Redes de títulos de artigos científicos variáveis no tempo

Marcelo do Vale Cunha[1,4], Marcos Grilo Rosa[2,6], Inácio de Sousa Fadigas[2], José Garcia Vivas Miranda[3,6], Hernane Borges de Barros Pereira[1,5,6]

[1]Programa de Modelagem Computacional, SENAI Cimatec, Salvador, BA, Brasil

[2]Universidade Estadual de Feira de Santana, Feira de Santana, BA, Brasil

[3]Universidade Federal da Bahia, Salvador, BA, Brasil

[4]Instituto Federal da Bahia, Salvador, BA, Brasil

[5]Universidade do Estado da Bahia, Salvador, BA, Brasil

[6]Programa de Doutorado Multiinstitucional e Multidisciplinar em Difusão do Conhecimento Universidade Federal da Bahia (Sede), Salvador, BA, Brasil

marcelovale@ifba.edu.br, {fadigas,grilo}@uefs.br, vivas@ufba.br, hbbpereira@gmail.com

Abstract. *In this paper we present a research on the existence of relationship patterns in a scientific community and its trends over time. Networks are formed by cliques from the titles of scientific articles published in the Nature journal. The method has its theoretical basis in Time-Varying Graph (TVG). We consider a window bimonthly (week by week) and advancement in time shows the evolution of networks of words from the titles of papers published. The analysis is carried out using indicators timeless. The results compare the pattern of vertex connections of each network TVG at different times. Networks have the small-world phenomenon.*

Resumo. *Este trabalho investiga a existência de padrões de relacionamento em uma comunidade científica e suas tendências ao longo do tempo. As redes são formadas por cliques a partir dos títulos de artigos científicos publicados no periódico Nature. O método tem sua base teórica em* Time-Varying Graph *(TVG). Nós consideramos uma janela bimensal (semana por semana) e o avanço no tempo mostra a evolução das redes de palavras dos títulos dos artigos publicados. A análise é feita utilizando indicadores atemporais. Os resultados comparam o padrão de conexão de vértices de cada rede do TVG em épocas diferentes. As redes apresentam o fenomeno* small-world.

1. Introdução

Recentemente, alguns trabalhos em redes sociais e complexas têm focado na evolução temporal de redes. Amblard et al. 2011 investiga as relações de coautoria e citações entre autores de artigos científicos. Na biologia, Silva et al. 2012 analisa a evolução temporal de sinais cerebrais em redes de neurônios de ratos, de comportamento livre.

Redes sociais estão definidas formalmente por um conjunto V de Vértices (ou atores da rede), que são amarrados por um ou mais tipos de relações [Wasserman and

Faust 1994]. Este conjunto de vértices, representa objetos reais (e.g. pessoas, instituições, palavras, neurônios, etc.) e denotaremos por $n = |V|$ a cardinalidade do conjunto V. O conjunto $\mathcal{E}$ das relações entre os elementos de V é chamado conjunto de arestas, que denotaremos por $m = |\mathcal{E}|$ a cardinalidade deste conjunto. A estrutura de uma rede social de n atores pode ser modelada por um grafo $G = (V, \mathcal{E})$ e sua análise pode ser feita através de índices estatísticos, que dependem unicamente de informações contidas nos dois conjuntos citados acima.

Estudos mais recentes têm usado esta modelagem com a inclusão de, pelo menos, mais um conjunto dentro do grafo original, com elementos que representam o tempo. Isto torna o grafo variável no tempo. Desta forma, segundo Amblard et al. 2011, os vértices de uma rede dinâmica podem aderir, atrair, competir e até cooperar com outros vértices. Eles podem ainda, desaparecer e até afetar a forma e solidez de seu sistema de relacionamentos.

Um periódico científico pode servir de palco para vários tipos de relações sociais (e.g. coautoria, citações, vocabulário comum, etc.). Por exemplo, pode ser visto como um conjunto de artigos que representam um conhecimento comum de uma comunidade de cientistas, que publica, lê e cita artigos desta mesma comunidade. Desta forma, pode-se modelar uma rede social que tem como relacionamento entre seus atores o vocabulário comum utilizado para compor as publicações dos autores. Todo artigo possui um título e este é composto por palavras selecionadas pelos autores, buscando uma representação fidedigna das ideias que são apresentadas no corpo do trabalho. Através das palavras contidas nesses títulos, pode-se construir redes de palavras a fim de se perceber a relação de um trabalho com outro, de um campo do saber com outro e de um grupo de cientistas com outro.

Qualquer texto escrito pode ser transformado em uma rede de palavras. Caldeira et al. 2006 foram uns dos primeiros autores a considerar as palavras de uma sentença de um texto como vértices de uma clique. De acordo com este raciocínio, a adição de vértices em redes de textos escritos se dá pela adição de cliques. A abordagem estática para redes de títulos foi estudada por por Fadigas et al. 2009 e Pereira et al. 2011. Estes trabalhos propuseram regras para tratamento das palavras e para a formação de redes de títulos.

Este trabalho investiga a evolução temporal de vértices, arestas e índices de um grafo formado por palavras impressas nos títulos de artigos científicos publicados no periódico *Nature*, de 07 de Janeiro de 1999 à 18 de dezembro de 2008. A escolha deveu-se à possibilidade de comparação com trabalhos de mesma natureza, i.e. os dois últimos supracitados. Os vértices são as palavras e as arestas conectam pares de palavras que ocorreram em um mesmo título. Os títulos são agrupados em um documento de texto de formato *.txt*, como um discurso escrito, para assim serem construídas as redes. Após sua construção, uma rede é analisada utilizando índices clássicos de redes e índices da abordagem em cliques, proposto recentemente por Fadigas and Pereira 2013. Os resultados permitem visualizar a dinâmica do adensamento das redes. Esta abordagem contribui para revelar, ao longo do tempo, a importância dos conceitos, expressos nas palavras dos títulos.

Este artigo está organizado da seguinte forma: A seção 2 trata sobre a construção

das redes de títulos, como rede de cliques. A seção 3 trabalha os índices de redes utilizados. Na seção 4 é feita uma descrição da metodologia do estudo. A seção 5 apresenta alguns resultados. Por fim, a seção 6 evidencia as conclusões para esta pesquisa.

2. Construção das Redes

As redes de títulos de artigos científicos deste trabalho são formadas utilizando a abordagem de construção de redes de cliques. Um grupo de vértices é dito como uma clique se todos eles estiverem ligados entre si. A premissa proposta por Caldeira 2005, Teixeira 2007, Aguiar 2009, Fadigas et al. 2009 e Pereira et al. 2011 diz que palavras que ocorrem juntas em uma mesma sentença teriam sido evocadas de forma associativa na construção de uma ideia a ser apresentada. Com isso, pode-se construir uma rede onde as palavras são representadas como os vértices e as arestas são criadas entre pares de palavras que ocorrem em uma mesma sentença em um discurso escrito. A sentença é vista como a menor unidade de significado de um texto e cada palavra pode ter um significado diferente a depender das palavras que estejam ao seu redor.

No contexto deste trabalho, o título de um artigo científico é visto como uma sentença de um discurso escrito e as palavras de cada título como vértices. Consequentemente, a dinâmica de construção de uma rede de títulos é dada por justaposição e/ou sobreposição de cliques. De acordo com Fadigas and Pereira 2013 designa-se de justaposição o processo no qual duas cliques são ligadas por apenas um vértice comum. Quando a ligação ocorre com dois ou mais vértices comuns, chama-se o processo de sobreposição. Observe a construção da rede na Figura 1. Todas as palavras de um mesmo título são interligadas, formando uma clique. Contudo, títulos diferentes podem conter palavras iguais ou palavras de mesma forma canônica. Neste caso, os cliques são unidos, justapondo a palavra comum. Antes de construir as redes, é necessário realizar um

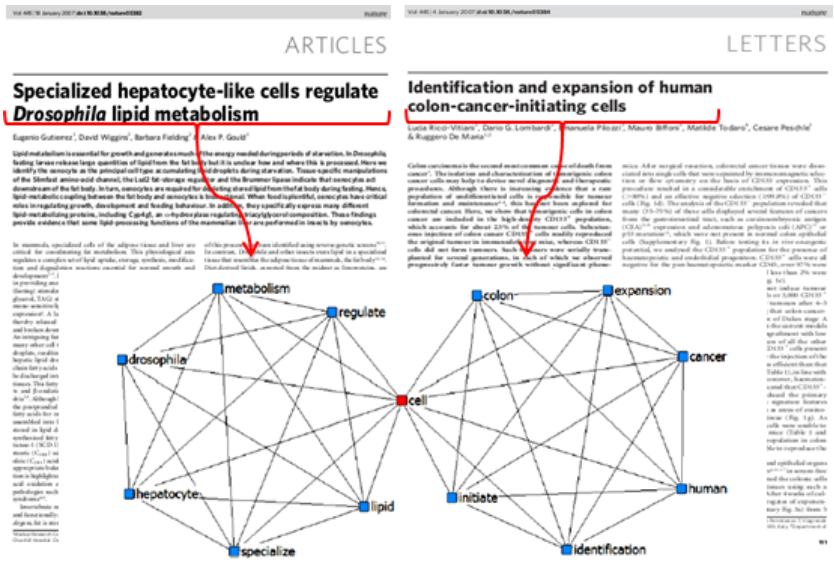

Fig. 1. Primeira página de dois artigos da Nature. Em destaque, seus títulos e o processo de construção da rede. Adaptado de Pereira et al. 2011

pré-tratamento dos títulos. Esta etapa consiste em organizar os títulos em um arquivo de texto, em que cada linha contem um único título. A partir daí modifica-se cada título,

se necessário, de acordo com as regras propostas por Fadigas et al. 2009 e Pereira et al. 2011. Na sequência, o texto é submetido a um tratamento computacional que classifica, modifica e elimina palavras quando necessário. Elas são classificadas de acordo com seu significado léxico (e.g. pronome, advérbio, adjetivo, nome). As palavras gramaticais como artigos e preposições não têm signifcado semântico relevante para a construção das redes e, portanto, são eliminadas. As palavras lexicais, restantes, são mantidas e os verbos reduzidos à sua forma canônica (i.e: initiating $\rightarrow$ initiate). Feito isto, as redes são montadas. Todo este processo é executado com o uso de programas computacionais, em que a base foi o uso do conjunto de softwares do pacote livre UNITEX [1].

3. Indicadores Atemporais

Para grafos variáveis no tempo existem um cojunto de indicadores úteis para analisar sua evolução estrutural. Este conjunto é divido em *índices atemporais* e *índices temporais*[2] [Santoro et al. 2011]. Consta aqui apenas a abordagem de índices atemporais. Esta abordagem está dividida em *Índices Clássicos* e *Índices de Redes de Cliques*.

3.1. Índices Clássicos

A **Distribuição de Graus**, $P(k)$, representa a probabilidade de distribuição de conexões dos vértices da rede. O **Grau Médio** de uma rede, $\langle k \rangle = \sum_{i=1 \epsilon V}^{n} k_i = \frac{2m}{n}$, representa a média dos graus de seus vértices. O **grau** k_i de um vértice i é a quantidade de arestas m_i que incidem nele, ou seja a quantidade de relacionamentos que ele faz.

O **Caminho Mínimo Médio** de uma rede é dado por $\langle \ell \rangle = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij}$ e representa a média de todos os caminhos mínimos ℓ_{ij} entre dois nós i e j da rede. Admite-se que cada aresta ao longo do caminho que ligam dois vértices tem comprimento $\ell = 1$. O **Diâmetro** de uma rede é o máximo valor de ℓ_{ij}.

O **Coeficiente de Aglomeração Médio** de uma rede, $\langle C \rangle = \frac{1}{n} \sum_{i}^{n} C_i$, em que $C_i = \frac{m_i}{\frac{k_i(k_i-1)}{2}}$, é a aglomeração de um vértice i e representa a proporção entre o número de arestas de seus vizinhos m_i e o número máximo de arestas possíveis. Quanto maior for a quantidade de ligações entre os vizinhos de i, maior será a sua aglomeração. Caso um vértice esteja ligado a somente 1 outro vértice, sua aglomeração é 0.

A **Densidade** de uma rede é dada por $\Delta = \frac{m}{\frac{n(n-1)}{2}}$ e representa a razão entre o numero de arestas m de uma rede e o número máximo de arestas possíveis que ela pode ter. Este valor varia de 0, se não existem arestas no grafo, até 1, quando todas as arestas possíveis estão presentes na rede.

3.2. Abordagem baseada em índices para redes de cliques

Para esta pesquisa, também foram usados alguns índices de coesão para redes de cliques, propostos por Fadigas and Pereira 2013. Quando tem-se uma configuração inicial de cliques desconectadas, o número de arestas m_0 é dado pela soma da quantidade

[1] Disponibilizado pela Rede Relex Brasil. Além dele foi utilizado o programa Ambisin e Netpal, desenvolvido por Caldeira 2005

[2] e.g. *Jornada*, *distância temporal*, *excentricidade.*

de arestas em cada clique. Designa-se n_q o número de cliques, q_i o tamanho (número de vértices) da i-ésima clique e n_0 o número total de vértices do estado inicial das cliques isoladas. Fadigas and Pereira 2013 apresenta estruturas de cliques minimamente conectadas: *Estrela, Círculo, Camada e Linha*. Uma rede de cliques minimamente conectada é um conjunto de cliques que forma um único componente.

Para classificar uma rede de cliques real em uma das quatro estruturas teóricas proposta por Fadigas and Pereira 2013, faz-se o uso do *Diâmetro de Referência* normalizado em escala logarítmica (Equação 1):

D^*_{ref}	Estrutura Teórica de Cliques
$0.00 - 0.25$	Layout Estrela
$0.26 - 0.75$	Layout Círculo ou Camada
$0.76 - 1.00$	Layout Linha

$$D^*_{ref} = \frac{ln(D/2)}{ln(n_q/2)} \tag{1}$$

Tab. 1. Classificação de uma rede de cliques através do Diâmetro de referência. Fonte: [Fadigas and Pereira 2013]

Pode-se reescrever os índices *clássicos* (ou *atemporais*) em função dos parâmetros acima citados para redes de cliques reais, considerando todas as justaposições e sobreposições, e o quanto estes valores diferem dos mesmos índices para a mesma rede, só que com cliques desconectadas.

O **Grau Médio** de uma rede de cliques desconectadas é dado por 2:

$$\langle k_{q0} \rangle = \frac{\sum_{i=1}^{n_q} q_i(q_i - 1)}{n_0} \tag{2}$$

Para quantificar a **variação do grau médio** da rede de cliques, comparada com a mesma rede, só que desconectada, é dada por 3:

$$\upsilon(\langle k \rangle) = \frac{\langle k \rangle - \langle k_{q0} \rangle}{\langle k_{q0} \rangle} \tag{3}$$

A **Densidade** de uma rede de cliques desconectada é dada por 4:

$$\Delta_{q0} = \frac{2m}{n_0(n_0 - 1)} = \frac{\sum_{i=1}^{n_q} q_i(q_i - 1)}{n_0(n_0 - 1)} \tag{4}$$

A **variação de densidade** é dada por:

$$\upsilon(\Delta) = \frac{\Delta - \Delta_{q0}}{\Delta_{q0}} \tag{5}$$

De acordo com Fadigas and Pereira 2013, a expressão 5 mede o "adensamento" da rede, em relação ao seu estado inicial (rede de cliques desconectadas).

4. Metodologia, aquisição de dados e TVG

4.1. *Time-Varying Graph*

Em redes sociais tradicionais, a utilização de grafos estáticos (com vértices e arestas fixos), no estudo de problemas de conectividade da rede, consegue representar bem as relações entre os objetos envolvidos na rede e, em geral, caracterizar bem uma comunidade. Todavia, em redes dinâmicas, onde existe uma constante mudança dos objetos que compõem a rede e suas interações, fazem-se necessárias várias representações que não possuam vértices e arestas fixos [Santana 2012].

Nos últimos anos, alguns trabalhos têm apresentado diversas formas de estudar redes variáveis no tempo. Casteigts et al. 2011 teve como objetivo unir e formalizar os diversos conceitos e métricas utilizados no estudo das redes dinâmicas, criando assim o conceito de Time-Varying Graph (TVG). Um TVG pode ser entendido como um grafo estático $G = (V, \mathcal{E})$ acrescido de outros parâmetros que representam funções ou conjuntos temporais: ς (i.e. função de latência), Υ(i.e. função de presença) e Γ(i.e. tempo de vida). Assim um TVG é a quíntupla $\mathcal{G} = (V, \mathcal{E}, \Upsilon, \varsigma, \Gamma)$, onde V e $\mathcal{E}$ representam respectivamente o conjunto de vértices e arestas; a função $\Gamma \subset \mathbb{N}$ representa o tempo de vida do sistema. A função de latência ς indica quanto tempo necessita para que uma aresta esteja disponível em um instante $t \in \Gamma$, em outras palavras, é o tempo necessário para estabelecer o relacionamento entre dois vértices, em um dado instante t. $\Upsilon : \mathcal{E} \times \Gamma \rightarrow \{0, 1\}$ é definido como uma função de presença e garante a existência de uma dada aresta em um dado instante de tempo t.

4.2. Método e Aquisição de dados

Os títulos dos artigos científicos foram coletados do períodico Nature, no período que vai de 1999 até 2008. A revista é publicada semanalmente. Sendo assim, inicialmente, os títulos foram agrupados por semana em 507 arquivos de texto. Posteriormente, para a construção da *janelas temporais*, estes arquivos são agrupados em grupos de 8 arquivos, ou seja 8 semanas. Consideremos a rede de títulos de artigos publicados nos meses de janeiro e fevereiro de 1999. Todos eles compõe a 1ª janela do TVG, ou seja $t = 1$. Da mesma forma, os títulos das publicações que compreendem o mês de Janeiro, exceto a 1ª semana, todo mês de fevereiro e a 1ª semana de março compõe a 2ª janela do TVG, ou seja $t = 2$. Isso se repete até a ultima janela $t = 507$ que corresponde aos títulos das publicações dos meses de Novembro e Dezembro do ano de 2008. Para melhor adaptar nossa proposta ao referencial teórico sobre TVG proposto por Casteigts et al. 2011, foram consideradas as seguintes condições:

- A função de latência ς é constante para todo o TVG. Não faz sentido quantificá-la, portanto este parâmetro não será levado em consideração em nossa análise.
- O tempo de vida do sistema de nossa amostra é o conjunto $\Gamma = \{t_1, t_2, ..., t_i, t_{i+1}, ..., t_{507}\}$. Em que cada t_i corresponde ao intervalo de tempo de $01\ semana$. $|\Gamma| = 507\ semanas$.
- O tempo se inicia na 1ª semana de Janeiro de 1999.
- A função de presença pode ser simplificada e melhor entendida com o uso de uma janela temporal, semelhante à utilizada no trabalho de Silva et al. 2012 e aos *footprints* do trabalho de Santoro et al. 2011. Dessa forma, o TVG em questão será um conjunto de grafos estáticos $\mathcal{G}^{[t_i,t_j]} = (V, \mathcal{E}^{[t_i,t_j]})$. De tal forma que $\forall e \in \mathcal{E}, e \in \mathcal{E}^{[t_i,t_j]} \Leftrightarrow \exists t \in [t_i, t_j], \Upsilon(e, t) = 1$. Ou seja, $\mathcal{G} = \{G^{[t_i,t_j]}, G^{[t_{i+1},t_{j+1}]}, G^{[t_{i+2},t_{j+2}]}, ...\}$
- Cada *janela* temporal de observação será definida por $\tau_i = [t_i, t_{i+7}]$. Deste modo o TVG pode ser escrito em função dessas janelas: $\mathcal{G} = \{\mathcal{G}^{\tau_1}, \mathcal{G}^{\tau_2}, \mathcal{G}^{\tau_3}, ..., \mathcal{G}^{\tau_{507}}\}$.
- A escolha de uma janela de $8\ semanas$ deu-se para suavizar as flutuações dos valores dos índices a cada mudança de janela. Isto, no entanto, não ofusca as tendências dos valores dos índices no TVG como um todo.

5. Resultados e Discussão

5.1. Abordagem baseada em índices Clássicos

A Figura 2 exibe os valores dos indicadores *atemporais* (eixo das ordenadas), em função do tempo, dado em semanas (eixo das abcissas). Cada ponto no gráfico representa a rede de uma janela $\tau_i = [t_i, t_{i+7}]$ de oito semanas de publicações, que se inicia a partir da semana i, correspondente abscissa do ponto. Vale lembrar que a primeira janela do TVG ($\tau_1 = [t_1, t_8]$) se inicia em $t = 1$, 1º de janeiro de 1999.

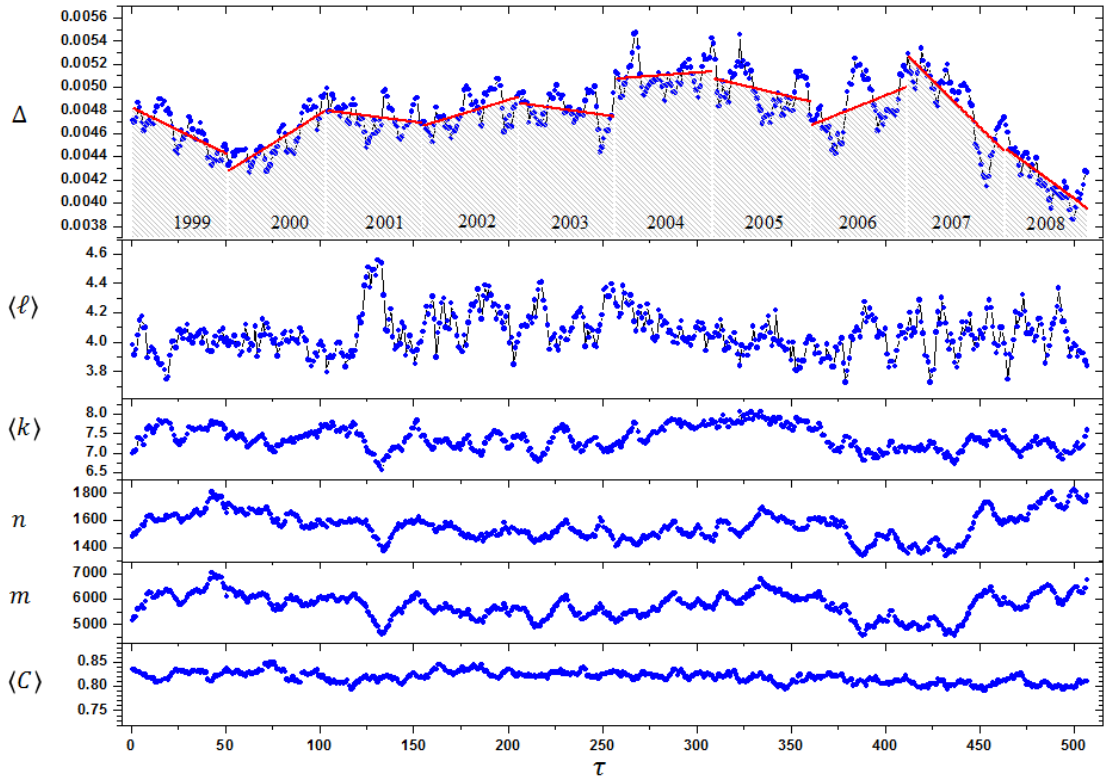

Fig. 2. Evolução dos índices das janelas temporais entre 1999 e 2008. As linhas retas no gráfico de Δ representam o melhor ajuste linear para janelas de 1 ano

A densidade, é relativamente baixa para redes com base em títulos de trabalhos científicos [Fadigas and Pereira 2013]. Entretanto, é possível verificar tendências de crescimento, decrescimento e constância em torno dos valores de densidade ao longo do tempo, nas linhas retas do gráfico da Figura 2. Percebe-se, de acordo com estas inclinações[a], que em média:

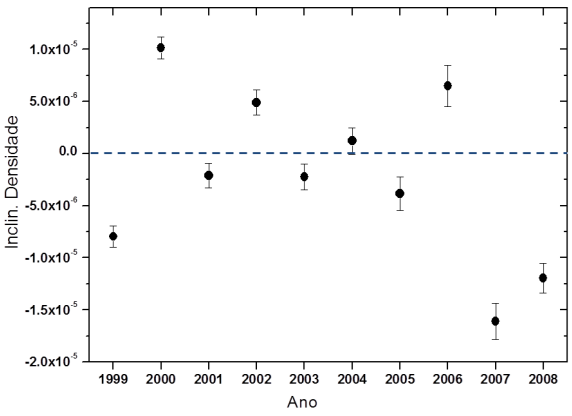

Fig. 3. Valores das inclinações do gráfico de Δ por ano.

[a] Seus valores podem ser consultados no gráfico da Figura 3.

- Em 2007 e 2008 a densidade evolui semelhante à 1999. Ou seja, a rede tende a ser mais esparsa.

- Em 2000 e 2006 ocorre o contrário, a rede tende a ser mais densa.

- De 2001 à 2005 a rede tende a manter a densidade de suas relações entre seus vértices, já que suas inclinações são próximas de zero.

Pode-se reescrever a expressão da densidade da Seção 3, como em função do tempo $t \in \Gamma$ (Eq. 6), assim:

$$\Delta_t = \frac{m_t}{\frac{n_t(n_t-1)}{2}} = \frac{2\langle k \rangle_t}{(n_t-1)} \simeq \frac{2\langle k \rangle_t}{n_t} \qquad (6)$$

Sabe-se que para todos os subgrafos do TVG em questão $n \gg 1$, por isso foi usado a aproximação $n_t - 1 \simeq n_t$. De acordo com a Equação 6, para épocas em que as redes tiveram o mesmo valor de grau médio $\langle k \rangle_t$, a densidade é apenas efeito do tamanho da rede. Neste caso, quanto maior o número de vértices n_t menor será o valor da densidade Δ_t.

Para exemplificar, considere as janelas τ_{257}, τ_{372} e τ_{419}. Como se pode ver no gráfico da Figura 2 e na Tabela 3, as redes para estes instantes possuem o mesmo valor de grau médio, ou seja, para estes três momentos da história de publicação da Nature, os "relacionamentos" entre os vértices (o grau médio) em média se manteve, mas o "poder de relacionamento" das redes (a densidade) inicialmente diminuiu e em seguida aumentou. Isto se deve ao fato de que inicialmente a diversidade de palavras da janela (número de vértices) aumentou e em seguida diminuiu. De maneira análoga podemos comparar a janela τ_{18} com a janela τ_{445}. Percebe-se que o poder de relacionamento destas redes é mesmo. Houve uma redução no vocabulário e no "relacionamento médio" entre as palavras dos títulos (Tabela 3).

O comportamento dos gráficos de vértices e arestas no tempo (Figura 2) são similares, quase que sobrepostos, respeitando as diferenças entre escalas. Isto sugere que, em algumas épocas, n e m são em média proporcionais, ou seja:

$$m_t \propto n_t \qquad (7)$$

Este resultado nos mostra que em vários períodos de tempo, maiores que τ_i, a proporção em entre vértices e arestas, ou seja, o grau médio $\langle k \rangle_t$ das redes com o passar do tempo, em média, se mantém constante. Este mesmo resultado pode ser entendido observando o gráfico da Figura 4.

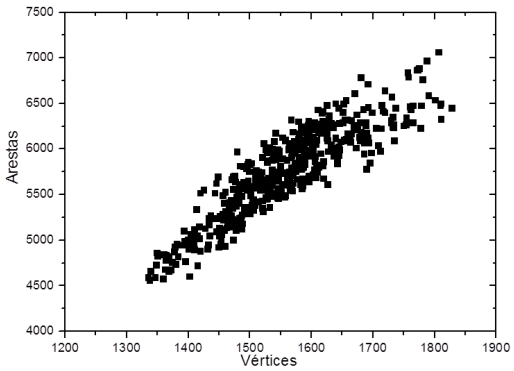

Fig. 4. O número de vértices cresce em média proporcionalmente ao número de arestas

Neste caso, pode-se rearrumar a Equação 6 da densidade para:

$$\Delta_t \cong \frac{2\langle \overline{k} \rangle_{\tau_k}}{n_t} \qquad (8)$$

Ou seja, já que o grau médio das redes das janelas variaram muito pouco, para alguns períodos de tempo $\tau_k = [t_k, t_z] > \tau_i$, a densidade Δ_t das janelas τ_i é, em média, função exclusiva do numero de vértices da rede de cada janela. Sendo $(t_k, t_z) \in \Gamma$.

Em se tratando do TVG como um todo, ou seja para um período de tempo τ de 507 semanas, cabe as seguintes perguntas: Cada índice possui um valor médio bem definido? Ou seja, o padrão de conexões das palavras dos títulos da revista Nature, ao longo desses 10 anos (exibidos no gráfico da Figura 2), é apenas flutuação de um valor médio (para cada índice)? Existe correlação temporal para estas flutuações? Para responder à estas perguntas, foi realizado mais dois procedimentos: Teste de Normalidade de Shapiro and Wilk 1965 e aplicação do Método DFA de Peng et al. 1994. A Tabela 2 mostra o resultado destes testes.

Teste de Normalidade	Δ	$\langle k \rangle$	n	m	D	$\langle C \rangle$	$\langle \ell \rangle$
W	0.99	0.98	0.99	0.99	0.88	0.99	0.98
$p-valor$	0.00006	0.00003	0.00387	0.00020	0.00000	0.00499	0.00000
Método DFA	Δ	$\langle k \rangle$	n	m	D	$\langle C \rangle$	$\langle \ell \rangle$
$\mathcal{H}$	0.661	0.640	0.783	0.767	-	0.550	0.499
$erro$	0.013	0.006	0.007	0.006	-	0.008	0.016
R^2(ajuste)	0.994	0.999	0.999	0.999		0.998	0.985

Tab. 2. Teste de Normalidade: Valores da Estatística de Shapiro-Wilk, com o nível de confiança de $\alpha = 0.05$. Método DFA: Expoente de Hurst $\mathcal{H}$ das séries temporais dos índices de redes, erros associados e R^2 dos ajustes

Como se pode ver na tabela todos os (p-$valores$)$<\alpha$. Sendo assim, este teste rejeita a hipótese de normalidade nas distribuições de frequências de todos os indicadores. Portanto a resposta para as duas primeiras perguntas é que a média dos valores de cada índice não é representativa, pois não segue uma distribuição normal. Resta agora saber se existe correlação temporal para os índices. A análise do expoente de Hurst na Tabela 2 e no gráfico da Figura 5 nos permite identificar que, com exceção do *Diâmetro* (não foi identificada auto-afinidade em sua série temporal), todos os indices de redes têm um padrão auto-afim, no intervalo de 4 a 21 semanas, em suas séries.

O único índice que possui inclinação $\mathcal{H} \simeq 0.5$ é o *caminho mínimo médio* $\langle \ell \rangle$. Isto significa que, para o intervalo de tempo considerado, apesar de correlacionado, sua série temporal não possui "memória" e segue uma caminhada aleatória. Entretanto, as outras quantidades possuem memória para o intervalo de tempo de 4 a 21 semanas, com destaque para n e m, que possuem as correlações mais altas, seguidas de $\langle k \rangle$. Os valores de $\mathcal{H} > 0.5$ representam correlações de longo alcance persistentes, ou seja, uma tendência positiva no passado é mais provável de continuar positiva e vice-versa. Caso as séries fornecessem $\mathcal{H} < 0.5$ os dados teriam uma correlação anti-persistente.

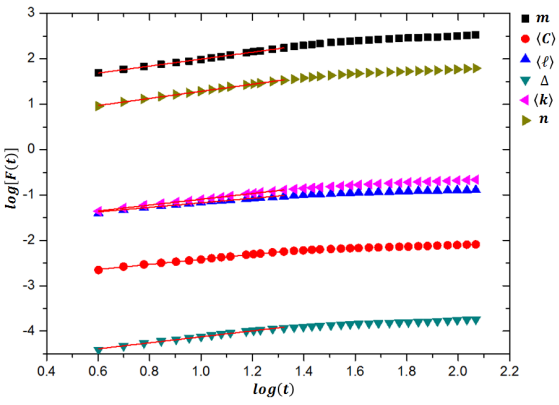

Fig. 5. Valores do logaritmo da função F_{DFA} em função do logaritmo do tempo, dado em semanas.

5.2. Índices para Abordagem de Redes de Cliques

A Tabela 3 mostra os valores dos índices[3] baseados na abordagem de redes de cliques para algumas janelas do TVG e o confronto deles com alguns índices clássicos.

Semana	*18*	*25*	*106*	*249*	*257*	*268*	*339*	*372*	*419*	*445*	*499*
Δ	0.0049	0.0044	0.0048	0.0044	0.0051	0.0055	0.0047	0.0045	0.0053	0.0049	0.0039
Δ_{q0}	0.0020	0.0021	0.0021	0.0022	0.0027	0.0026	0.0021	0.0022	0.0024	0.0023	0.0017
$\upsilon(\Delta)$	1.4541	1.1060	1.2963	1.0277	0.8809	1.1390	1.2238	1.0303	1.2280	1.1429	1.3159
$\langle k \rangle$	7.82	7.24	7.65	7.18	7.20	7.76	7.75	7.20	7.20	7.42	6.98
$\langle k_{q0} \rangle$	5.11	5.05	5.14	5.09	5.31	5.38	5.26	5.17	4.94	5.16	4.68
$\upsilon(\langle k \rangle)$	0.53	0.43	0.49	0.41	0.36	0.44	0.47	0.39	0.46	0.44	0.49
n_0	2558	2416	2466	2349	1958	2095	2489	2332	2061	2278	2810
n	1606	1637	1598	1619	1405	1420	1652	1617	1350	1523	1811
n_q	481	459	480	458	368	391	477	465	429	460	594
n_0/n	1.59	1.48	1.54	1.45	1.39	1.48	1.51	1.44	1.53	1.49	1.55
n_0/n_q	5.3	5.3	5.1	5.1	5.3	5.4	5.2	5.0	4.8	5.0	4.7
D	10	10	8	10	11	11	11	11	10	14	10
D_{ref}	0.29	0.30	0.25	0.30	0.33	0.32	0.31	0.31	0.30	0.36	0.28
$\langle C \rangle$	0.82	0.83	0.81	0.82	0.83	0.82	0.82	0.82	0.80	0.81	0.80
$\langle C \rangle_{rd}$	0.006	0.005	0.004	0.003	0.005	0.006	0.005	0.005	0.005	0.006	0.004
$\langle \ell \rangle$	3.81	4.08	3.91	4.11	4.26	4.08	4.06	3.83	4.13	4.16	3.95
$\langle \ell \rangle_{rd}$	3.99	3.99	3.84	3.98	3.91	3.73	3.85	3.95	3.89	3.82	4.05
$\%Cp_{maior}$	90.9%	90.3%	91.3%	89.9%	88.6%	94.3%	92.5%	85.4%	90.9%	91.4%	89.1%

Tab. 3. Índices de redes complexas e índices de rede de cliques para algumas janelas do TVG.

A análise de seus valores, permite observar que:

- A linha que apresenta (n_0/n_q) mostra que, em média, aproximadamente 50% dos vértices das janelas apresentadas constituem palavras que se repetem.
- Para todo o TVG $D_{ref445} \leq D_{ref} \leq D_{ref106}$. Ou seja os extremos possuem o maior e o menor[4] D_{ref}, respectivamente.
- A janela τ_{257}, apesar de ter um dos maiores valores de Δ, possui um valor pequeno para $\upsilon(\Delta)$ e para $\upsilon(\langle k \rangle)$. Ou seja, em média, o relacionamento dos vértices e o poder de relacionamento do estado inicial das cliques isoladas já eram, relativamente, altos antes das cliques se juntarem.
- As janelas τ_{372} e τ_{419} possuem o mesmo valor de $\langle k \rangle$. O aumento de Δ, de uma janela para outra, se deu pela redução de n. Isto reforça o que já foi discutido: Para este TVG, o aumento no vocabulário dessas janelas de tempo, na maioria das vezes, torna a rede mais esparsa, já que as palavras, em média, relacionam-se da mesma forma.

[3] A maior funcionalidade destes índices é ver o quão uma rede de cliques real difere de sua configuração inicial (i.e. estado inicial das cliques isoladas)

[4] Caso a rede de cliques fosse minimamente conectada, teria um layout tipo *Estrela*

- A janela τ_{499} possui o menor valor de Δ das redes que formam o TVG. Janelas próximas à ela representam a tendência mais atual[5] da revista, i.e. um vocabulário cada vez maior a medida que o tempo passa.

Segundo Watts and Strogatz 1998, uma rede apresenta o efeito small world se $\langle C \rangle \gg \langle C \rangle_{rd}$ e se $\langle \ell \rangle$ é comparável com $\langle \ell \rangle_{rd}$. Nesta definição, $\langle C \rangle_{rd}$ é o coeficiente de aglomeração médio para uma rede aleatória com mesmo grau médio $\langle k \rangle$ e mesmo número de vértices n. Analogamente, $\langle \ell \rangle_{rd}$ é o caminho mínimo médio para a rede aleatória correspondente. Logo, de acordo com os resultados da Tabela 3, podemos inferir que todas[6] as redes exibem o fenômeno small world.

6. Considerações finais

A abordagem TVG é ideal para se perceber tendências no comportamento do relacionamento das palavras. O teste de Normalidade aplicado pôde constatar que não existem valores médios para os índices, de forma a representar todo o TVG. O método DFA foi útil para perceber que existe uma forte correlação no número de vértices e arestas no intervalo $4 \leq \tau \leq 21$. Isto significa, por exemplo, que para um dado tempo $t \in \Gamma$, se o vocabulário da Nature aumentou nos últimos 2 meses, então existe uma alta probabilidade dele continuar aumentando a partir do próximo mês, e essa tendência se mantém fortemente correlacionada até aproximadamente 4 meses depois.

A abordagem de rede de cliques é adequada para este estudo. Seus resultados podem mostrar o quanto mudam os indicadores a partir da junção das cliques, uma vez em uma configuração inicial isoladas. Os títulos de artigos são apenas grupos de palavras isolados que interagem entre si e representam bem a ideia dos estudos dos pesquisadores que trabalharam em cada artigo. Após publicados por uma revista e analisados sob a ótica de Redes, percebe-se a influência de palavras comuns na junção de ideias. Os altos coeficientes de aglomeração e a constatação do fenômeno small-world nos leva a supor que é alta a probabilidade de que duas palavras ligadas a uma outra, estejam, elas mesmas, ligadas entre si. Estudos futuros sobre a frequência dessas palavras poderão revelar algo sobre a capacidade delas interagirem com grupos de temáticas diferentes, que podem indicar áreas do conhecimento humano.

Cabe-nos indicar que os comportamentos observados se repetem em outros periódicos (e.g. Science). Entretanto, a limitação do número de páginas impede-nos de ampliar a discussão.

Referências

[Aguiar 2009] Aguiar, M. S. (2009). Redes de palavras em textos escritos: Uma análise da linguagem verbal utilizando redes complexas. Programa de pós-graduação em físca, Universidade Federal da Bahia, Salvador.

[Amblard et al. 2011] Amblard, F., Casteigts, A., Flocchini, P., Quattrociocchi, W., and Santoro, N. (2011). On the temporal analysis of scientific network evolution. In *CASoN*, pages 169–174. IEEE.

[5]Entende-se por "mais atual", nessa base de dados, as ultimas janelas deste TVG, ou seja, publicações no final do ano de 2008

[6]Para isso as redes precisam estar conectadas ou que tenham um componente com a maioria dos vértices da rede. Na tabela, $\%Cp_{Maior}$ representa a porcentagem do maior componente da rede

[Caldeira 2005] Caldeira, S. (2005). Caracterização da rede de signos linguísticos: Um modelo baseado no aparelho psíquico de freud. Mestrado interdisciplinar em modelagem computacional, Fundação Visconde de Cairu, Salvador.

[Caldeira et al. 2006] Caldeira, S. M. G., Petit Lobão, T. C., Andrade, R. F. S., Neme, A., and Miranda, J. G. V. (2006). The network of concepts in written texts. *The European Physical Journal B*, 49(4):523–529.

[Casteigts et al. 2011] Casteigts, A., Flocchini, P., Quattrociocchi, W., and Santoro, N. (2011). Time-varying graphs and dynamic networks. In Frey, H., Li, X., and Rührup, S., editors, *ADHOC-NOW*, volume 6811 of *Lecture Notes in Computer Science*, pages 346–359. Springer.

[Fadigas et al. 2009] Fadigas, I., Henrique, T., Pereira, H., Senna, V., and Moret, M. (2009). Análise de redes semânticas baseada em títulos de artigos de periódicos científicos:o caso dos periódicos de divulgação em educação matemática. *Educação Matemática Pesquisa*, 11(1):167–193.

[Fadigas and Pereira 2013] Fadigas, I. and Pereira, H. (2013). A network approach based on cliques. *Physica A: Statistical Mechanics and its Applications*, 392(10):2576 – 2587.

[Peng et al. 1994] Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., and Goldberger, A. L. (1994). Mosaic organization of dna nucleotides. *Phys. Rev. E*, 49(2):1685–1689.

[Pereira et al. 2011] Pereira, H., Fadigas, I., Senna, V., and Moret, M. (2011). Semantic networks based on titles of scientific papers. *Physica A: Statistical Mechanics and its Applications*, 390(6):1192 – 1197.

[Santana 2012] Santana, A. (2012). Caracterização da jornada máxima em redes dinâmicas. Programa de pós-graduação em matemática, Universidade Federal da Bahia, Salvador.

[Santoro et al. 2011] Santoro, N., Quattrociocchi, W., Flocchini, P., Casteigts, A., and Amblard, F. (2011). Time-varying graphs and social network analysis: Temporal indicators and metrics. *CoRR*, abs/1102.0629.

[Shapiro and Wilk 1965] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

[Silva et al. 2012] Silva, B., Miranda, J., Corso, G., Copelli, M., Vasconcelos, N., Ribeiro, S., and Andrade, R. (2012). Statistical characterization of an ensemble of functional neural networks. *European Physical Journal B*, 392:85–358.

[Teixeira 2007] Teixeira, G. M. (2007). Redes semânticas baseadas em discursos orais: Uma proposta metodológica baseada na psicologia cognitiva utilizando redes complexas. Mestrado interdisciplinar em modelagem computacional, Fundação Visconde de Cairu, Salvador.

[Wasserman and Faust 1994] Wasserman, S. and Faust, K. (1994). *Social Network Analsis*. Cambridge: Cambridge University Press.

[Watts and Strogatz 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of'small-world'networks. *Nature*, 393(6684):409–10.