Redes Complexas - Motifs na Wikipédia

Gustavo Vrech, Luiz Gonzaga

Instituto de Física de São Carlos, USP, 13566-590, Brazil

20 de Junho de 2011

1 Introdução

Um motif é um pequeno elemento de um todo, que pode ser analisado para dar informações sobre todo um conjunto. Em música, um motif, ou motivo é um pequeno trecho de melodia que permite a um ouvinte identificar qual música está sendo reproduzida escutando apenas estas poucas notas. No caso das redes complexas, um motif é uma sub-rede que de algum modo é representativa para a rede, um pequeno elemento básico como um tijolo, que de algum modo caracteriza toda a construção.

Um motif é definido como uma sub-rede de uma rede maior, um subconjunto de qualquer tamanho. Tipicamente procuram-se os motifs de pequeno tamanho, aproximadamente de três a cinco nós ou arestas. Qualquer rede poderia ser subdividida em todos os seus motifs, e assim feita o levantamento estatístico sobre sua composição tal qual um espectro de freqüências, porém as redes que possuem motifis mais significativos, aqueles que aparecem muito mais do que os outros evidenciam o real potencial prático dos motifs.

Neste trabalho mostraremos como o conceito de motifs pode ser aplicado para caracterizar redes complexas, mais especificamente aplicaremos o conceito para caracterizar uma rede formada pela estrtura de páginas da wikipedia, onde escolhemos doi temas que a principio acreditamos sejam completamente diferentes. Tais redes reflete as relações autores-páginas e páginas-páginas existentes na wikipedia, assim sera possivel identificar uma rede em relação a outra apenas olhando para os motifs.

2 Reprodução do Experimento

Para reproduzir o experimento apresentado no artigo, foi necessário eleger dois nichos de redes dentro da Wikipédia. Uma vez que o artigo original compara as redes de jogadores de futebol (cujo resultado salientou que um autor contribui para várias páginas) com as redes de sociólogos (salientando desta vez uma topologia onde vários autores se juntam para construir uma única rede) foram procuradas redes que preservassem essa característica. Uma vez que seria utilizado para a aquisição das redes a Wikipédia em inglês, foram escolhidos como fontes para as redes de autoria os artigos referentes as redes complexas, sendo que estes era esperado um grande número de coautores dentro de uma página, e como segunda fonte páginas das cidades da região de São Carlos, onde era

esperado que um único autor contribuísse com diversas páginas.

De forma muito semelhante ao artigo original, foram procurados motifs de dois a cinco nós, sendo possíveis dois tipos de nós, os de autor e os de página. Um nó autor seria ligado a um nó página caso o autor tivesse contribuído àquela página, e dois nós páginas seriam ligados caso existisse hiperligação entre elas. Não era possível ligação entre dois nós autores. Uma vez que estávamos interessados na relação autoria/página destas estruturas, foram desprezados motifs que não continham autor algum. Tais restrições, tal qual ao artigo original, resultou em 134 motifs diferentes.

O estudo sobre a reprodução do trabalho pode ser resumido em três passos: Aquisição das redes complexas a partir da Wikipédia; Aquisição dos motifs presentes a partir de um algoritmo sampling ou amostragem; e a discussão dos resultados obtidos. O algoritmo foi implementado no ambiente de programação livre scilab (http://www.scilab.org/).

3 Aquisição das redes complexas a partir da Wikipédia

Podemos, resumidamente, definir o objetivo do nosso trabalho como um estudo de caracterização de redes, no contexto de redes complexas. Onde a partir de duas redes distintas extraímos características que evidenciam uma determinada rede em relação a outra. A técnica explorada neste trabalho, como já foi discutido anteriormente, utilliza-se do fato de que pequenos padrões, que ocorrem com alta frequência numa rede, podem ser utilizados para evidenciar esta rede em relção a uma outra rede qualquer. Estes pequenos padrões são conhecidos, por quem estuda o assunto, como *motifs*.

Devemos portanto defenir nossas redes, ou seja, escolher duas redes distintas na qual estudaremos as características que as evidenciam, com esse objetivo escolhemos as redes que ocorrem dentro da estrtura de páginas da wikipédia.

3.1 Utilização do software PageAnalyser

Para obter a estrutura da Wikipédia, e com objetivo de polpar esforços em assuntos que não acrescentariam grandes conhecimento relevantes ao assunto em estudo (Redes Complexas), fizemos uso de um software, já pronto, desenvolvido em linguagem ruby. Este software encontra- se disponível na página do desenvolvedor (https://github.com/ChrisSalij/PageAnalyzer). O pacote de código como um todo, cai sobre versão GNU General Public License version. A razão para isto é que o software Weka está incluído neste software e é licenciado sob GPL. A fim de redistribuir o código Weka este programa como um todo deve ser licenciado sob a GPL também.

Este programa requer, para uma execução correta, que a máquina primeiramente esteja com todos os pacotes e bibliotecas necessárias para a execução de um programa, escrito em linguagem ruby, assim como os Germs Curb e HappyMapper, onde o primeiro é necessário para o código de raspagem (um sub-grupo do software que não inclui o software Weka) e o segundo é necessário para inúmeras funções.

Algumas configurações foram necessárias, para que as informações extraídas da página fossem focadas ao tema em questão (páginas relacionadas as cidades da região de São Calos e páginas relacionadas a artigos que tratam de Redes Complexas). As principais configurações a serem feitas encontram-se no diretório de configurações. Tais informações são sobre login para o boot na

wikipédia que se encontra no arquivo bot-login.r (figura 1), se tiver registrado, número de variáveis

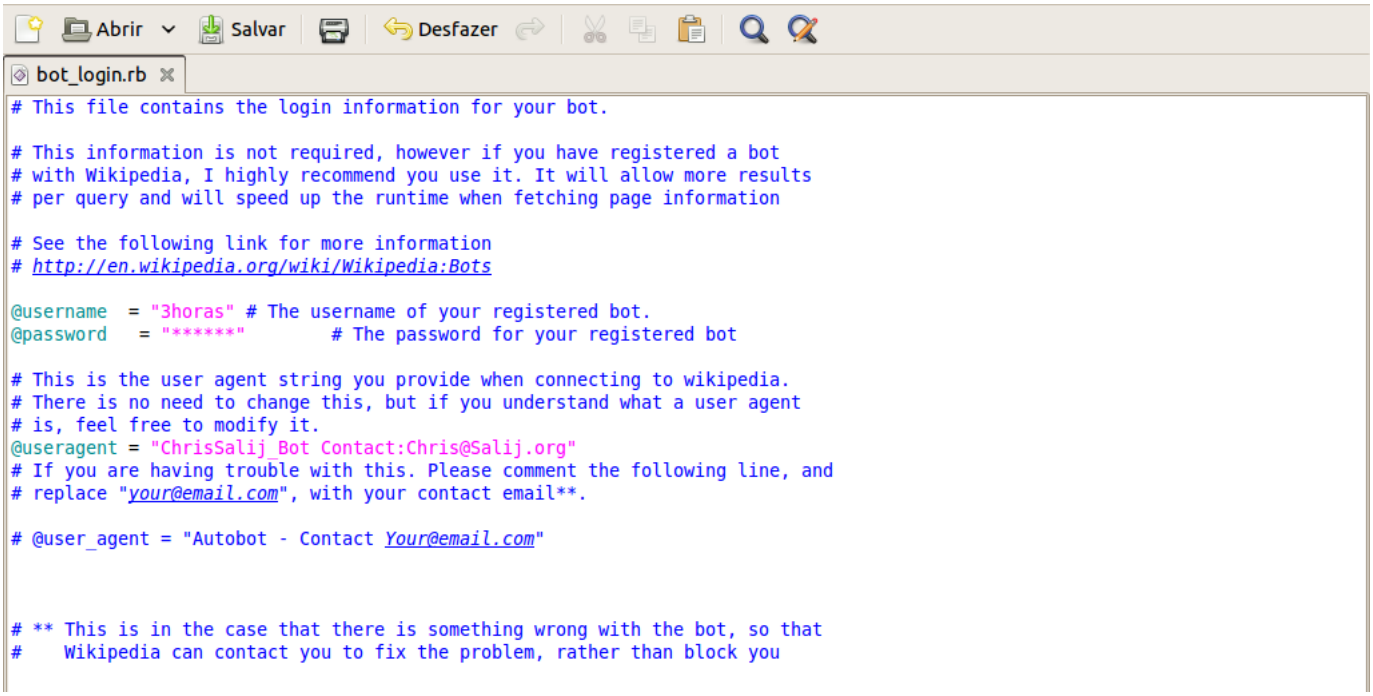

Figura 1: Arquivo bot-login.r

que controlam a quantidade de dados coletados e onde eles serão armazenados que se encontram no arquivo config.rb (figura 2)

e as páginas de onde as informações serão extraídas que tambem se encontram pages-list.txt (figura 3)

3.2 Tratamento dos Dados

Após ter baixado todos os programas bem como seus pacotes e as bibliotecas necessárias, deve-se interpretar os resultados que se obtem após executar o programa, tais resultas encontram-se espalhados por dois diretórios o primeiro deles é o downloaded-data que contém diversos arquivos, um para cada página indicada no arquivo pages-list.txt. Estes arquivos contém informações das páginas que referenciam a página em questão assim como seus id (número que identifica cada página dentro da wikipédia). O segundo diretório mais importante é o parsed-data que também contém diversos arquivos, um para cada página indicada no arquivo pages-list.txt, alem de arquivos contendo informações de usuários que contribuem para aquelas página da lista. Um exemplo dem um arquivo estraído é dado na figura 4

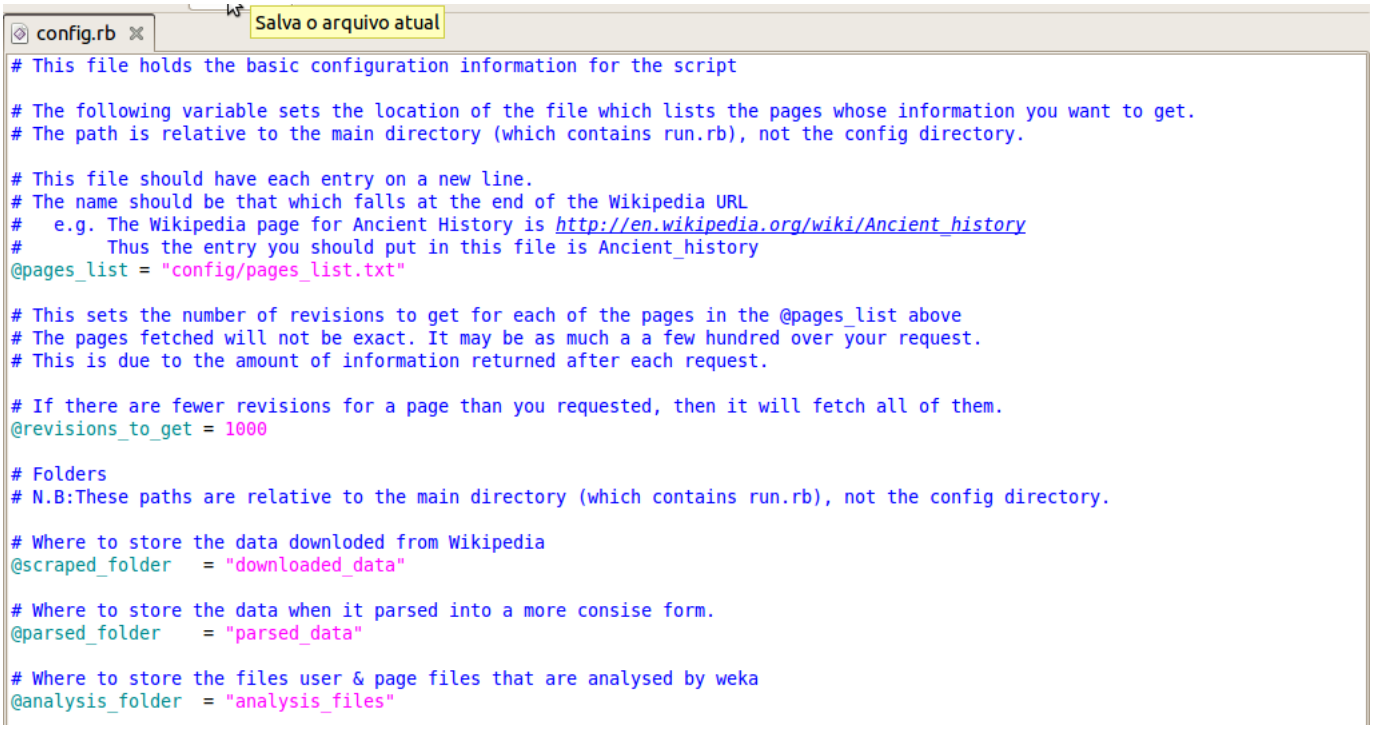

Figura 2: Arquivo config.rb

Figura 3: Arquivo pages-list.txt

A figura 4 mostra um arquivo referente a página da cidade de Americana. Podemos observar que temos um arquivo com estrutura *xml*, portanto para que as informações sejam extraídas e utilizadas para formar a rede é preciso de um programa que possa tratar uma estrutura xml. Uma linguagem bastante utilizada, na qual tambem foi utilizada no trabalho, para tratar tais estrutura é a linguagem php. Com um pequeno programa escrito em php podemos extrair as informações contidas nos documentos xml.

O arquivo mostrado na figura 4, que se encontra no diretório downloaded-data, contém os links que referênciam a cidade de Americana, junto com os links encontra-se o id da página e um número (ns) que indentifica que tipo de link está refenciando a página, se este número for zero significa

Figura 4: Arquivo contendo informações de Americana

que se trata de uma outra página caso este número for 6 o tipo de link é uma imagem, em geral *.png.

Portanto vemos que há diversos tipos de links, mas nós estamos interessados em olhar apenas um tipo bem específico de link, cujo estes links são outras páginas. Isto para nós é interessante pois o objetivo do trabalho é obter uma rede complexa, onde há dois tipos de nós na rede, os nós de autores e os nós de páginas. Entre estes nós há relações que denotaremos, em nossa rede, por uma aresta que liga dois distintos nós, existindo portanto dois tipos de ligações: ligações autor-página e ligações página-página. Observe que não há ligações entre autor-autor pois este tipo de estrtura não ocorre nas páginas wikipédia.

Precisamos agora obter informações dos autores que trabalham nas páginas estudadas, estas informações encontra-se dentro do diretório parsed-data onde há dois tipos de arquivos, um deles contem as informações de todos autores de contribuem para uma específica página. O arquivo referente a cidade de Americana é mostrado na figura 5.

Vemos apartir da figura 5 que o arquivo contém o nome do usuário e o número da revisão feita por ele, dentro da página cidade de Americana. Dois detalhes foram levados em conta ao extrair as informações de tais arquivos, nesta lista de autores percebemos que haviam autores se repetindo, pois a lista refere-se as revisões feita na págia e os autores que as fizeram e não uma lista específica de autores. Isto porque cada autor pode ter realizado mais de uma revisão dentro daquele artigo. Assim o programa, escrito em linguagem php, levou este detalhe em conta evitando assim redundâncias.

O outro detalhe crucial, para obtenção da rede, é que nem todos os autores listados no arquivo

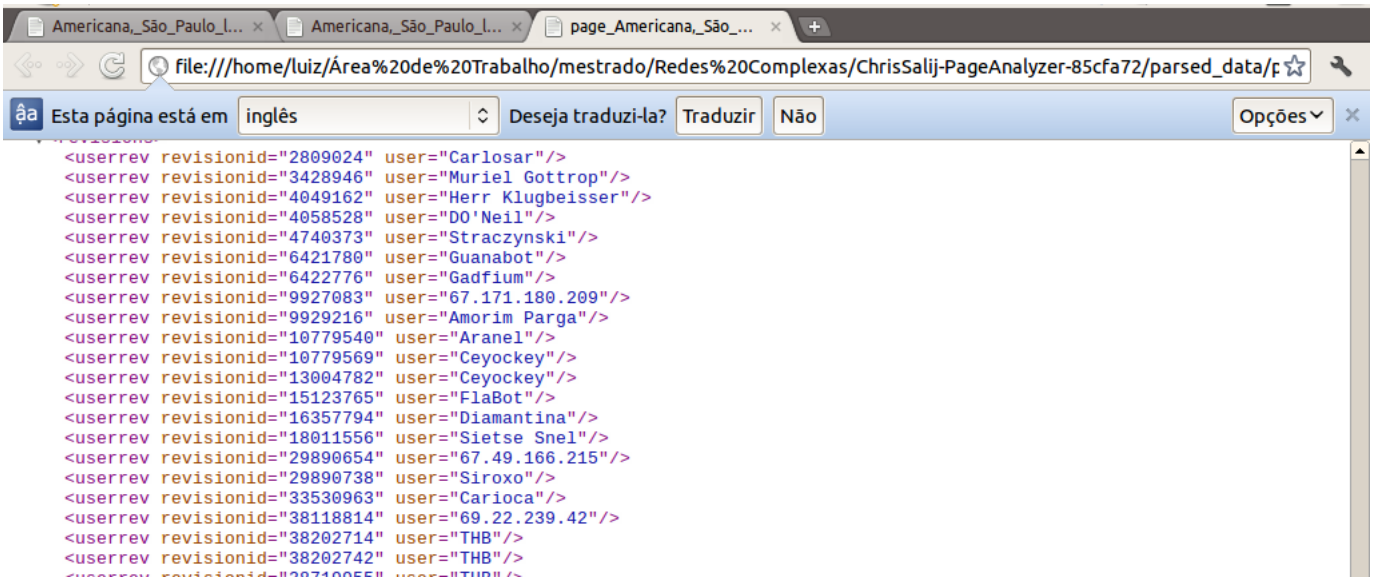

Figura 5: Arquivo contendo informações das revisões

acima são realmente autores que contribuem para a página, alguns deles são usuários virtuais que existem para monitorar as revisões feita nas páginas. Portanto não devemos levar em conta, na contrução da rede, tais usuários. O problema seria facilmente resolvido se houvesse uma informação dizendo se o autor relamente contribui na construção da página ou se é apenas um usuário "root". Mas felizmente esta informação encotra-se implícita no mesmo diretório. Como já havia dito, o diretório parsed-data contém uma lista com todos os arquivos de usuários que contribuem para as páginas estudadas. Estes arquivos, identificados pelo nome do usuário, contém todas informações necessárias para nosso propósito, assim através deles foi possivel identificar os usuários "root" e portanto excluí-los de nossa análises.

A figura 6 mostra um arquivo de usuário:

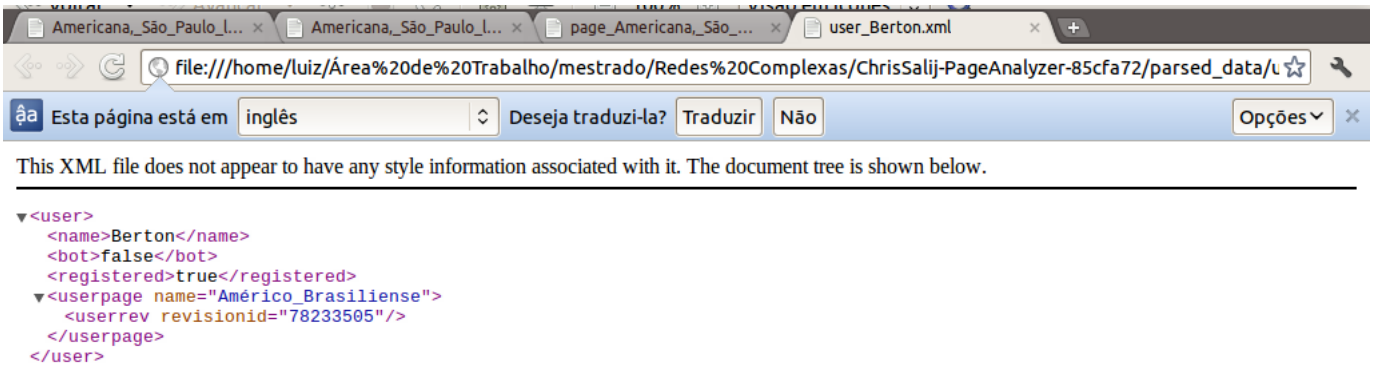

Figura 6: Arquivo contendo informações dos autores

Como vemos pela figura 6 o arquivo gerado possui o nome do autor e o nome da página que ele revisou, assim como o número da revisão. Mas a informação que nos mais interessa é o que está dentra das tags “bot” isto é um identificador, ele é quem determina se o usuário em questão é “root”, ou seja, de inspeção ou se ele reamente é usuário real que contriu na construção da página. Quando o valor for False quer dizer que o autor contribui, caso contrário o autor é considerado virtual e nã deve ser levado em consideração.

3.3 Construção da Rede

Após extrair as informações necessárias para o estudo, devemos transformá-las numa forma que possam ser tratadas no contexto de redes complexas. Para isso existem diversas maneiras de representar redes complexas. Neste trabalho utilizamos para representá-las uma estrutura bem conhecida na área de computação, que é chamada de Matriz de Adjacências. Neste tipo de estrutura colunas e linhas representam os nós (Autores e Páginas) e os valores dos elementos de matriz representam as ligações, sendo que o valor *zero* representa a ausência de ligação e o valor *um* representa a acorrência de uma ligação. Todo o tratamento dos dados, gerado pelo PageAnalyser, assim como a construção da matriz de Adjacências foram feitas com auxílio de um programa escrito em linguage php, uma peque na parte do pragrama é apresentado na figura 7:

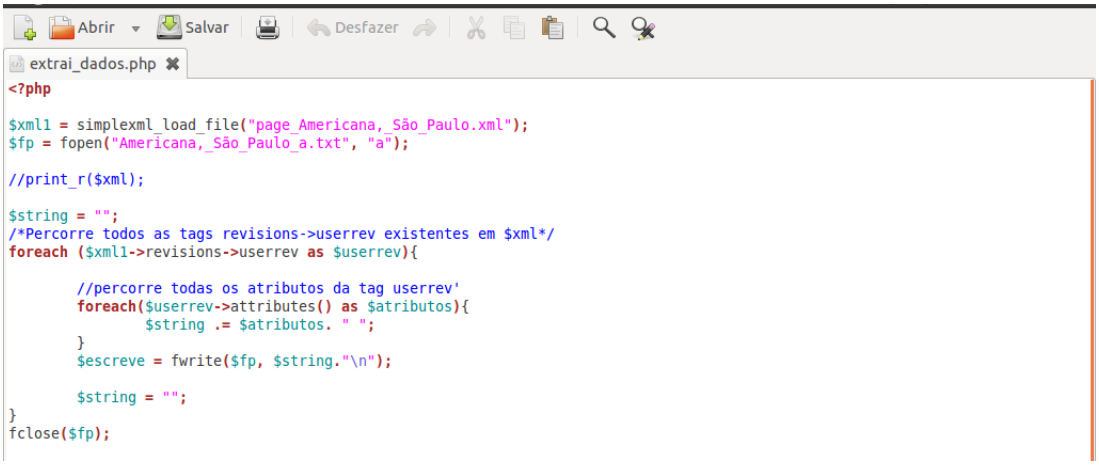

Figura 7: Programa em linguagem PHP

A função do trecho de código apresentado é extrair informações que encontram-se dentro das tags XML, no presente caso o trecho de código extrai informações da página da cidade de Americana, onde retiramos as informações que nos interessa da estrutura XML (neste autores que contribuiram para a página de Americana) e reescrevemos estas informações num segundo arquivo .txt que poderá ser tratado por qualquer outra linguagem (por exemplo C ou java).

4 Algoritmo utilizado

O algoritmo utilizado no presente trabalho para detectar os motifs presentes na rede da Wikipédia é denominado algoritmo de sampling ou amostragem, e é descrito em mais detalhes em (bioinformatics). Sua idéia básica, tal qual métodos de monte-carlo, é baseado no sorteio de nós aleatórios

das redes, assim como o sorteio de uma subestrutura a qual este nó é presente. Tal subestrutura é um motif, e repetindo-se diversas vezes tal algoritmo é possível determinar a distribuição estatística dos motifs da rede.

Conforme evidenciado no organograma apresentado na figura (Fig-organograma), o algoritmo tem cinco passos básicos, sendo que estes serão descritos com mais detalhes nas seções subseqüentes: Sortear um nó; sortear um vizinho; Guardar configuração; Verificar Isomórfico e atualizar dados.

4.1 Sortear nó

Ponto de partida do algoritmo é escolhido aleatoriamente um dos nós presentes da rede. Tal nó será o ponto inicial de detecção do motif neste ponto da iteração.

As redes são tratadas como suas matrizes de adjacências, sendo estas o parâmetro de entrada para o algoritmo. De acordo com as restrições apresentadas no trabalho inicial, todas as redes devem necessariamente ter um nó de autor, fazendo assim com que todos os nós iniciais sejam nós de autor, impossibilitando desta forma que qualquer motif sem um nó autor seja detectado. Tal sorteio é feito trivialmente no ambiente computacional utilizando, utilizando apenas a lista de nós autor gerados inicialmente.

Diversos artigos defendem heuristicamente a adição de um fator multiplicativo ao realizar a amostragem, corrigindo diversos efeitos que ocorrem ao eleger aleatóriamente um nó para a amostragem. Tal fator foi ignorado no presente trabalho devido às características intrínsecas do método utilizado: estamos interessados em comparar diretamente os motifs de duas redes distintas, porém de tamanho semelhantes, fazendo com que quaisquer fatores multiplicativos sejam aplicados às duas redes, gerando apenas encargo computacional extra sem revelar nenhum tipo novo de informação. Na seção de discussão dos resultados será explorada tal característica mais afundo.

4.2 Sortear vizinho

Escolhido um nó inicial, o algoritmo procurará por todos os nós que interagem com este, sejam este nós que estejam conectados ou se conectam a este (No caso de redes direcionadas). Isso é feito verificando-se na matriz de adjacência as linhas referenciadas por aquele nó – evidenciando os nós o qual o nó inicial é ligado, assim como as colunas – que representam quais nós que eventualmente estão conectados àquele nó.

A partir dos vizinhos do nó inicial é escolhido aleatoriamente um novo nó. A priori o nó escolhido pelo algoritmo é sempre um nó autor, fazendo com que o primeiro passo escolha um nó página para fazer parte da amostragem do subconjunto. Os nós conectados ao segundo nó são então adicionados ao conjunto de nós conectados ao primeiro nó - este novo conjunto será o conjunto de eventuais amostragem que o algoritmo consultará ao fazer sucessivas iterações.

4.3 Guardar configuração

Os nós escolhidos aleatoriamente pelo passo anterior são armazenados, até que o laço seja repetido k-1 vezes, resultando num subconjunto da rede de tamanho k. Tal subconjunto será um motif de tamanho k. Para armazenar o subconjunto foi eleito um método que transforma o subconjunto de tamanho k numa sub-matriz de adjacência de lado k. Este tratamento foi escolhido devido às características computacionais do passo seguinte, a verificação de isomorfismos.

4.4 Verificar se é Isomórfico

Antes de atualizar a base de dados com a amostragem de motif feita anteriormente, é necessário saber de o motif amostrado não é isomórfico a outro já amostrado.

Figura 8: (Organograma do algoritmo utilizado para detecção de motifs) - Inicialmente sorteia-se um nó da rede, e em seguida sorteia-se k vizinhos relacionados, onde k é o tamanho dos motifs que estão sendo detectados. Atualiza-se então a estatística relacionada a este subconjunto de tamanho k na base de dados. Tal procedimento é repetido n-1 vezes, sendo este parâmetro n dependente do tamanho da rede analisada

Uma rede é isomófica a outra se ambas são equivalentes, isto é, se para cada nó da primeira rede, assim como para cada ligação, exista equivalência total na segunda rede, preservando todas as características (como número de nós, topologia das ligações e eventualmente peso e direção das arestas). É comumente denominada bijeção com preservação de arestas.

Inicialmente utilizou-se um algoritmo de força-bruta que trocava iterativamente as linhas e colunas da matriz de adjacência do motif tentando reduzi-lo de uma forma conveniente. Porém a presença de dois tipos de nós (os nós autor e página) fez com que tal algoritmo tivesse um tempo de execução demasiadamente longo. Optou-se então por importar uma função do software Matlab para o ambiente de programação utilizado, função esta que utilizava métodos mais avançados para detectar isomorfismos, sendo que estes fogem do escopo deste trabalho.

A função utilizada tinha como entrada a matriz de adjacência do motif, e retornava uma nova matriz de adjacência numa forma canônica, de tal forma que quaisquer redes isomórficas fornecidas na entrada retornariam como a mesma matriz canônica.

4.5 Atualizar base de dados

A partir da matriz de adjacência do motif calculado anteriormente é feita a estatística de quantas vezes aquele motif em específico ocorreu na rede.

São amostrados um número n de motifs para caracterizar a rede, sendo que este número n deverá ser ao menos cinco vezes maior do que o número de nós presente na rede.

Do ponto de vista computacional, é necessária uma estrutura inomogênea para manter a estatística dos motifs apresentados, na presente implementação foi utilizado o recurso list() do scilab, onde eram armazenadas todas as matrizes de adjacência, assim como a sua freqüência de aparecimento dentro da rede.

5 Discusão dos Resultados

As redes amostradas foram então submetidas ao algoritmo apresentado, resultando assim nas freqüências de aparecimento de cada motif possível nas duas redes. Utilizando a mesma metodologia apresentada no artigo original, foram calculadas as freqüências relativas de cada um dos motifs da rede, sendo S_i^w o i-ésimo motif detectado da rede w . A partir desta calculou-se para todos os motifs i a grandeza:

$$\log\left(\frac{S_i^{cidade}}{S_i^{complexa}}\right) \tag{1}$$

Os valores em módulo desta grandeza foram ordenados e o valor original foi graficado na figura 9. Note que tal método faz com que, caso um motif seja muito mais freqüente na rede de cidades do que na rede formado pelas redes complexas, a grandeza será bastante positiva, enquanto que caso a freqüência seja maior na rede das redes complexas, a grandeza terá valor fortemente negativo.

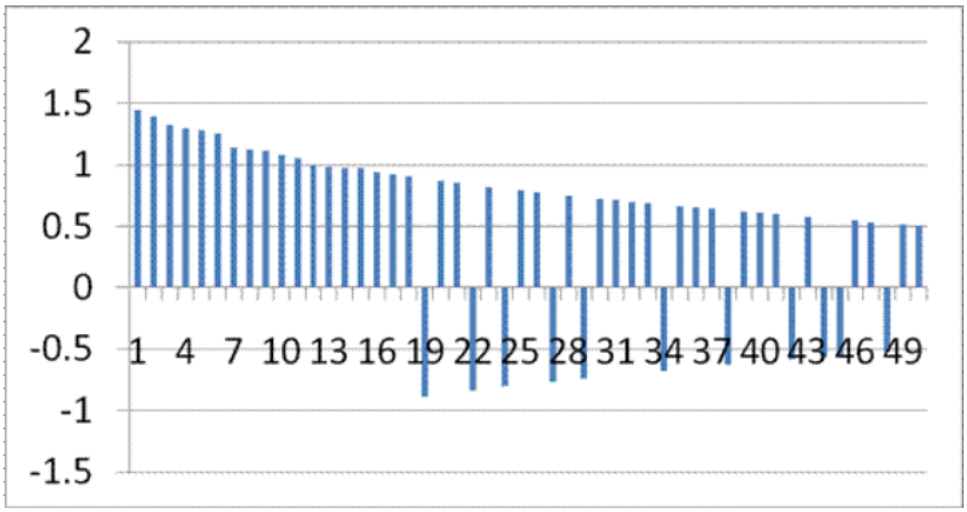

Figura 9: $\log\left(\frac{S_i^{cidade}}{S_i^{complexa}}\right)$ vs. i

Os três motifs de valor mais positivo, e assim maior freqüência na rede de cidades do que na rede de redes complexas podem ser vistos na figura 10. Na figura 11 são exibidos os motifs de valor mais negativo, ou seja, aqueles que aparecem com maior freqüência na rede formada pelas páginas das redes complexas na Wikipédia. Nesta representação os círculos azuis representam autores, enquanto que os quadrados amarelos representam os nós das páginas.

Os resultados obtidos foram coerentes com o que se esperava inicialmente, pode-se observar que no caso das redes formadas pelas cidades um único autor participa da edição de diversas páginas –

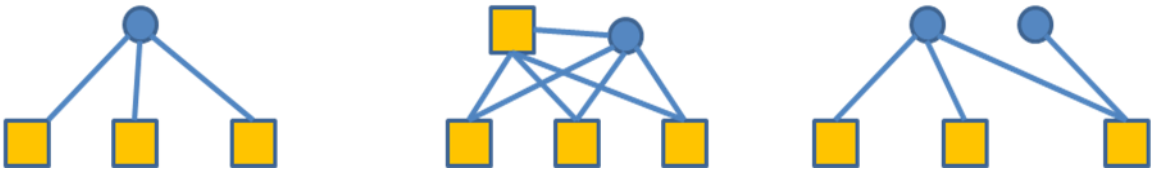

Figura 10: Os três motifs mais expressivos na rede de cidades

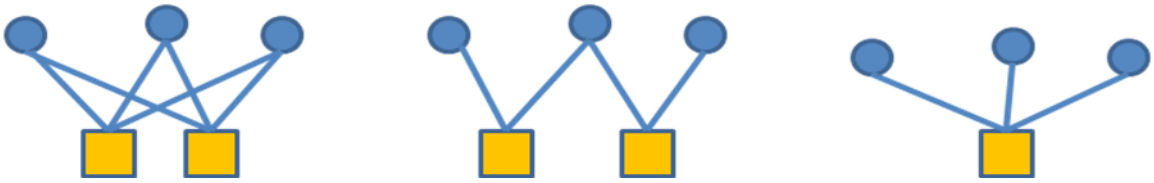

Figura 11: Os três motifs mais expressivos na rede de páginas sobre Redes Complexas

provavelmente um morador desta região fluente em língua inglesa que por alguma razão interessou-se em manter tais páginas. Já no caso da rede de páginas formadas por informações sobre as redes complexas, nota-se que em geral vários autores contribuem para a formação de uma única página, como trata-se de informações científicas é natural que um usuário complemente informações inicialmente tratadas pelo usuário anterior.

Um resultado particular interessante é o segundo motif da figura 10(a). Neste motif é possível observar um nó página ligado com todos os nós páginas do motif, e com um autor. Este nó página de alta conectividade representa uma cidade grande da região, que faz nesta rede um papel semelhante a um hub, é muito provável que páginas de cidades pequenas citem acidade maior como referencia.

É importante salientar a forma como estes resultados foram obtidos, fazendo-se o logaritmo da razão entre as freqüências relativas de dois diferentes motifs. Diferentes trabalhos ao se lidar com motifs utilizam técnicas refinadas para analisar os resultados, como a presença de fatores multiplicativos e diferentes tipos de medidas. Aqui porém foi utilizado apenas o método de amostragem simples – note que nem ao menos foram agrupados motifs de tamanhos diferentes, é feita a comparação inclusive entre motifs de tamanhos diferentes.

Tal metodologia resulta em resultados errôneos na maior parte das vezes, porém aqui é feita a comparação entre duas redes, sendo assim desnecessário. O arcabouço estatístico necessário para estudar uma única rede se faz necessário para evitar que um determinado motif tenha peso exagerado na rede, e em geral é feita comparação com uma rede aleatória com uma medida denominada significância. Porém ao se comparar duas redes todo esse raciocínio se faz desnecessário, uma vez que o gráficoe videncia apenas as diferenças entre as duas redes. Tal gráfico não mostra qual motif é mais presente um uma ou outra rede, nem ao menos a freqüência relatica de aparecimento de um único motif, porém mostra a comparação entre as freqüências relativas das duas redes, dando importância às diferenças, desta forma um motif com alta freqüência de aparecimentos nas duas redes (supostamente importante na topologia de ambas) não terá valor apreciável no gráfico, por que este motf não caracteriza uma importante diferenciação entre as duas redes.

6 Conclusões

Motifs são subredes presentes dentro de redes maiores, que de alguma forma caracterizam esta rede. O presente trabalho procurou reproduzir o método de comparação entre duas redes de acordo com motifs característicos. No artigo original os autores compararam duas redes de informação

formadas pelos autores e páginas da Wikipédia (http://www.wikipedia.org/), a rede de páginas referentes a jogadores de futebol inglês e sobre sociólogos.

O presente trabalho optou por comparar redes formadas pelas páginas da Wikipédia em inglês sobre as cidades da região de São Carlos e as páginas referentes às redes complexas. A metodologia utilizada foi semelhante ao trabalho original, inicialmente foi necessário extrair as redes da Wikipédia, implementar um algoritmo que analisasse tais redes e retorna-se os motifs que as compõe, e por fim analisar os resultados obtidos.

Para extair as redes foram utilizados um software pronto para extrair a estrutura de páginas da Wikipédia, PageAnalyser, tal software encontra-se disponível na página do desenvolvedor, e um segundo software, desenvolvido pelo grupo em PHP na qual trata as arquivos XML que o PageAnalyser produz. Com isso podemos tratar o assunto no contexto de Redes Complexas, formando com os dados uma Matriz de Adjacências, na qual foi utilizada para extração dos *motifs*.

O algoritmo utilizado para extrair os motifs é conhecido como algoritmo de *sampling* ou amostragem consiste em escolher aleatóriamente um nó da rede, e a partir deste escolher nós conectados a este de tal modo a formar um sub- conjunto ou motif. Ao realizar este procedimento diversas vezes a rede será caracterizada pela requencia relativa de aparecimento de cada um dos motifs. É importante salientar que deve existir preocupação para se lidar com sub-redes isomórficas.

Por fim, o método proposto utiliza o logaritmo da razão entre as freqüências de aparecimento de um determinado motifs nas duas redes. Tal método permite diversas simplificações estatísticas ao lidar com os motifs, uma vez que não é necessário nenhuma preocupação quanto a significância do motif, ou suas medidas estatísticas: Apenas a razão do aparecimento de um motif para as duas redes.

O método utilizado salienta as diferenças topológicas das duas redes em termos dos motifs. Tal método mostrou-se eficiente do ponto de vista de diferenciação, levando em consideração seu simples implemento em comparação aos métodos tradicionais ao se lidar com motifs. Como pode ser visto na figura 9 o método de fato diferencia duas redes, no caso das redes formadas pelas cidades pode-se claramente notar que um único autor participa da elaboração de várias páginas, enquanto que no caso das redes complexas vários autores tendem a contribuir com o mesmo material.

Também fica evidente a discondancia que existe na área a respeito dos métodos que são utilizados. Diversos artigos utilizam a mesma idéia básica, porém a literatura da área apresenta uma série de outras soluções ad hoc de acordo com o problema tratado, como por exemplo o presente trabalho.

Referências

[1] NEWMAN, M. E. J. *The structure and function of complex networks*. SIAM Reviews, v. 45, n. 2, p. 167 – 256, 2003.

[2] N. Kashtan, S. Itzkovitz, R. Milo and U. Alon *Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs*, Vol. 20 no. 11 2004, pages 1746–1758 doi:10.1093/bioinformatics/bth163

[3] Site do PageAnalyser *http://github.com/ChrisSalij/PageAnalyzer*

[4] Tutorial sobre PHP *http://www.php.net/*

[5] Site da Wikipedia *http://pt.wikipedia.org/wiki/Wiki*

[6] Guangyu Wu, Martin Harrigan, Pádraig Cunningham: *A Characterization of Wikipedia Content Based on Motifs in the Edit Graph*. School of Computer Science & Informatics, University College Dublin, Technical REport UCD-CSI-2011-02.

[7] N. Kashtan, S. Itzkovitz, R. Milo and U. Alon *Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs*, Vol. 20 no. 11 2004, pages 1746–1758 doi:10.1093/bioinformatics/bth163