degree*) corresponde à soma dos pesos das arestas do incidente em um nó v e está relacionada à "popularidade" dos nós [Beveridge and Shan 2016].

A busca por novas medidas e seu aprimoramento é um objeto de estudo atual. Por exemplo: em relação à medida de centralidade k-shell [Kitsak et al. 2010] possui limitações para informar as posições topológicas dos nós melhor classificados. Existe uma proposta de uma nova medida hierárquica especial que funciona como uma melhoria do k-shell nesse aspecto [Zareie and Sheikhahmadi 2018].

As medidas de centralidade podem ter desvantagens na tarefa de identificar os nós mais significativos de uma rede complexa no caso de cenários nos quais as decisões são baseadas em vários critérios. Uma estratégia alternativa de avaliação pode ser, por exemplo, ranquear nós baseada em grupos [Yang et al. 2020]. O modelo NDNS tem um foco semelhante, ou seja, classifica os nós em uma estratégia hierárquica baseada em grupos, como uma alternativa às medidas de centralidade.

As medidas de centralidade têm limitações para encontrar nós relevantes de redes complexas quando a tarefa envolve cenários descritos por múltiplas redes. Uma estratégia pode ser utilizar uma abordagem baseada no tempo capaz de entender como a relevância dos nós aumenta e diminui nas redes sendo aplicável a sistemas dinâmicos [Fire and Guestrin 2020]. Uma proposta de algoritmo baseado em *k-means* para redes ligadas utilizando a modelagem de blocos também discute as vantagens de lidar com uma coleção de redes vinculadas, e.g. multiníveis, ao invés de uma única rede [Žiberna 2020]. O modelo NDNS também lida com várias redes para descrever cenários e propõe uma abordagem alternativa para combinar essas redes.

2.3 O Modelo NDNS e os Trabalhos Relacionados

O modelo NDNS, ao contrário de outros trabalhos cuja aplicabilidade é restrita a situações específicas [Balaniuk et al. 2013; Hu et al. 2013; Bansal, Gaur, and Singh 2016; Costa and Aparicio 2011; C. V. S. Silva and Ralha 2011, 2010; L. A. D. Silva 2016; Neville et

al. 2005; Carvalho 2014; Cunha, Rodrigues, and Bugarin 2014], é útil para situações genéricas com a vantagem de ser aplicável a uma vasta gama de circunstâncias.

Alguns trabalhos lidam apenas com tipos específicos de dados, por exemplo, informações textuais [Skillicorn and Purda 2012; L. A. D. Silva 2016], ou indicadores de dados [Virdhagriswaran and Dakin 2006; Bhowmik 2008; Neville et al. 2005]. O NDNS pode gerenciar qualquer tipo de informação que possa ser modelada em uma rede, tornando-o mais aplicável e flexível.

3 O modelo NDNS

Nesta seção, serão utilizadas redes e teoria dos conjuntos na formalização do NDNS.

A hipótese é: *considerando um conjunto de redes, com elementos sendo os nós e suas relações como arestas, que, isoladamente, representem partes de um cenário; a intersecção dessas redes revela os nós mais relevantes de todo o cenário sendo o grau de importância dos nós diretamente proporcional ao número de vezes em que cada nó aparece nessas intersecções*.

A Figura 1 mostra o esquema de funcionamento do NDNS. A construção das redes mapeia os dados originais, tendo os elementos como nós, e.g., empresas, e as arestas as relações entre esses elementos, e.g., participação conjunta em licitações. Essas redes, conjuntamente, representam um único cenário.

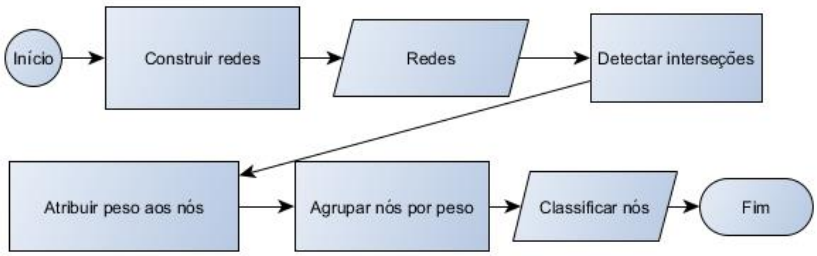


Figura 1. Esquema geral de trabalho NDNS.

O primeiro passo corresponde ao estabelecimento de um conjunto de redes N

$$N = \{N_o, \ldots, N_{j-1}\} \tag{1}$$

onde $\{N_o, \ldots, N_{j-1}\}$ são as redes que representam todo o cenário. Assim,

$$N_i = \{\{v_1, v_2, \ldots, v_y\} \mid N_i \in N\} \tag{2}$$

define uma rede N_i como um conjunto de nós ($\{v_1, v_2, \ldots, v_y\}$).

O segundo e terceiro passos definem a função W

$$W(v_i) = \sum_{x=0}^{|N|-1} |\{v_i \mid v_i \in N_x \wedge N_x \in N\}| \tag{3}$$

que associa um peso $W(v_i)$ a cada nó v_i, indicando o número de vezes um nó aparece em cada uma das redes que compõem N, i.e. suas intersecções. O peso indica a suposta relevância de um nó. Quanto maior o peso associado, maior a relevância do nó.

A quarta etapa agrupa os nós de acordo com seus respectivos pesos:

$$G(j) = \{v_i \mid W(v_i) = j\}, \quad (4)$$

que classifica os nós de acordo com sua suposta relevância. Quanto maior o valor de j, maior a suposta relevância do nó pertencente ao grupo. Para fins de análise, considerar-se-á como relevantes os grupos com $j \geq 2$.

Como critério de ranqueamento dos nós dentro de cada grupo, será considerado o grau dos nós, considerando que a quantidade de arestas é diretamente proporcional à importância do nó no grupo. O quinto passo, portanto, classifica os nós em seus grupos.

Define-se o conjunto de arestas *E* como o produto cartesiano, dois a dois, de todos os conjuntos que compõem *N*, assim:

$$E = \{N_x \times N_y \mid N_x, N_y \in N \wedge x \neq y\}, \quad (5)$$

logo, E é composto por pares ordenados (v_i, v_j) representando as arestas que conectam todos os nós de todas as redes pertencentes a N. Utiliza-se arestas não direcionadas para conectar os nós, logo, a ordem dos nós nos pares ordenados não é relevante. A relevância R de um nó v_i (Equação 8) é dada pela divisão do grau Gr nó v_i (Equação 6) pelo maior grau $MaxGr$ dentre todos os nós pertencente a pares ordenados de E (Equação 7). Assim, define-se $Gr, MaxGr$ e R como:

$$Gr(v_i) = \left|\{(v_i, v_j) \mid (v_i, v_j) \in E \wedge v_i \neq v_j\}\right|, \quad (6)$$

$$MaxGr(E) = max(\{Ne(v_i) \mid (v_i, v_j) \in E\}), \quad (7)$$

$$R(v_i, E) = \frac{Ne(v_i)}{MaxNe(E)}. \quad (8)$$

Considerando o exemplo da Figura 2, tem-se $N = \{N_0, N_1, N_2\}$ (Equação 1), onde $N_0 = \{v_1, \ldots, v_6\}$, $N_1 = \{v_4, v_5, v_7, v_8, v_9, v_{12}\}$ e $N_2 = \{v_5, v_6, v_8, v_9, v_{10}, v_{11}\}$ (Equação 2).

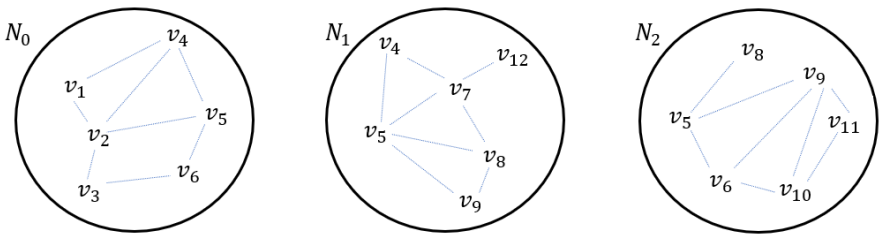


Figura 2. Exemplo de três redes genéricas com intersecções não vazias.

Em seguida aplica-se a função W (Equação 3) - para todos os nós. Por exemplo, o peso do nó v_4 é dado por $W(v_4) = |\{v_4 \mid v_4 \in N_0 \wedge N_0 \in N\}| + |\{v_4 \mid v_4 \in N_1 \wedge N_1 \in N\}| + |\{v_4 \mid v_4 \in N_2 \wedge N_2 \in N\}| = 1 + 1 + 0 = 2$ Os pesos são calculados para todos os nós presentes nas redes (Figura 3).

Pode-se observar que os nós $v_1, v_2, v_3, v_7, v_{10}, v_{11}$ e v_{12}, apesar de pertencerem aos conjuntos *N0, N1* e *N2*, não têm um peso associado *W*. Isso acontece porque esses nós têm peso igual a um e, presumivelmente, não são relevantes. Nos nós v_4, v_5, v_6, v_8 e v_9, os pesos indicam o número de vezes que esses nós aparecem em um conjunto N_i. Por exemplo, $W(v_4) = 2$ porque v_4 está presente nas redes *N0* e *N1* (Figura 3).

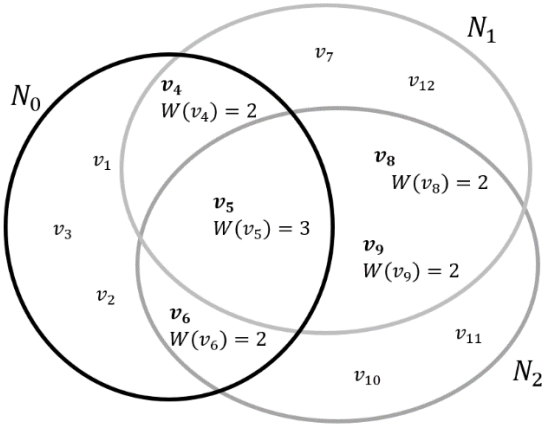


Figura 3. Exemplo de aplicação do modelo NDNS para três redes {N_0, N_1, N_2}, com seus nós {v_1, …, v_{12}} e o pesos associados $W(v_x)$.

Divide-se então esses nós em grupos, de acordo com sua suposta relevância (Equação 4), isto é, $G(2) = \{v_4, v_6, v_8, v_9\}$ e $G(3) = \{v_5\}$. As arestas são representadas como pares ordenados no conjunto E (Equação 5), e.g., (v_1, v_4), refere-se à aresta que conecta nós v_1 e v_4 no conjunto N_0 (Figura 2).

O modelo NDNS divide os nós em grupos que têm sua própria ranqueamento. Isso significa que o primeiro critério para apontar os nós relevantes é o grupo e, depois disso, a ranqueamento dos nós naquele grupo. Em outras palavras, o último nó ranqueado em um grupo $G(i+1)$ é, presumivelmente, mais relevante do que o primeiro nó classificado do grupo $G(i)$, mesmo que, numericamente, o ranking do nó pertencente a $G(i)$ (Equação 8) seja superior ao do nó pertencente ao grupo $G(i+1)$. Isso significa que, para um nó ser considerado relevante, não é só necessário que ele tenha uma alta ranqueamento, mas também que ele deva estar presente no maior número de redes possível.

Como o grupo $G(3)$ tem apenas um nó (v_5) serão classificados os nós do grupo $G(2)$, para os quais aplica-se a Equação 6 a todos os elementos, a saber: $Gr(v_4) = 4$, $Gr(v_6) = 1$, $Gr(v_8) = 2$ e $Gr(v_9) = 5$. Para obter o ranking normalizado, aplica-se a Equação 7, em nosso exemplo $MaxGr(E) = 6$ e, finalmente, aplica-se a Equação 8 a todos os nós do grupo $G(2)$, a saber: $R(v_4, E) = 0.66$, $R(v_6, E) = 0.16$, $R(v_8, E) = 0.33$, and $R(v_9, E) = 0.83$. Dessa forma verifica-se que v_9 é, presumivelmente, o nó mais relevante do grupo $G(2)$, seguido por v_4, v_8, e v_6 (Tabela 1).

Tabela 1– Ranqueamento dos nós do grupo $G(2)$.

nodes	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}
v_1		√		√								
v_2	√		√	√	√							
v_3		√										
v_4	√	√			√		√					
v_5		√		√		√	√	√	√			
v_6					√							
v_7				√	√				√			√
v_8					√				√			
v_9					√		√	√		√	√	

v_{10}									√		√	
v_{11}									√	√		
v_{12}							√					
$Gr(v_i)$	2	4	1	4	6	1	4	2	5	2	2	1
$R(v_i, E)$	0.33	0.66	0.16	0.66	1	0.16	0.66	0.33	0.83	0.33	0.33	0.16

4 Aplicando o modelo NDNS à operação do Licitante Fantasma

A Lei de acesso à informação (Lei no 12.527, de 18/11/2001) exige que todos os dados de licitações públicas estejam disponíveis por meio de uma API[1] (*Application Programming Interface*). Utilizou-se essa API na extração dos dados de todas as licitações federais no Estado de Mato Grosso do Sul no período de 2013 a 2016. Esses dados subsidiaram a construção das redes complexas que analisamos neste artigo.

O modelo NDNS lida com várias redes complexas que, juntas, descrevem um cenário onde a escolha natural neste trabalho a atribuição de cada ano de licitações a uma rede, sendo as empresas os nós e as arestas indicando a participação conjunta dessas empresas em uma mesma licitação.

O passo seguinte foi aplicar às redes as equações de 1 a 4, com o objetivo de determinar os pesos associados a cada nó e agrupá-los segundo seu grau de relevância, onde a relevância dos nós cresce à medida que o peso aumenta.

As interseções dessas quatro redes resultam em três grupos de nós — $G(2)$ a $G(4)$ —, divididos de acordo com seus pesos associados. Na Figura 4, pode-se notar que o número de nós aumenta significativamente à medida que o peso diminui sendo, portanto, o peso do grupo inversamente proporcional ao número de nós nele contido.

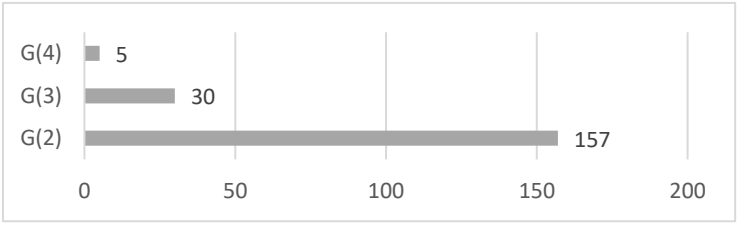


Figura 4 - Número de nós por grupo.

As seções seguintes comparam o modelo NDNS com as medidas de centralidade: *betweenness* (BW), *eigenvector* (EI), *weighted degree* (WD), *page rank* (PG) e sua média geométrica normalizada, intitulada de *score* (SC). Para balizar essa análise, se propõe também uma métrica de avaliação.

4.1 Métrica de avaliação proposta

O modelo NDNS divide os nós em grupos de acordo com sua relevância. Para fins de comparação com outras medidas, toma-se o número de nós em cada grupo e o mesmo número de nós classificados pelas outras medidas (centralidades), preservando sua ordem. A descrição do processo encontra-se na Tabela 2, sendo *i* o maior peso alcançado e *j* o número de medidas de centralidade a serem comparadas.

[1] http://compras.dados.gov.br/docs/home.html

Tabela 2 – Ranqueamento dos nós para *i* grupos e *j* centralidades.

	Centralidade 1	**Centralidade 2**	**…**	**Centralidade *j***
Grupo $G(i)$				
Nó 1	ranking do nó 1	ranking do nó 1	…	ranking do nó 1
...	…	…	…	…
Nó n	ranking do nó n	ranking do nó n	...	ranking do nó n
Grupo $G(i-1)$				
Nó n+1	ranking do nó n+1	ranking do nó n+1	...	ranking do nó n+1
...	...	...	...	...

Aplicando a ranqueamento ao estudo de caso tem-se que, para o grupo $G(4)$ com 5 nós, todas as outras medidas (centralidades), tomam os 5 primeiros nós melhor classificados. Dessa forma, é possível comparar o grupo $G(4)$ com os nós de melhor ranqueamento em todas as outras medidas de centralidade. O grupo $G(3)$ tem 30 nós, portanto, toma-se os nós classificados entre 6 e 35, em todas as outras medidas. Esse processo é o mesmo para o grupo $G(2)$.

Para fins de análise, o que realmente importa é quanto dinheiro cada empresa ganha, sendo, portanto, coerente a proposta de uma métrica que considere valores como parâmetro de relevância das empresas. A relevância dos grupos será aferida pelos valores obtidos pelas empresas pertencentes a cada grupo, verificando se os grupos mais relevantes foram capazes de apontar as empresas com maiores ganhos.

A capacidade de apontar empresas condenadas também foi usada como um parâmetro de eficácia. Esse critério visa verificar se a separação dos nós em grupos reflete a capacidade desses grupos de apontar a existência de fraude. Ou seja, se os grupos mais relevantes puderam apontar um número maior de empresas condenadas.

4.2 O peso como parâmetro para definir grupos no modelo NDNS

A Figura 5 mostra a distribuição normalizada dos valores ganhos pelas empresas em licitações, divididas por ano e grupo. Ou seja, tomou-se o valor máximo de R$ 402.925,58, relativo à soma dos valores ganhos por empresas do grupo $G(4)$ no ano de 2016, e obteve-se os demais valores normalizados para cada grupo/ano. Verifica-se que o crescimento dos valores é coerente com o grau de relevância dos grupos — $G(2)$ a $G(4)$. A única exceção ocorre no grupo $G(4)$ em 2014.

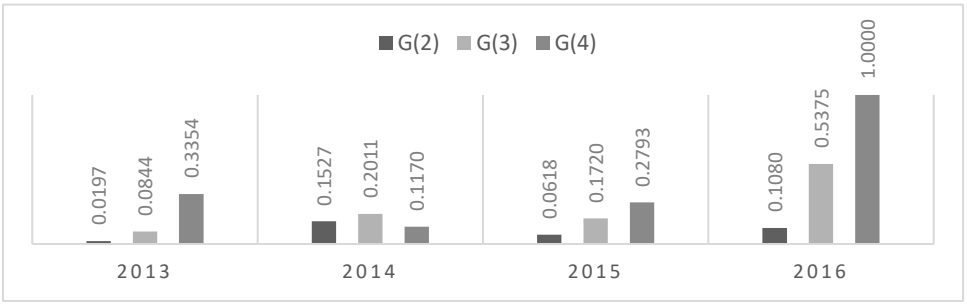
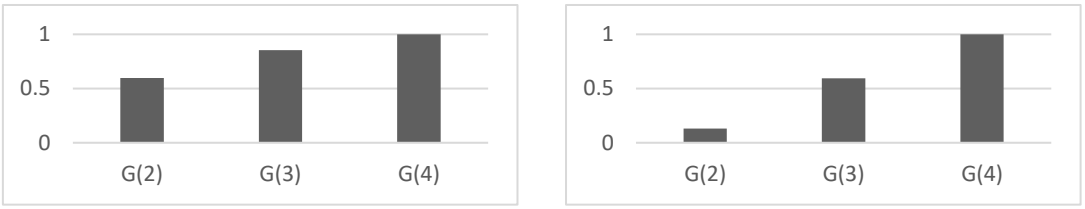


Figura 5 – Distribuição de valores normalizados por grupos.

A Figura 6 mostra a distribuição normalizada dos valores ganhos pelas empresas em licitações, considerando todo o período. Verifica-se que o grupo $G(4)$ possui os valores mais altos que diminuem nos grupos $G(3)$ e $G(2)$, respectivamente. É importante perceber que a separação de empresas em grupos mostra sua coerência não apenas no modelo NDNS, mas também quando consideradas todas as medidas de forma conjunta.



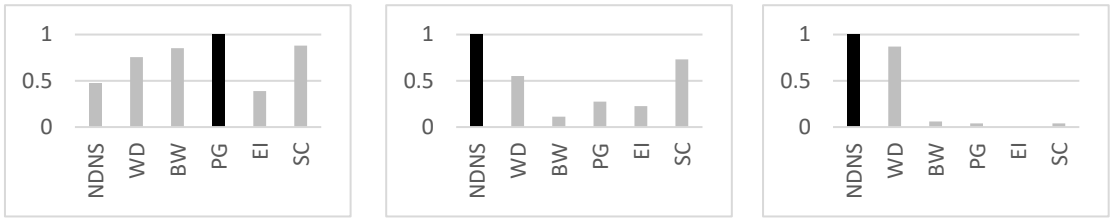
NDNS, WD, BW, PG, EI e SC somadas NDNS

Figura 6 – Distribuição de valores normalizados por grupos.

4.3 Relevância por Grupo

A Figura 7 traz a análise comparativa da relevância de cada medida da centralidade por grupo. Observa-se que, para os grupos $G(3)$ e $G(4)$, o modelo NDNS foi capaz de apontar as empresas mais relevantes, segundo a métrica proposta (Seção 4.1).

Para o grupo $G(2)$, a centralidade *page rank* foi capaz de apontar os nós mais relevantes, mas não pode ser considerado um resultado ruim, uma vez que o grupo $G(2)$ supostamente deve ter as empresas menos importantes. Considerando isso, a distribuição dos valores desejados é exatamente essa, ou seja, as empresas mais importantes (maiores ganhos) estão nos grupos também mais relevantes e as empresas menos importantes (menores ganhos) nos grupos menos importantes. Assim, o fato de o modelo NDNS apontar empresas com menor ganho no grupo $G(2)$ é o resultado esperado.

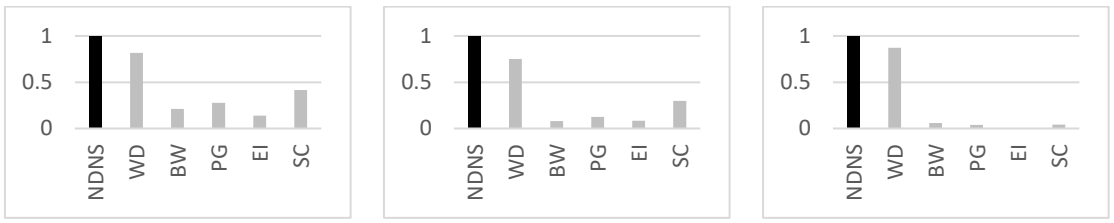


$G(2)$ $G(3)$ $G(4)$

Figura 7 - Análise comparativa da relevância normalizada por grupo.

4.4 Relevância por grupo (uma análise cumulativa)

A Figura 8 traz a análise comparativa cumulativa da relevância por grupo. Pode-se notar que o modelo NDNS conseguiu apontar as empresas mais relevantes (maiores ganhos), obtendo o melhor desempenho geral, considerando a métrica proposta.



$G(2)$ - $G(4)$ $G(3) - G(4)$ $G(4)$

Figura 8 – Análise comparativa cumulativa de relevância normalizada por grupo.

4.5 Análise de empresas (precisão, revocação, F_1 e acurácia)

A Figura 9 considera apenas as empresas condenadas. Nesta análise, verifica-se que o modelo NDNS acançou uma precisão de 0,93 e uma revocação de 0,94 contra 0,51 e 0,4

das segundas medidas mais bem posicionadas, respectivamente *score* (SC) e *Page Rank* (PG). Com relação à média harmônica da precisão e da revocação (F_1), o modelo obteve 0.93, contra 0.38 da segunda medida mais bem posicionada, o *Page Rank* (PG).

Quanto à acurácia, que leva em consideração tanto as empresas condenadas como as não condenadas, o modelo NDNS obteve a segunda melhor marca (0,53), sendo superado apenas pelo *Eigenvector* (EI), com 0,75.

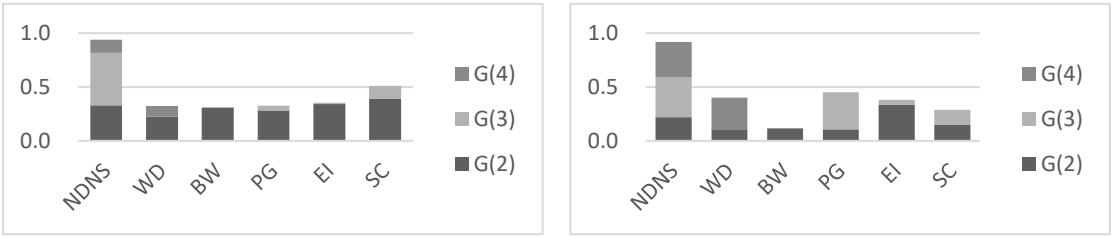


Figura 9 – Análise da revocação (esquerda) e precisão (direita).

4.6 O ranqueamento dos nós em cada grupo

O modelo NDNS classifica os nós pela aplicação das equações 6, 7 e 8 a todos os grupos. Como exemplo, tem-se o grupo com menor número de nós — $G(4)$ (Tabela 3).

Tabela 3 – Ranqueamento de NDNS para $G(4)$.

	Grau	*Total Normalizado*
PGA Servicos Terceirizados Eireli - - Epp	184	1.000
Lideranca Limpeza e Conservacao Ltda	178	0.967
Total Administracao de Servicos Terceirizados Ltda - Epp	117	0.636
Planalto Limpeza e Conservacao de Ambiente - Eireli - Epp	109	0.592
Clima Teck Climatizacao Ltda - Epp	50	0.272

Como parâmetro comparativo, assume-se a posição em todas as medidas de centralidade de cada nó presente no grupo $G(4)$ (Tabela 4) e calcula-se média das posições. Logo, quanto menor a média, maior a relevância do nó. Verifica-se que a ranqueamento proposta (Tabela 3) coincide com a obtida pelo cálculo da posição média (Tabela 4).

Tabela 4 – Ranqueamento dos nós NDNS $G(4)$ em cada medida de centralidade.

Nós do grupo $G(4)$	BC	WD	EV	PR	SC	Média
PGA Servicos Terceirizados Eireli - - Epp	84	2	49	14	8	31.4
Lideranca Limpeza e Conservacao Ltda	72	1	164	13	9	51.8
Total Administracao de Servicos Terceirizados Ltda - Epp	92	60	108	61	29	70
Planalto Limpeza e Conservacao de Ambiente - Eireli - Epp	154	45	106	68	39	82.4
Clima Teck Climatizacao Ltda - Epp	49	211	150	46	53	101.8

5 Conclusão

Foi proposto neste trabalho o modelo NDNS, uma abordagem centrada em grupos para encontrar nós relevantes em cenários descritos por várias redes complexas. Como prova de conceito, foi utilizado o estudo de caso de conluio entre empresas participantes de licitações públicas (Controladoria Geral da União, 2019).

Dada uma métrica proposta, o modelo se mostrou superior às medidas de centralidade avaliadas (*betweenness*, *eigenvector*, *weighted degree*, *page rank* e sua média geométrica normalizada) na detecção de empresas fraudulentas. Essa análise foi feita de forma segmentada, por grupos de nós ($G(2)$ a $G(4)$), onde o modelo apresentou resultados superiores na grande maioria das situações abordadas.

Além disso, o modelo alcançou uma precisão de 93%, uma revocação de 94%, uma acurácia de 53% e um F_1 de 93%, na detecção de valores fraudulentos, sendo esses valores bem superiores aos alcançados pelas medidas de centralidade avaliadas. Exceto quanto à acurácia, onde o modelo atingiu a segunda posição. O modelo também apresentou um ranqueamento que se mostrou coerente com os das mesmas medidas.

Como próximas etapas, pretende-se aplicar o modelo NDNS a outros estudos de casos para atestar sua eficácia. O cálculo da complexidade do algoritmo é necessário para garantir sua viabilidade. Pretende-se também estender o modelo para redes multinível de forma a propor uma centralidade aplicável a redes desse tipo.

A base de dados, os scripts, análises, gráficos e as redes complexas que subsidiaram à elaboração deste trabalho encontram-se disponíveis em https://data.mendeley.com/datasets/s8p55kp2dp/draft?a=922eefee-28ed-40ea-b9f4-fda46123c08e.

Referências

Balaniuk, Remis, Pierre Bessiere, Emmanuel Mazer, and Paulo Cobbe. 2013. "Collusion and Corruption Risk Analysis Using Naïve Bayes Classifiers." In *Communications in Computer and Information Science*. https://doi.org/10.1007/978-3-642-42017-7_7.

Bansal, Rashi, Nishant Gaur, and Shailendra Narayan Singh. 2016. "Outlier Detection: Applications and Techniques in Data Mining." In *Proceedings of the 2016 6th International Conference - Cloud System and Big Data Engineering, Confluence 2016*. https://doi.org/10.1109/CONFLUENCE.2016.7508146.

Beveridge, Andrew, and Jie Shan. 2016. "Network of Thrones." *Math Horizons*. https://doi.org/10.4169/mathhorizons.23.4.18.

Bhowmik, Rekha. 2008. "Data Mining Techniques in Fraud Detection." *Journal of Digital Forensics, Security and Law*. https://doi.org/10.15394/jdfsl.2008.1040.

Boccaletti, Stefano, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. 2006. "Complex Networks: Structure and Dynamics." *Physics Reports*. https://doi.org/10.1016/j.physrep.2005.10.009.

Bonacich, Phillip. 2007. "Some Unique Properties of Eigenvector Centrality." *Social Networks*. https://doi.org/10.1016/j.socnet.2007.04.002.

Carvalho, José Carlos Oliveira. 2014. *Por Dentro Das Fraudes: Como São Feitas, Como Denunciá-Las, Como Evitá-Las.* Edited by Digitaliza Brasil. 1st ed.

Costa, Carlos J., and Manuela Aparicio. 2011. "Using Data Mining to Help Auditors." In *Creating Global Competitive Economies: A 360-Degree Approach - Proceedings of the 17th International Business Information Management Association Conference, IBIMA 2011*.

Cunha, Flávia Ceccato, Rodrigues, and Maurício Soares Bugarin. 2014. "Lei de Benford e Auditoria de Obras Públicas: Uma Análise de Sobrepreço Na Reforma Do Maracanã." *Revista Do TCU*, 48–53. https://revista.tcu.gov.br/ojs/index.php/RTCU/article/view/63.

Diário Digital. 2017. "PF e CGU Deflagram Operação 'Licitante Fantasma,'" March 21, 2017. http://www.diariodigital.com.br/policia/pf-e-cgu-deflagram-operacao-licitante-fantasma/155891/.

Federal, Polícia. 2017. "PF Desarticula Organização Criminosa Que Fraudava Licitações No

MS.” 2017. http://www.pf.gov.br/agencia/noticias/2017/03/pf-desarticula-organizacao-criminosa-que-fraudava-licitacoes-no-ms.

Fire, Michael, and Carlos Guestrin. 2020. “The Rise and Fall of Network Stars: Analyzing 2.5 Million Graphs to Reveal How High-Degree Vertices Emerge over Time.” *Information Processing and Management*. https://doi.org/10.1016/j.ipm.2019.05.002.

Ghedini Ralha, Célia, and Carlos Vinícius Sarmento Silva. 2012. “A Multi-Agent Data Mining System for Cartel Detection in Brazilian Government Procurement.” *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2012.04.037.

Hu, Bo, Nuno Carvalho, Loredana Laera, Vivian Lee, Takahide Matsutsuka, Roger Menday, and Aisha Naseer. 2013. “Applying Semantic Technologies to Public Sector: A Case Study in Fraud Detection.” In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-642-37996-3_23.

Kitsak, Maksim, Lazaros K. Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hernán A. Makse. 2010. “Identification of Influential Spreaders in Complex Networks.” *Nature Physics*. https://doi.org/10.1038/nphys1746.

Koh, Robin, EW Edmund W Schuster, Indy Chackrabarti, and Attilio Bellman. 2003. “White Paper: Securing the Pharmaceutical Supply Chain. 2003.” *AUTO-ID CENTER, Massachusetts Institute …*. https://doi.org/10.1007/s10611-006-9009-5.

Mankiw, N Gregory, and Mark P. Taylor. 2011. *Principles of Economics, Second Edition*. *Book*. https://doi.org/10.1017/CBO9780511511455.

Neville, Jennifer, Özgür Şimşek, David Jensen, John Komoroske, Kelly Palmer, and Henry Goldberg. 2005. “Using Relational Knowledge Discovery to Prevent Securities Fraud.” In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/1081870.1081922.

Otte, Evelien, and Ronald Rousseau. 2002. “Social Network Analysis: A Powerful Strategy, Also for the Information Sciences.” *Journal of Information Science*. https://doi.org/10.1177/016555150202800601.

Silva, Carlos Vinícius Sarmento, and Célia Ghedini Ralha. 2010. “Utilização de Técnicas de Mineração de Dados Como Auxílio Na Detecção de Cartéis Em Licitações.” In *XXX Congresso Da Sociedade Brasileira de Computação*, 1–14. Belo Horizonte / Brazil.

———. 2011. “Agmi - An Agent-Mining Tool and Its Application to Brazilian Government Auditing.” In *WEBIST 2011 - Proceedings of the 7th International Conference on Web Information Systems and Technologies*. https://doi.org/10.5220/0003333905350538.

Silva, Luis Andre Dutra. 2016. “Utilização de Deep Learning Em Ações de Controle.” *Revista TCU*, 18–23. https://revista.tcu.gov.br/ojs/index.php/RTCU/article/view/1321.

Skillicorn, D. B., and L. Purda. 2012. “Detecting Fraud in Financial Reports.” In *Proceedings - 2012 European Intelligence and Security Informatics Conference, EISIC 2012*. https://doi.org/10.1109/EISIC.2012.8.

Virdhagriswaran, Sankar, and Gordon Dakin. 2006. “Camouflaged Fraud Detection in Domains with Complex Relationships.” In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/1150402.1150532.

Yang, Hu, Jar Der Luo, Ying Fan, and Li Zhu. 2020. “Using Weighted K-Means to Identify Chinese Leading Venture Capital Firms Incorporating with Centrality Measures.” *Information Processing and Management*. https://doi.org/10.1016/j.ipm.2019.102083.

Zareie, Ahmad, and Amir Sheikhahmadi. 2018. “A Hierarchical Approach for Influential Node Ranking in Complex Social Networks.” *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2017.10.018.

Žiberna, Aleš. 2020. “K-Means-Based Algorithm for Blockmodeling Linked Networks.” *Social Networks*. https://doi.org/10.1016/j.socnet.2019.10.006.