Uma análise da colaboração científica numa área da pós-graduação brasileira por meio da modelagem estatística de redes sociais usando ERGM: Estudo de caso

Jorge H. C. Fernandes[1,2], João P. A. Maranhão[2], César A. B. de Andrade[2], Ricardo B. Sampaio[2]

[1]Departamento de Ciência da Computação – Universidade de Brasília (UnB)
Caixa Postal 4466 – 70910-900 – Brasília, DF – Brasil

[2]Pós-Graduação em Ciência da Informação – Universidade de Brasília

{jhcf,rbsam}@unb.br, jpamaranhao@yahoo.com.br, caborges72@gmail.com

Abstract. *Presents an application of statistical methods to the analysis of a social network in order to provide statistically significant information about structural and nodal factors that may influence scientific collaboration. The research universe was the colaboration network for publications in periodicals by researchers linked to brazilian post-graduate programs in the field of Information Science. The research demonstrates a practical application of methods and tools for Exponential Random Graph Models (ERGM). The results and analysis allows probabilistic explanations of the behavior of researchers in terms of network self organization factors and also due to endogenous attributes of the researchers such as gender, program affiliation, field of doctoral study, years since doctoral degree, research focus and others.*

Resumo. *Apresenta uma aplicação de métodos estatísticos à análise de uma rede social, que provê informações estatisticamente significativas sobre aspectos que podem influenciar a colaboração científica. O universo da pesquisa foi a rede de colaboração para publicações em periódicos efetivada por pesquisadores afiliados aos programas de pós-graduação na área da Ciência da Informação no Brasil. A pesquisa demonstra uma aplicação prática de métodos e ferramentas para modelagem de grafos aleatórios da família exponencial (ERGM). Os resultados e sua análise permite explicações probabilísticas sobre o comportamento dos pesquisadores em termos de fatores de auto-organização de redes sociais, bem como devida a atributos endógenos dos pesquisadores, tais como gênero, afiliação a programas de pós-graduação, área de doutorado, tempo decorrido desde a obtenção do grau de doutor, foco de pesquisa e outros.*

1. Introdução

A colaboração científica tem sido objeto de estudo de autores na área da ciência de redes [Newman 2004], que se valem das características e organização das bases de publicações científicas com grandes volumes de dados para desenvolver estudos sobre estruturas das redes e padrões de colaboração entre os objetos desse estudo, que são os pesquisadores e suas publicações conjuntas. A grande maioria dos estudos realizados nesse universo de colaboração científica investiga a organização macroscópica das redes, considerando aspectos descritivos tais como grau médio, diâmetro, centralização e densidade

[Sampaio et al. 2015], enquanto outros consideram a auto-organização complexa ou topológica das redes, por meio de fenômenos tais como percolação, relevância de laços fortes e fracos, ligação preferencial e estruturas de mundo pequeno [Kronegger et al. 2012], além de leis de potência [Barabási et al. 2002]. Poucos estudos sobre colaboração científica tem empregado a modelagem estatística das redes sociais [Harris et al. 2012], que avalia de forma probabilística os laços de colaboração entre pesquisadores, considerando não apenas os fenômenos de auto-organização estrutural de redes complexas, mas também os múltiplos processos sociais aninhados entre si [Lusher et al. 2013, 21].

A proposta do estudo aqui relatado foi apresentar uma aplicação da modelagem estatística de redes sociais ao universo de pesquisadores de uma área específica da ciência no Brasil, a Ciência da Informação. Para tal, são utilizados os dados estruturais das redes de colaboração, bem como os atributos individuais dos pesquisadores, disponíveis em bases de dados abertos tais como nas Plataformas Sucupira[1] e Lattes[2]. Alguns desses atributos são gênero, afiliação a programas de pós-graduação, área de doutorado, tempo decorrido desde a obtenção do grau de doutor, foco de pesquisa e outros.

Os objetivos do presente estudo são dois: (1) apresentar de forma tutorial um guia que facilite o acesso de pesquisadores interessados no desenvolvimento de modelos estatísticos de redes sociais; (2) identificar os aprofundamentos analíticos que podem ser obtidos por meio da modelagem estatística de redes utilizando modelos da família de grafos aleatórios exponenciais, chamada de ERGM ou p*.

O restante deste artigo é organizado em mais seis seções: 2 - O estudo das redes de colaboração científica por meio da análise de redes sociais; 3 - O desenvolvimento de modelos probabilísticos de análise de redes sociais; 4 - A descrição da metodologia de coleta e análise de dados empregada; 5 - Os resultados e sua análise inicial; e 6 - Conclusões sobre o estudo desenvolvido.

2. A análise de redes sociais de colaboração científica

A análise da colaboração científica se concentra em padrões de relações entre os autores, permitindo identificar a disponibilidade e o fluxo de troca de recursos entre os mesmos [Wasserman and Faust 1994]. A análise efetuada neste estudo teve como foco os padrões de colaboração e as características que contribuem para reforçar ou enfraquecer essas possibilidades de colaboração.

A coautoria de um documento é um registro oficial do relacionamento entre dois ou mais autores ou organizações [Glanzel 2002] e apesar do debate antigo a respeito do seu significado e interpretação [Katz and Martin 1997], a análise de coautoria tem sido amplamente utilizada para entender e avaliar os padrões de colaboração científica.

A colaboração científica pode, então, ser definida como a interação que ocorre em um contexto social entre dois ou mais pesquisadores, que facilita o compartilhamento de significado e a realização de tarefas com relação a um objetivo mutuamente compartilhado [Sonnenwald 2007]. Em redes de coautoria, os vértices representam os autores ou as organizações às quais esses estão vinculados, e dois ou mais autores estão conectados se eles compartilham a autoria de uma ou mais publicações [Newman 2004].

[1] https://sucupira.capes.gov.br/sucupira/
[2] http://lattes.cnpq.br/

3. Modelos ERGM (*Exponential Random Graph Models*)

3.1. Histórico

Desenvolvidos entre 1970 e 1990, os modelos de grafos aleatórios exponenciais (ERGM) - também chamados de modelos p* - têm mostrado ser uma das classes mais promissoras de modelo estatístico, capaz de expressar várias propriedades estruturais de redes sociais [Snijders et al. 2006].

Modelos ERGM podem ser aplicados à análise de [Lusher et al. 2013] redes não-dirigidas ou dirigidas , estratificadas em múltiplas camadas (multivariadas) , bipartidas e longitudinais. Suas aplicações mais comuns são no estudo de redes intra e interorganizacionais. Modelos ERGM não são aptos a modelar redes com pesos em laços relacionais.

3.2. Definição

Conforme apresentam [Robins et al. 2007], em um modelo ERGM os possíveis laços entre atores de uma rede social são tratados como variáveis aleatórias, que podem ser modeladas sob várias suposições de dependências. Essas suposições de dependência possibilitam entender como e porque surgem - e também não surgem - os laços entre atores de uma rede social, e são primariamente expressas por meio da maior ou menor probabilidade de ocorrência de configurações endógenas de vários tipos (relacionamentos mútuos entre atores, assimetria de relações, formação de triângulos, estrelas de grau k etc). As suposições sobre a probabilidade de ocorrência de um laço também podem ter origem exógena à rede, devidas ao valor assumido por um atributo de um nó da díade, ou por uma combinação de valores - seja diferença ou similaridade - dos atributos de ambos os nós de uma díade. As configurações (endógenas) e os efeitos de atributos (exógenos) são inseridas em um modelo ERGM por meio de termos de modelagem (exemplificados na seção 3.4).

Informado por hipóteses socialmente embasadas sobre a formação de laços em uma rede empiricamente obtida, um analista pode inserir esses termos em um modelo exploratório (uma fórmula), na busca de se encontrar de forma generativa ou simulada uma combinação de ponderações (os parâmetros) para a estatística (quantidade de ocorrências) de cada termo, que melhor expresse as probabilidades de se gerar uma classe estatística de rede estruturalmente equivalente a uma rede empiricamente obtida.

3.3. Problematização para uma rede empiricamente obtida

A título de exemplo, os autores buscaram analisar por meio de modelagem ERGM uma rede de colaboração entre os 249 pesquisadores da área de ciência da informação vinculados a pós-graduações *stricto sensu* no Brasil, objeto da investigação relatada neste artigo. Uma apresentação gráfica da rede é feita na figura 1.

Cada vértice da rede representa um pesquisador. As linhas indicam as colaborações científicas entre os pesquisadores. Cada pesquisador está vinculado a uma linha de pesquisa, que faz parte de um programa vinculado a uma universidade ou instituição de pesquisa. Outros atributos de cada pesquisador foram obtidos a partir dos seus registros na Plataforma Lattes.

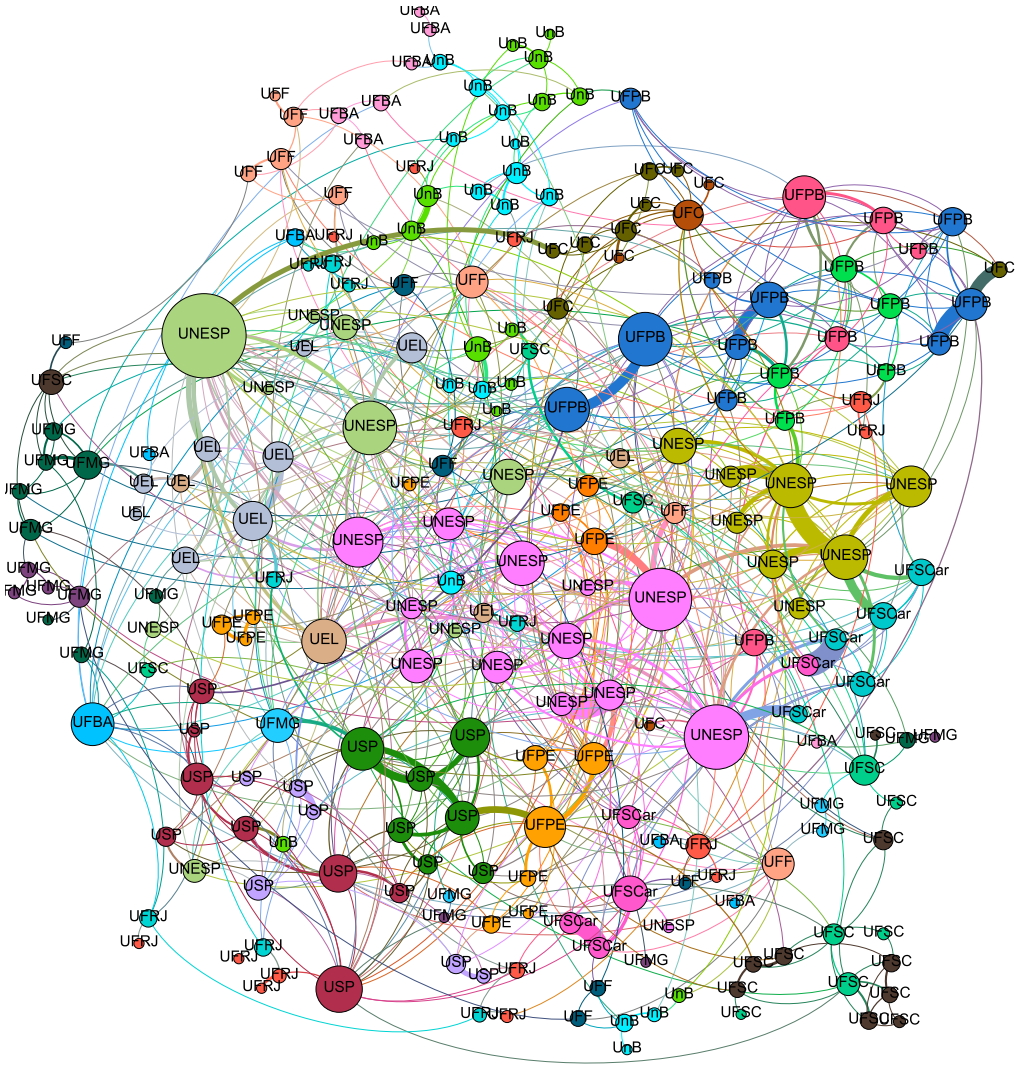

Figura 1. Rede de colaboração entre pesquisadores da Ciência da Informação no Brasil.

3.3.1. Algumas questões que podem ser respondidas por um modelo estatístico de rede

Perante o desafio de desenvolvimento de um modelo ERGM para a rede indicada, apresentam-se algumas questões básicas, como: (1) Qual é a propensão típica dos pesquisadores da rede a estabelecer um relacionamento com outro pesquisador, independentemente de aspectos como gênero, área de doutorado etc? (2) Quais tipos de relacionamentos triádicos são mais comuns na rede? (3) Qual a tendência de um pesquisador a se relacionar com uma multiplicidade de colaboradores? (4) Qual a influência de um determinado atributo de um pesquisador, tal como gênero, programa ao qual está filiado, região do país onde mora, em sua propensão a estabelecer colaborações com outros pesquisadores? (5) Quais outras configurações disponíveis em um modelo ERGM podem ser selecionadas para análise de aspectos como homofilia, propinquidade, seletividade etc, que podem eventualmente estar presentes nessa rede não dirigida em estudo?

Os detalhes envolvidos na modelagem estatística dessa rede específica são apre-

sentados na seção 4, sendo que o restante desta seção apresenta princípios gerais da modelagem usando ERGM, especificamente nos aspectos referentes a: (1) Seleção de termos ERGM; (2) Sensibilidade de modelos ERGM; (3) Ciclo de desenvolvimento de modelos ERGM; (4) Análise de adequação geral de um modelo; e (5) Interpretação de parâmetros estimados.

3.4. Seleção de termos ERGM

Para cada questão formulada visando análise de uma rede empírica um analista que desenvolve um ERGM deve lançar mão de um ou mais termos que expressam possíveis características generativas da rede. Os termos selecionados são inseridos em uma soma de termos, constituindo uma fórmula. Por meio de um pacote de software de modelagem ERGM é possível estimar-se as ponderações mais adequadas (parâmetros) para cada um dos termos inseridos na fórmula, visando assim definir quantitativamente, com maior ou menor grau de confiança, a contribuição individualizada de cada termo para a geração de um conjunto de grafos que possuam características estruturais similares às da rede empiricamente obtida (figura 1).

Termos Básicos O pacote `statnet` [The Statnet Development Team 2017] de análise de ERGM é utilizado na plataforma R, e apresenta termos que permitem a exploração de aspectos estruturais básicos (configurações endógenas) de redes sociais tais como número de arestas (`edges`), densidade (`density`), mutualidade (`mutual(attrname)`) e assimetria (`asymmetric(attrname)`) de díades em redes dirigidas. Os termos que permitem explorar mutualidade e assimetria numa rede dirigida podem ser parametrizados pelo valor de atributos nodais quantitativos ou qualitativos. Por exemplo, numa rede de preferências entre alunos de uma sala de aula os atributos de gênero e raça podem ser avaliados quanto à propensão para criação de relacionamentos mútuos ou assimétricos.

Termos que modelam efeitos exógenos de atributos nodais O pacote `statnet` apresenta termos que permitem estimar o efeito de valores de atributos nodais na propensão ao estabelecimento de laços não dirigidos ou dirigidos entre atores em uma rede. O termo `nodecov(attrname)` permite estimar a probabilidade de estabelecimento de laços entre dois nós, baseada na soma dos valores de um atributo quantitativo presente nesses nós (ex: idade, tempo de trabalho). O termo `nodematch(attrname)` permite estimar a probabilidade de estabelecimento de laços entre dois nós, baseada na igualdade do valor de um atributo categórico desses nós (ex: área de doutoramento, programa de pós-graduação ao qual está vinculado, região do Brasil na qual trabalha, gênero etc). O termo `absdiff(attname)` permite estimar a probabilidade de estabelecimento de laços entre dois nós baseada na diferença absoluta de valores de um atributo quantitativo dos nós (por exemplo, tempo de doutoramento).

Termos de formas paramétricas para análise de parceiros compartilhados Dificuldades para a estimativa adequada dos parâmetros para um modelo ERGM tem sido amenizadas pela introdução de configurações geometricamente ponderadas [Snijders et al. 2006], denominadas de termos paramétricos no pacote `statnet`. A utilização desses termos confere uma estabilidade significativa ao desenvolvimento de

um modelo ERGM. De forma geral, os modelos ERGM mais sofisticados utilizam termos paramétricos tais como `gwdsp(alpha)`, `gwesp(alpha)`, `gwdegree(decay)` e `altkstar(alpha)`.

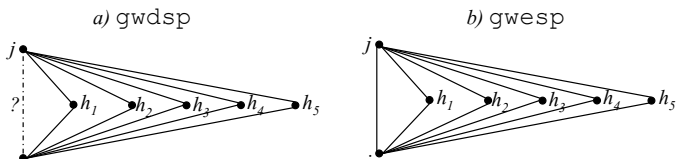

Figura 2. Configurações modeladas pelos termos gwdsp e gwesp.

O termo `gwesp(alpha)` permite estimar de forma geometricamente ponderada a probabilidade de estabelecimento de laços com terceiros nós, entre uma díade que já se relaciona. Graficamente, o termo `gwesp(alpha)` explora a chance de ocorrência de configuração numa rede na forma da figura 2b. Intuitivamente, o termo `gwesp` permite responder à seguinte pergunta: qual a chance de que dois nós que se relacionam estabelecerem uma ligação com terceiros nós (h_1, h_2, ..., h_n), de forma crescente, formando 1-triângulos ($\{(i,j),(i,h_1),(j,h_1)\}$), 2-triângulos ($\{(i,j),(i,h_1),(j,h_1),(i,h_2),(j,h_2)\}$), ..., k-triângulos ($\{(i,j),(i,h_1),(j,h_1),(i,h_2),(j,h_2),...,(i,h_k),(j,h_k)\}$)?

O valor `gwesp.alpha` permite definir inicialmente a curvatura geométrica de um termo paramétrico. O resultado de estimativa de um termo paramétrico apresenta dois valores, sendo o primeiro indicador da probabilidade de geração de um 1-triângulo no caso de `gwesp`. O segundo resultado estimado é o alpha, que indica a curvatura geométrica das probabilidades de geração de k-triangulos.

Outros efeitos O pacote `statnet` apresenta atualmente 64 diferentes termos que permitem explorar a influência de diversas estatísticas na probabilidade de estabelecimento de laços entre nós de uma rede social. Está fora do escopo desse artigo apresentar detalhadamente as possibilidades de uso desses termos, e remetemos ao trabalho de [Morris et al. 2008] para tal detalhamento.

A família de softwares PNET [Wang et al. 2009] também apresenta um amplo conjunto de termos que podem ser explorados no desenvolvimento de um modelo ERGM, especialmente no suporte à analise de redes com múltiplas relações, redes longitudinais e redes bipartidas, não presentes no pacote `statnet`.

3.5. Sensibilidade de Modelos ERGM

O desenvolvimento de um modelo ERGM requer uma abordagem cuidadosa para selecionar e validar os termos, e para analisar os valores obtidos por estimativas de valores para esses termos.

Na estimativa de valores é empregado um processo estocástico para a geração de redes onde cada um dos termos escolhidos atua de forma simultânea a todos os demais termos presentes no modelo. Dessa forma, a interpretação da influência de um termo específico deve levar em consideração que os efeitos de todos os demais termos também já foram computados, especialmente aqueles termos cujos efeitos são sobrepostos aos demais.

Portanto, as estatísticas (contagem da ocorrência de termos numa rede empiricamente obtida ou estocasticamente simulada) e ponderações para cada um dos termos já isolam o efeito individual de cada configuração.

3.6. Ciclo de desenvolvimento de um modelo ERGM

Três das maiores dificuldades no desenvolvimento de um modelo ERGM são:

Facilidade de degeneração do Modelo Os procedimentos de estimativa de parâmetros são muito sensíveis à escolha dos termos do modelo, e facilmente ocorre a degeneração do procedimento sem que haja convergência;

Baixo grau de confiança dos parâmetros estimados Mesmo após a convergência do algoritmo de estimativa dos parâmetros, a confiabilidade dos valores para um ou mais termos pode ser baixa, seja devido à insuficiência de dados da rede, à deficiência nos algoritmos de estimação, ou a interações entre os termos. Nesse caso, um ou mais dos termos devem ser removidos do modelo, ou não analisados.

Geração de efeitos colaterais não investigados Mesmo após a obtenção de um modelo consistente, que gere uma classe de redes que exibem estruturas estatisticamente equivalentes à rede empiricamente obtida, é possível que o modelo esteja enviesado por configurações não exploradas.

Essas dificuldades exigem que o ciclo de desenvolvimento de um modelo ERGM seja composto pelas seguintes etapas: (1) Simulação; (2) Estimativa; (3) Análise de adequação geral do modelo; e (4) Interpretação dos parâmetros estimados. Essas fases do ciclo de desenvolvimento são apresentadas a seguir. Códigos na linguagem R e com o uso do pacote statnet são introduzidos para fins de exemplificação.

3.6.1. Simulação

A simulação permite que um analista ganhe intuição sobre o efeito provocado pelos termos e seus parâmetros em um modelo ERGM. Busca-se simular redes que apresentem estrutura similar à rede empiricamente analisada. O script seguinte gera uma rede não dirigida com 249 nós, e explora o efeito produzido pelos parâmetros (termos) seguintes: -5 (`edges`), 2 (`gwesp`) e 1 (`gwesp.alpha`) na simulação de uma rede. A execução do comando `summary.statistics` gera uma contagem da quantidade de configurações para os termos correspondentes. O grafo simulado possui 492 arestas. No caso específico de gwesp, é apresentada a contagem da quantidade de 214 1-triângulos (`esp#1`), 72 2-triângulos (`esp#2`), 20 3-triângulos (`esp#3`) e 2 4-triângulos (`esp#4`). Não foram encontradas ocorrencias de triângulos acima de 4.

```
>library(statnet)       # carrega a biblioteca statnet
>parametros<-c(-5,2,1) # parâmetros iniciais
>rede.inicial <- network(249,directed=FALSE) # rede inicial
>formula <- formula(rede.inicial ~ edges + gwesp(2)) # formula
>rede.simulada<- simulate(formula, nsim=1, coef=parametros)
>summary.statistics(rede.simulada ~ edges + gwesp(2))
edges  esp#1  esp#2  esp#3  esp#4  esp#5  esp#6  esp#7  esp#8  esp#9
492    214     72     20      2      0      0      0      0      0
esp#10 esp#11 esp#12 esp#13 esp#14 esp#15 esp#16 esp#17 esp#18 esp#19
0      0      0      0      0      0      0      0      0      0
esp#20 esp#21 esp#22 esp#23 esp#24 esp#25 esp#26 esp#27 esp#28 esp#29
```

```
0      0      0      0      0      0      0      0      0      0
esp#30
0
```

3.6.2. Estimativa

A estimativa dos valores dos parâmetros é feita com base na utilização de uma rede empiricamente obtida ou simulada. Aplica-se a função ergm, no caso do pacote statnet, a fim de se obter uma convergência do modelo, de modo que sejam gerados parâmetros para cada termo, e que esses parâmetros sejam obtidos com elevada confiabilidade. O script seguinte apresenta uma sequencia de comandos que ocorre após a sequencia anterior, onde um modelo simulado é agora utilizado como sendo uma rede empírica, por meio da qual se tentará estimar valores para os parâmetros, realizando-se uma operação conceitualmente inversa à simulação.

```
>model.fit <- ergm(rede.simulada ~ edges + gwesp(1) + gwdsp(1))
Starting maximum likelihood estimation via MCMLE:
Iteration 1 of at most 20:
The log-likelihood improved by 0.7781
...
```

Quando se consegue uma convergência no algoritmo MCMLE (*Monte Carlo Maximum Likelihood Estimation*) usado pela função `ergm`, os valores dos parâmetros encontrados são sumarizados, como exemplificado de modo simplificado no quadro abaixo. Para cada parâmetro obtido são indicados o erro padrão e a estimativa do intervalo de confiança dos valores encontrados. Três asteriscos (***) indicam que o grau de confiança da estimativa é 10^-4, ou acima de 99,9%. Esse é usualmente o grau de confiança empregado para aceitação de estimativas para análise em um modelo ERGM.

```
>summary(model.fit)
edges        -4.70191    0.14316       0  <1e-04 ***
gwesp         1.30392    0.11814       0  <1e-04 ***
gwesp.alpha   0.62689    0.10639       0  <1e-04 ***
gwdsp        -0.04782    0.08783       0   0.586
gwdsp.alpha   0.89534    1.37951       0   0.516
```

Os valores estimados indicam que a rede analisada apresenta efeitos significativos ligados aos termos edges e gwesp, sendo que não foi encontrado valor confiável para a configuração gwdsp. Antes de se proceder à análise desses parâmetros deve ser efetuada uma análise de adequação geral do modelo.

3.7. Análise de adequação geral do modelo

Mesmo que parâmetros e configurações selecionadas pelo analista de uma rede empírica tenham sido encontrados com elevado grau de confiabilidade em um modelo estatístico gerado, existe a possibilidade de que o modelo possua efeitos extremos em outras configurações ou métricas de rede não investigadas. Desse modo, antes de se passar à análise dos parâmetros encontrados para um modelo de rede ERGM, deve ser efetuada uma análise de *goodness of fit*, que consiste em avaliar se há um adequado casamento entre as características das redes geradas pelo modelo estatístico e as características presentes na rede empírica, ou vice-versa. Várias métricas, tais como a distância geodésica

entre todas as díades da rede empírica são comparadas com a distribuição de frequência das distâncias geodésicas entre todas as díades de várias redes que pode ser simuladas com os parâmetros encontrados. Nesse caso, por meio da função `gof`, do pacote statnet, são produzidas análises, sumarizadas graficamente como na figura 3.

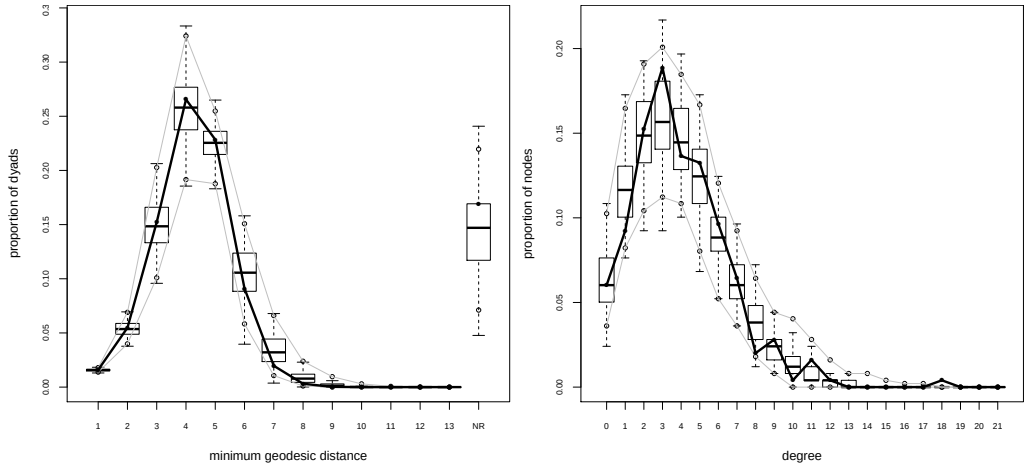

Figura 3. Resultados de análise de adequação geral de modelo (gof - *goodness of fit*) para uma rede empírica.

Os gráficos à esquerda e direita da figura 3 apresentam uma análise de adequação do modelo estatístico gerado na seção 3.6.2, relativamente à rede empírica que foi gerada por simulação e teve suas estatísticas analisadas na seção 3.6.1. O gráfico da esquerda compara os valores efetivos de mínima distância geodésica entre os nós da rede empírica (apresentados em linhas mais grossas) frente ao *box plot* da distribuição de frequência dos valores de mínima distância geodésica presentes em um conjunto de redes simuladas com base nos parâmetros sumarizados na seção 3.6.2. O gráfico da direita apresenta outra análise de adequabilidade, com base na distribuição de frequência de graus da rede empírica, frente às mesmas redes simuladas.

3.8. Interpretação dos parâmetros estimados

A interpretação dos valores obtidos pela modelagem ERGM de uma rede adequadamente modelada depende da compreensão de conceitos de regressão logística. Cabe informar que a introdução dos valores dos parâmetros `edges` e `gwesp` na equação 1 possibilita estimar, respetivamente, a chance de que um laço seja aleatoriamente estabelecido entre dois nós na rede empírica, e a chance de que dois nós que já se relacionam estabelecerem intencionalmente um 1-triângulo com um terceiro nó. Esses valores são de 0,9% e de 79%, respectivamente.

$$chance = e^{param}/(1 + e^{param}) \quad (1)$$

O restante deste artigo apresentará os resultados da análise da rede de colaborações entre os pesquisadores da Ciência da Informação no Brasil.

4. Metodologia de coleta e análise de dados

O universo de pesquisa teve como foco os pesquisadores que faziam parte dos programas de pós-graduação em Ciência da Informação registrados na Plataforma Sucupira da

CAPES, no final do ano de 2016. A partir desses registros, foram buscados os seus respectivos currículos na Plataforma Lattes, por meio da utilização do programa ScriptLattes [Mena-Chalco and Cesar-Jr 2009]. O ScriptLattes busca os dados contidos nos currículos dos pesquisadores e cria a rede de colaboração científica que foi aqui utilizada na aplicação dos estudos descritivos e estatísticos. É importante ressaltar que imprecisões nos dados contidos nos currículos podem resultar em alguns dados incorretos, especialmente devido à diferença de nomenclaturas de coautores ou mesmo no registro das publicações. Uma vez criada a rede de coautoria entre os pesquisadores escolhidos o tratamento dos dados se deu pela incorporação de atributos aos nós da rede. Esses atributos foram coletados manualmente a partir de análise dos respectivos currículos, com incorporação de dados como Unidade da Federação, Universidade, Programas de Pós-graduação e Linhas de Pesquisa às quais estavam vinculados, ano e área de obtenção dos títulos de graduação e doutoramento, o gênero e o foco de suas pesquisas na área da Ciência da Informação, de acordo com os grupos de trabalho definidos pela Associação Nacional de Pesquisa de Pós-Graduação em Ciência da Informação (ANCIB).

Foram realizadas cerca de cinquenta (50) simulações com diferentes combinações de termos ERGM, com durações que variaram de 15 minutos a 8 horas, em um notebook na plataforma Windows 7 com 8 GB de RAM e CPU Intel i5-2467M, a fim de se obter termos que produzissem uma convergência do modelo, cujos parâmetros são apresentados e analisados na seção 5.

5. Resultados e Análise

O melhor modelo para a rede foi o obtido pela fórmula a seguir.

```
ergm.model <- formula(rede ~
          edges + gwesp(1) + nodematch("Universidade") +
          nodematch("AreaDou") + nodematch("Regiao") +
          nodematch("Genero") + absdiff("AnoDou", pow = 1) +
          nodematch("GT"))
```

Os parâmetros estimados são apresentados na tabela 1. Todos foram obtidos com elevado grau de confiança. O valor do erro de cada estimativa, entre parênteses, é menor que metade do valor absoluto da estimativa, permitindo sua utilização em análises. As análises de adequação geral do modelo foram realizadas do modo apresentado na seção 3.7, obtendo-se todos os resultados satisfatórios. A interpretação dos valores obtidos é efetuada a seguir.

Tabela 1. Parâmetros estimados para o modelo ERGM.

Parâmetro	Estimativa (Erro) Confiança
edges	-5.199 (0.103) ***
gwesp	0.775 (0.034) ***
gwesp.alpha	0.958 (0.0264)***
nodematch.Programa	1.253 (0.082) ***
nodematch.Area	0.255 (0.063) ***
nodematch.Regiao	0.279 (0.087) **
nodematch.Genero	0.245 (0.073) ***
absdiff.AnoDou	-0.026 (0.005) ***
nodematch.GT	0.561 (0.078) ***

O parâmetro de valor -5.199 para o termo edges indica ser cerca de 0,5% a chance de que dois pesquisadores quaisquer em Ciência da Informação no Brasil desenvolvam

colaboração entre si. O parâmetro de valor 0.775 para o termo gwesp indica ser de 68% a chance de que dois pesquisadores de ciência da informação no Brasil que já colaboram estabelecerem uma colaboração mútua com um terceiro pesquisador. O valor 0.958 para parâmetro gwesp.alpha, combinado com o valor 0.775 para o termo gwesp, indica ser de 32% a chance de que um 1-triângulo de colaboração entre pesquisadores evolua para um 2-triângulo. Esses mesmos valores indicam ser de 16% a chance de que um 2-triângulo de colaboração evolua para um 3-triângulo.

Os valores dos termos nodematch, aplicados aos atributos Programa, Area, Regiao, Genero e GT (grupo de trabalho na área), apresentam correlações positivas no estabelecimento de laços entre pesquisadores, sendo, entretanto, a vinculação ao programa o fator de maior peso na chance de colaboração entre os pesquisadores, sendo de 77% a chance de que isso ocorra. O valor praticamente zero para o termo absdiff indica que a diferença entre os tempos de doutorado de cada um pouco influencia o estabelecimento de colaborações entre pesquisadores da área.

Alguns dos resultados obtidos com a obtenção do modelo ERGM ajudam a confirmar, com uma maior precisão, informações que já se supunham existir em uma relação de colaboração científica. No caso do coeficiente de clusterização, apresentado por meio do parâmetro gwesp, fica evidente a propensão de um pesquisador colaborar com pares que já tenham relação com colaboradores seus anteriores.

Quanto aos atributos exógenos de Regiao, Area, GT e principalmente Programa, os resultados já podiam ser esperados uma vez que se pode assumir que a proximidade física e a similaridade de assuntos aumentam a propensão em colaborar. No caso dos parâmetros Genero e AnoDou (de obtenção do doutorado) um aprofundamento sobre esses tópicos se faz necessário do ponto de vista dos autores. No último caso, entender as relações entre orientandos e orientadores pode ser um fator relevante no entendimento das colaborações científicas. Quanto ao Gênero, esse tópico tem sido recorrente em grupos de estudo sobre a ciência, e modelos que ajudam a quantificar essas relações podem contribuir bastante para um melhor entendimento dessas relações.

6. Conclusões

Esse artigo apresentou os principais conceitos da classe de modelos estatísticos para análise de redes sociais denominada ERGM. Superadas as dificuldades de compreensão na utilização do modelo, os resultados obtidos geram um grande número de respostas para várias hipóteses sobre a formação de redes sociais como a investigada. Fenômenos de homofilia, propinquidade e seleção social podem ser identificados com facilidade.

Referências

Barabási, A., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3–4):590–614.

Glanzel, W. (2002). Coathorship Patterns and Trends in the Sciences (1980-1998): A Bibliometric Study with Implications for Database Indexing and Search Strategies. *Library Trends*, 50(3):461.

Harris, J. K., Provan, K. G., Johnson, K. J., and Leischow, S. J. (2012). Drawbacks and benefits associated with inter-organizational collaboration along the discovery-development-delivery continuum: a cancer research network case study. *Implementation Science*, 7(1):69.

Katz, J. and Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26(1):1–18.

Kronegger, L., Mali, F., Ferligoj, A., and Doreian, P. (2012). Collaboration structures in Slovenian scientific communities. *Scientometrics*, 90(2):631–647.

Lusher, D., Koskinen, J., and Robins, G., editors (2013). *Exponential Random Graph Models for Social Networks: Theory, methods, and applications*. Structural Analysis in the Social Sciences. Cambridge University Press, USA.

Mena-Chalco, J. P. and Cesar-Jr, R. M. (2009). scriptLattes: An open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, 15(4):31–39.

Morris, M., Handcock, M. S., and Hunter, D. R. (2008). Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects. *Journal of Statistical Software*, 24(4):1–24.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5200–5205.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Special Section: Advances in Exponential Random Graph (p*) Models*, 29(2):173–191.

Sampaio, R. B., Sacerdote, H. C. d. S., Fonseca, B. d. P. F., and Fernandes, J. H. C. (2015). A colaboração científica na pesquisa sobre coautoria: um método baseado na análise de redes. *Perspectivas em Ciência da Informação*, 20(4):79–92.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006). New Specifications for Exponential Random Graph Models. *Sociological Methodology*, 36(1):99–153.

Sonnenwald, D. H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology*, 41(1):643–681.

The Statnet Development Team (2017). Introduction to Exponential-family Random Graph (ERG or p*) modeling with statnet - Version 3.7.1. Technical report, University of Washington, USA.

Wang, P., Robins, G., and Pattison, P. (2009). PNet Program for the Simulation and Estimation of Exponential Random Graph (p*) Models : USER MANUAL. Technical report, Department of Psychology School of Behavioural Science University of Melbourne, Australia.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and applications*. Cambridge University Press, USA.