Bruno Leite Alves

Um Estudo sobre a Evolução Temporal de Redes Sociais

Proposta de dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Alberto H. F. Laender

Belo Horizonte
Maio de 2012

Sumário

1. Introdução

A interação entre os indivíduos vem crescendo ao longo dos anos. Cada vez mais enxergamos tais interações como redes, ou seja, pessoas ou coisas interligadas com um propósito. Exemplos de redes incluem redes sociais, redes de computadores ou redes biológicas. Essas redes de interações podem ser definidas como redes complexas, sendo possível serem modeladas utilizando uma poderosa ferramenta conhecida como grafos. Grafos são conjuntos de vértices conectados por arestas. Em redes complexas, podemos representar pessoas ou coisas através de vértices e a relação que os une pode ser representada como arestas. Redes sociais são um tipo de redes complexas, em que as pessoas podem ser representadas através de vértices e seus relacionamentos através de arestas.

Redes sociais são compostas por pessoas que se relacionam, podendo tal relacionamento ser por meio direto, por algum tipo de interesse em comum ou através de comunidades. Nos últimos anos as redes sociais ganharam força na Internet, este crescimento permitiu que vários serviços se destacassem, por exemplo, Wikipédia[1], Flickr[2], Facebook[3], Twitter[4], Linkedin[5], dentre outros. Tais serviços têm se tornado cada vez mais populares nos últimos anos, isto acontece porque as pessoas estão compartilhando mais informações, pessoais e profissionais, na Web. Por exemplo, amantes de filmes podem ir ao cinema ou realizar compras baseados em recomendações da IMBD[6] ou Netflix[7]. O Facebook pode conectar pessoas em comunidades que compartilham um mesmo interesse. O Flickr é uma plataforma Web que permite que os usuários compartilharem suas fotos favoritas com outros usuários. O Linkedin é uma grande rede coorporativa que permite que os usuários divulguem seus currículos profissionais e também que as empresas realizem divulgação de vagas e seleções de candidatos. As pessoas também podem obter e compartilhar conhecimentos através da plataforma Wikipédia. Quando as pessoas se juntam dessa forma para compartilharem conhecimentos ou experiências na Web, as plataformas e serviços usados acabam sendo beneficiados pela massa de que lá circulam. Desta forma, existe a possibilidade de capturar tais informações e a cada dia esta tarefa tem se tornado mais importante.

Além das redes sociais citadas, temos também sistemas biológicos e de informação que também podem ser descritos como redes, onde os nodos representam indivíduos e as arestas ou links representam a relação ou interação entre os nodos. Grandes esforços têm sido desprendidos para entender a evolução dessas redes [1, 2], a relação entre as topologias e funções [3,4] e a caracterização destas redes [5].

Redes de coautoria são formadas por pesquisadores que publicam trabalhos em fóruns científicos. Podemos modelar essas redes de coautoria como cada nodo correspondendo a um pesquisador na rede e uma aresta entre dois nodos indicando que os pesquisadores publicaram pelo menos um trabalho em conjunto. Os dados mantidos pelas bibliotecas digitais DBLP[8] e BDBComp[9], por exemplo, possibilitam este tipo de modelagem.

---

[1] http://www.wikipedia.org/
[2] http://www.flickr.com/
[3] http://www.facebook.com
[4] http://www.twitter.com/
[5] http://www.linkedin.com/
[6] http://www.imdb.com/
[7] http://www.netflix.com/
[8] http://www.informatik.uni-trier.de/~ley/db/
[9] http://www.lbd.dcc.ufmg.br/bdbcomp/

2. Motivação

Uma característica observada nas redes sociais é que elas evoluem através dos anos, independente do seu tipo. Por exemplo, no passado o MySpace[10] teve um crescimento exponencial no número de usuários, porém, em 2008 sofreu uma grande perda de usuários devido ao aumento de usuários do Facebook [6].

A Figura 1 mostra o número de publicações realizadas na DBLP ao longo dos anos. No caso da DBLP podemos observar que o número de publicações está sempre crescendo, uma vez que as informações das publicações são apenas adicionadas e nunca excluídas.

Estudos realizados na área de redes complexas, muitas vezes se baseiam em *snapshots* das redes, não considerando a sua evolução [3, 8, 9]. As características observadas nessas redes podem ser afetadas pelo seu crescimento ou declínio, por exemplo, se estudos tivessem sido realizados no MySpace considerando os dados antes de 2008, a rede teria métricas diferentes das obtidas após a migração dos usuários para o Facebook.

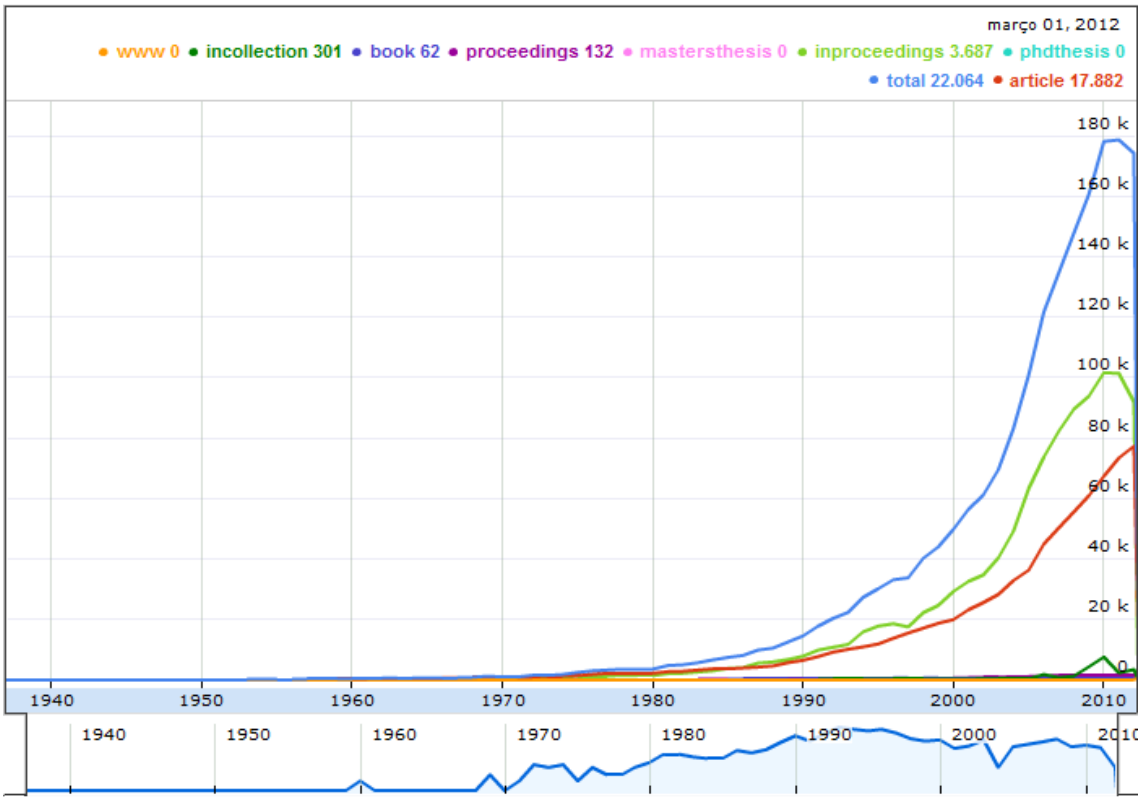

Figura 1. Evolução das publicações da DBLP através dos anos
Fonte: http://dblps.uni-trier.de/~mwagner/statistics/Publicationsperyear.html

3. Objetivo

Neste trabalho iremos realizar um estudo sobre a evolução temporal de redes complexas, a fim de verificar a correlação entre as características encontradas nas janelas temporais e na rede como um todo.

Considerando a questão temporal, este trabalho poderá identificar também como os links são formados ao longo do tempo, ou até mesmo o motivo deles serem desfeitos. O problema de predizer a criação de links entre nodos de uma rede é denominado na literatura como predição de links [7, 9].

[10] http://www.myspace.com/

4. Cenários de estudo

Este tipo de estudo é aplicável a redes com características temporais. Em especial, iremos inicialmente utilizar os dados da DBLP, SBBD (Simpósio Brasileiro de Bancos de Dados) e Twitter, devido à facilidade de obtenção dos dados e às características de cada rede.

4.1. DBLP

A DBLP é uma biblioteca digital que possui mais de 1,9 milhões de publicações. A biblioteca disponibiliza seus dados no formato XML, facilitando desta forma o seu uso no estudo. Utilizando os dados da DBLP é possível criar uma rede de coautoria em que os nodos correspondem aos autores e as arestas entre eles correspondem às colaborações de trabalhos.

A DBLP armazena várias informações sobre os pesquisadores e seus trabalhos publicados. Como a DBLP mantém a informação da data de publicação dos artigos, é possível obter o estado da rede em um dado instante, viabilizando desta forma, o seu uso para este estudo.

4.2. Twitter

O Twitter é uma rede social para *microblogging*, que permite que seus usuários realizem a postagem de pequenas mensagens e acompanhem também as mensagens de outros usuários. O Twitter permite também que os usuários se conectem através do sistema de seguidos e seguidores, dando origem a um grafo direcionado. Cada mensagem publicada na rede é conhecida como *tweet*. Os *tweets* podem ser propagados para outros usuários, sendo este termo conhecido dentro da blogosfera como *retweet*.

Os *tweets* possuem informações de quando foram publicados. Desta forma, podemos modelar a rede baseada nos *tweets* dos usuários e assim obter estados temporais dela.

4.3. SBBD

O SBBD, que em 2010 completou 25 anos de existência, é o maior e mais importante evento da América Latina na área de bancos de dados. Os trabalhos publicados nas edições do SBBD foram coletados e disponibilizados por [14]. Essa coleção possui informações sobre os artigos publicados, tais como autores, ano de publicação, dentre outras.

5. Fundamentação teórica

5.1. Redes complexas

As redes complexas permitem o estudo de vários fenômenos ou sistemas encontrados na natureza. Esses sistemas apresentam características que nos permitem modelá-los como um grafo. Grafos são compostos por vértices e arestas que representam ligações entre os elementos do sistema. Redes sociais ou redes de coautoria podem ser modeladas dessa forma, de modo que os vértices representem as pessoas e as arestas representem a interação entre eles, seja através de amizades ou coautorias em trabalhos.

Redes complexas vêm sendo usadas para realizar estudos em várias áreas, como redes biológicas e até mesmo a própria Web. Vários estudos compararam redes sociais utilizando técnicas de redes complexas [7, 10, 11, 12, 13]. Trabalhos de identificação de agrupamentos nas redes também vêm sendo realizados com o intuito de identificar

como os nodos se relacionam e como eles se agrupam em forma de comunidades [15, 16].

5.1.1. Métricas para o estudo de redes complexas

Métricas baseadas na topologia da rede podem ser usadas para identificar características da rede. A seguir descrevemos algumas métricas [12].

5.1.1.1. Grau dos vértices

Esta é uma característica importante na estrutura da rede e segue uma lei de potência. Sendo assim, a probabilidade de um nodo ter grau k é proporcional a $k^{-\alpha}$. Através de uma regressão linear, podemos obter o expoente α, comumente utilizado para comparar redes. Em grafos direcionados, é comum analisar o grau dos vértices levando em consideração as arestas de entrada e de saída.

5.1.1.2. Coeficiente de agrupamento

Este coeficiente é um indicador de conectividade do nodo. Como o próprio nome já diz, ele informa o quão agrupado um nodo se encontra da rede. Em outras palavras, é a razão entre o número de arestas que conectam um nodo *i* a seus vizinhos e o número máximo de arestas entre estes vizinhos.
O coeficiente de agrupamento representa a densidade de arestas que conectam o nodo *i* e seus vizinhos. Para calcular o coeficiente de agrupamento de uma rede é necessário calcular a média do coeficiente de agrupamento de todos os nodos.

5.1.1.3. Distância média e diâmetro

É possível calcular a distância média de um grafo, calculando a média de arestas em todos os caminhos mínimos existentes entre todos os pares dos nodos. Geralmente esta medida é calculada no maior componente conectado do grafo, uma vez que o grafo pode não ser totalmente conectado. Também no maior componente conectado, calculamos o diâmetro, que é a distância do maior caminho mínimo existente no grafo.

5.1.1.4. *Betweenness*

A *betweenness* é uma métrica de centralidade que mede a importância de um determinado nodo ou aresta na rede referente à sua localização, considerando o número de caminhos mínimos que por ali passam. Nodos ou arestas com maior valor de *betweenness* fazem parte de um número maior de caminhos mínimos e por isto são mais importantes da rede.

5.1.1.5. *Closeness*

Assim como a *betweenness*, esta medida também é uma métrica de centralidade e mede a distância geodésica (menor distância) média entre um vértice *i* a todos os outros vértices do grafo. Esta métrica informa a velocidade com a qual uma informação se propaga de um vértice para o resto da rede.

5.1.1.6. *PageRank*

O algoritmo *PageRank* [13] foi proposto inicialmente para ordenar páginas Web por uma máquina de busca. O algoritmo gera pesos para um determinado nodo *i* da rede, levando em conta a importância dos nodos que apontam para *i*. O *PageRank* considera que um nodo é importante se muitos nodos apontam para ele ou se existem nodos importantes apontando para ele.

6. Estudo preliminar

Em [14] é apresentado um estudo sobre a rede de coautoria do SBBD, que em 2010 completou 25 anos de existência. Foram coletados dados bibliográficos de todas as 25 edições, e em seguida, foram realizados alguns estudos estatísticos e foi construída também a rede de coautoria do SBBD.

As Figuras 2 a 6 mostram os estudos realizados em [14] juntamente com um estudo preliminar realizado neste trabalho, em que foram utilizadas janelas temporais de três e cinco anos. Nas Figuras podemos observar os resultados (a) sem janelas temporais, considerando (b) janelas temporais de 3 anos e (c) janelas temporais de 5 anos. Na Figura 2 podemos observar o diâmetro da rede, sendo que o pico não é atingindo ao final dos anos, conforme o estudo original. Na Figura 3 é mostrado que o maior componente conectado e o segundo maior componente conectado possuem menor diferença de tamanho considerando as janelas temporais, sendo em alguns períodos até de tamanho similares. O caminho mínimo médio é apresentado na Figura 4, sendo esta propriedade mais semelhante ao estudo original, mesmo os valores estando em uma escala menor, verifica-se que o comportamento da curva é semelhante, apresentando uma evolução no início e em seguida uma estabilização. Considerando a janela de 3 anos (Figura 4b), temos um decaimento no último ano, isto ocorre porque neste estudo consideramos somente o ano de 2010, devido a divisão de 3 em 3 anos. A Figura 5 mostra o coeficiente de agrupamento, na Figura 5a observa-se uma estabilidade na curva, já no estudo realizado utilizando janelas (Figuras 5b e 5c), observamos que as curvas apresentam um ganho através dos anos. Por fim, a Figura 6 apresenta a evolução do número de artigos e autores ao longo dos anos. No estudo original (Figura 6a), observa-se um crescimento considerável, já nos estudos utilizando janelas temporais (Figuras 6b e 6c), observa-se que este crescimento não é tão acelerado.

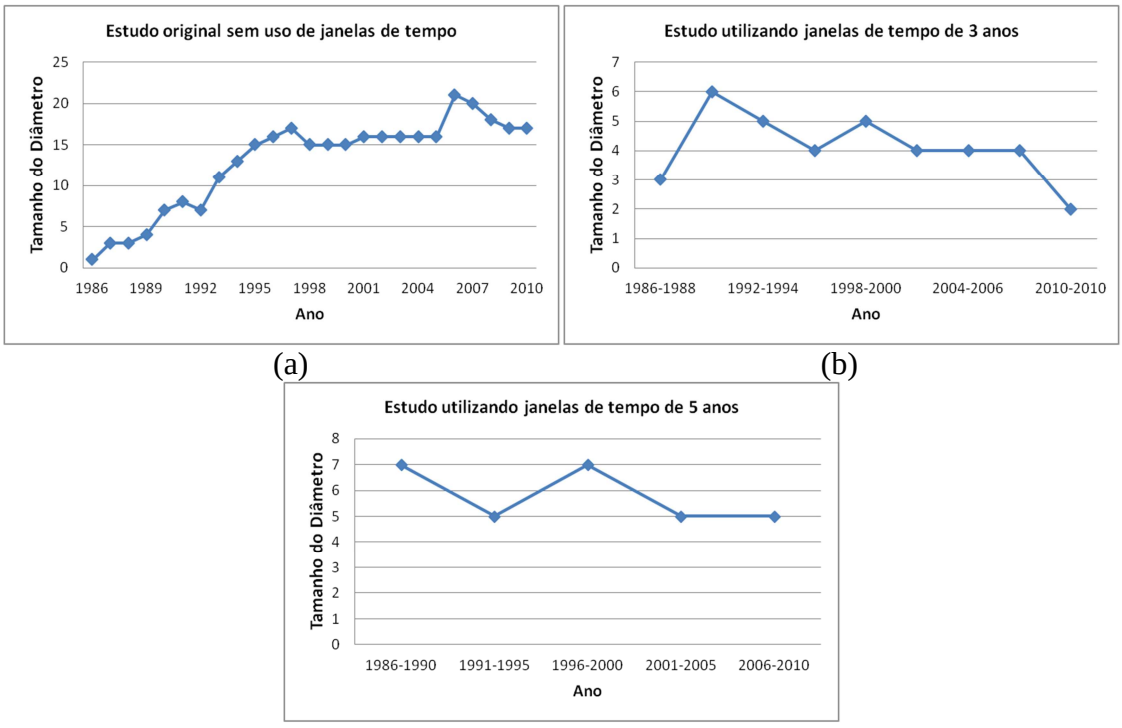

Figura 2. Evolução do diâmetro da rede: a) sem uso de janelas temporais, b) utilizando janelas de tempo de 3 anos e c) utilizando janelas de tempo de 5 anos

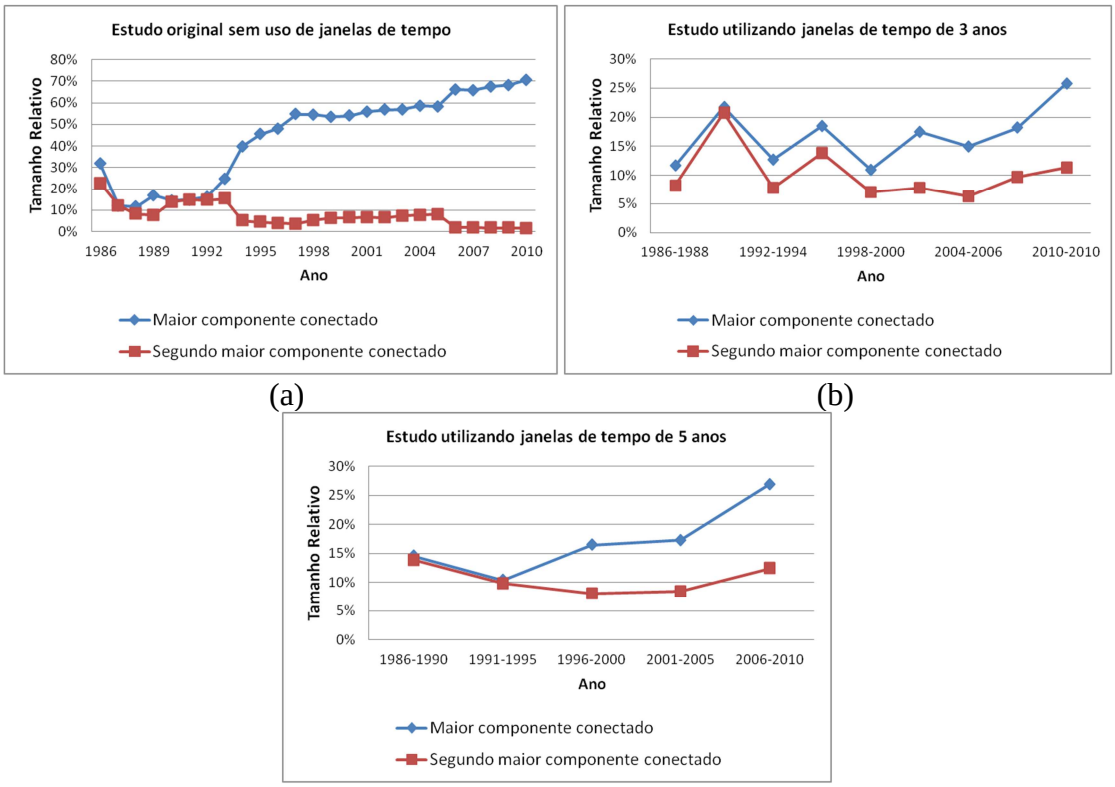

Figura 3. Tamanho relativo dos componentes conectados: a) sem uso de janelas temporais, b) utilizando janelas de tempo de 3 anos e c) utilizando janelas de tempo de 5 anos

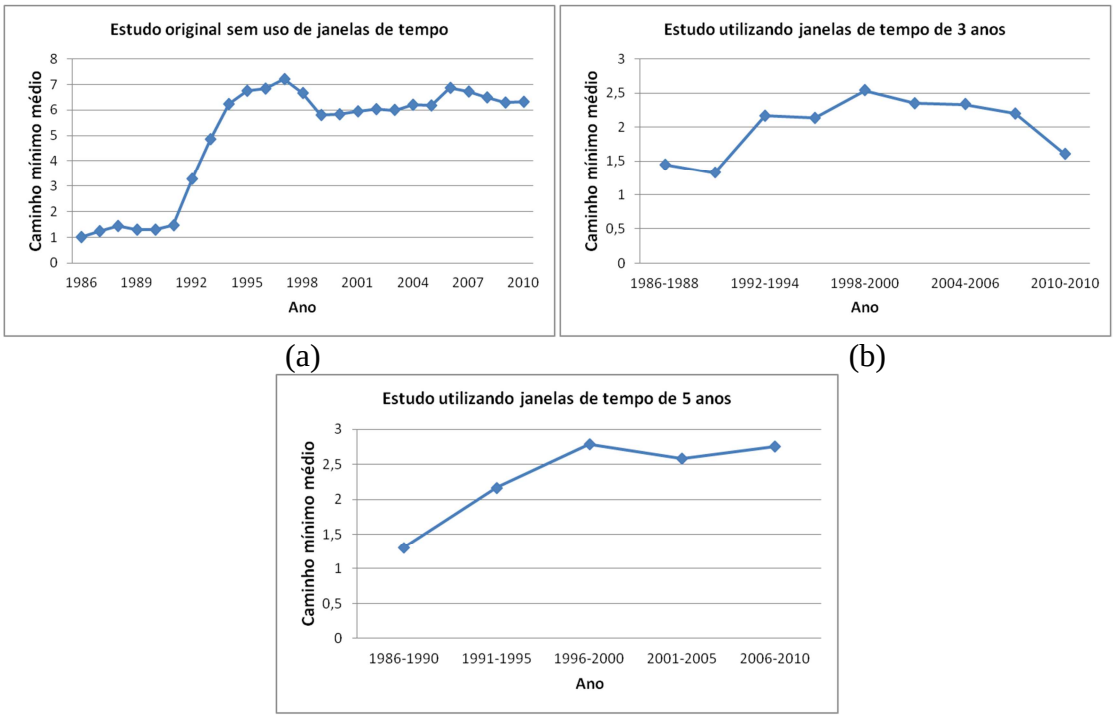

Figura 4. Evolução do caminho mínimo médio da rede: a) sem uso de janelas temporais, b) utilizando janelas de tempo de 3 anos e c) utilizando janelas de tempo de 5 anos

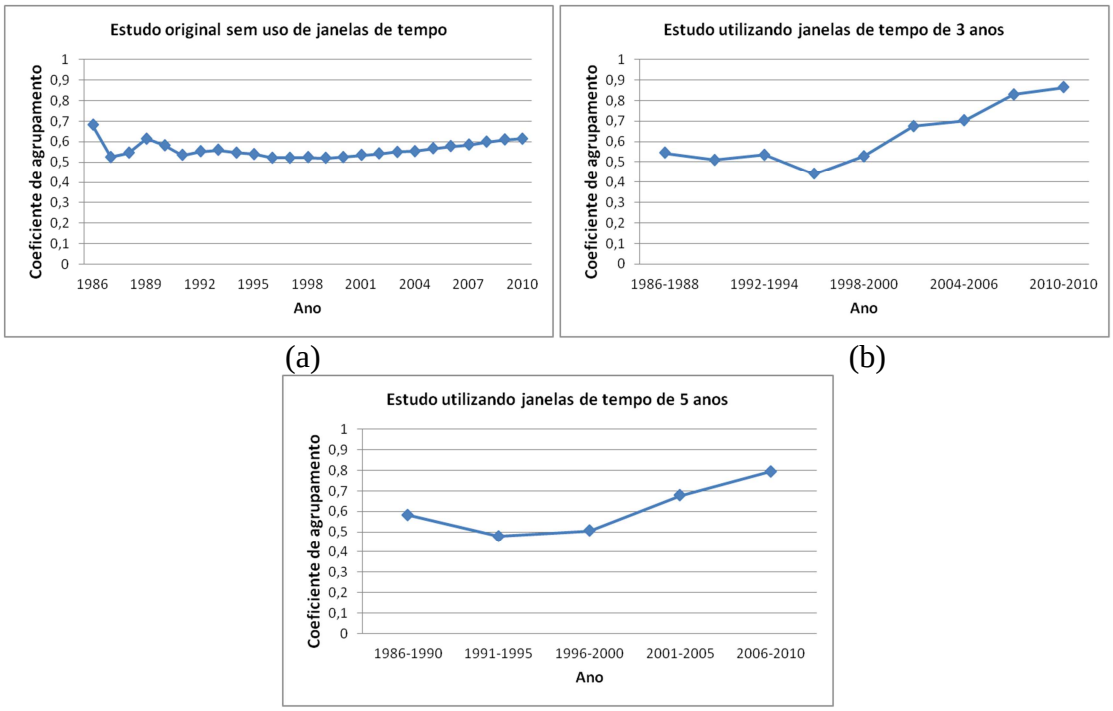

Figura 5. Evolução do coeficiente de agrupamento da rede: a) sem uso de janelas temporais, b) utilizando janelas de tempo de 3 anos e c) utilizando janelas de tempo de 5 anos

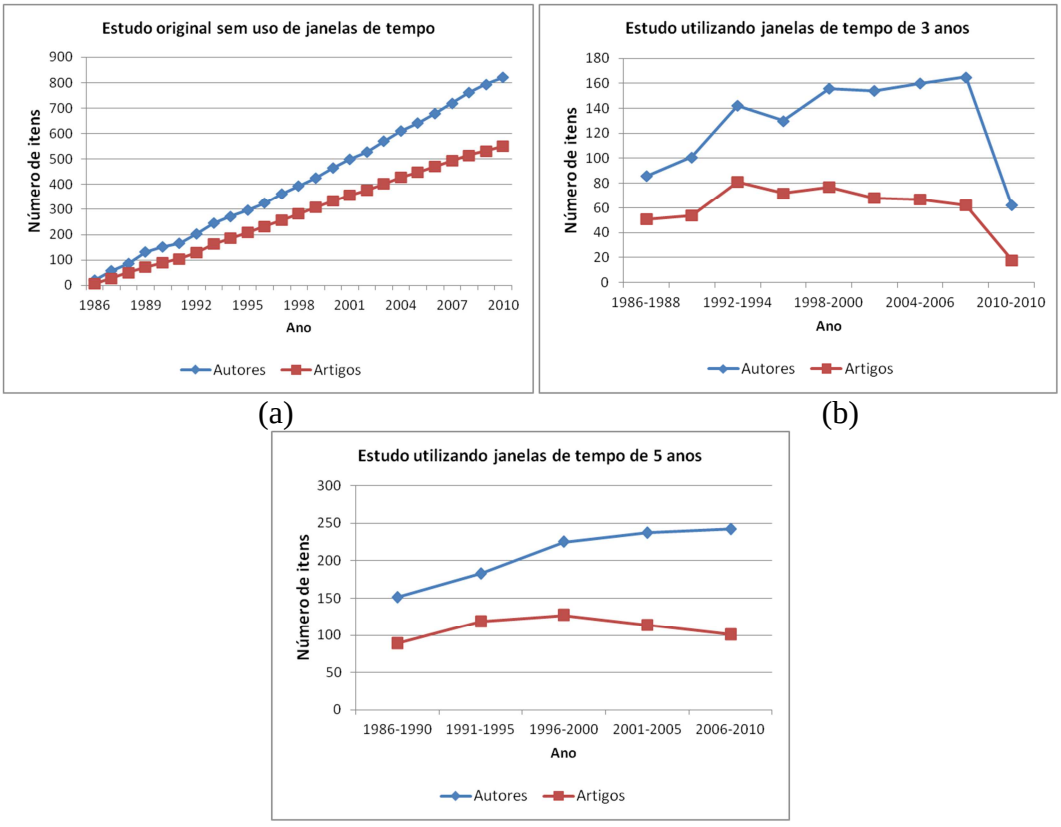

Figura 6. Evolução do número de autores e artigos ao longo do tempo: a) sem uso de janelas temporais, b) utilizando janelas de tempo de 3 anos e c) utilizando janelas de tempo de 5 anos

De forma geral podemos observar que mesmo utilizando janelas temporais de 3 e 5 anos, o número de autores apresentou um aumento a cada ano, o que mostra que novos autores aparecem na conferência ao longo dos anos, o que também é mostrado em [14]. No entanto, o número de artigos não apresenta um crescimento constante, conforme apresentado em [14], utilizando as janelas temporais poderíamos identificar os pontos de maior produtividade na rede, ou seja, os períodos que apresentam um maior número de artigos publicados, conforme pode ser observado na Figura 6c, tendo o intervalo entre 1996 e 2000 o maior número de artigos publicados.
O estudo considerando os maiores componentes conectados também apresenta uma diferença considerável entre o estudo utilizando toda a rede e o estudo utilizando janelas temporais, uma vez que analisando toda a rede, o maior componente conectado possui cerca de 70% da rede e o segundo maior componente conectado possui menos que 10% dos nodos (Figura 3a). Utilizando janelas temporais podemos observar que o maior componente contado possui cerca de um quarto dos nodos da rede, em ambas as janelas utilizadas, e o segundo maior componente possui entre 10 e 15% dos nodos (Figura 3b e 3c). Isto pode ocorrer porque ao longo dos anos, os autores podem não dar prosseguimento a trabalhos em conjuntos, por exemplo, um orientador e um orientando que publicaram trabalhos juntos durante o período de um curso de graduação, mestrado ou doutorado, e após esse curso não tiveram mais publicações em conjunto. Em estudos que consideram arestas formadas de forma acumulativa, conforme realizado em [14], não é possível identificar o efeito da ocorrência de arestas desfeitas ou que não foram reforçadas através dos anos, ou seja, este tipo de estudo não faz distinção de autores que realizaram apenas uma publicação em conjunto em toda a história da rede.
É possível verificar uma correlação entre a variação do tamanho do maior componente conectado, do diâmetro e do tamanho do caminho mínimo médio no estudo apresentado em [14]. No entanto, podemos verificar períodos que esta correlação não existe, uma vez que existem períodos em que o maior componente conectado aumenta enquanto o diâmetro e o caminho mínimo médio decaem. Vários fatores podem levar a este comportamento, por exemplo, durante um determinado período, os pesquisadores de uma determinada instituição, que realizam grandes contribuições para a rede, podem começar a colaborar mais entre si devido a um evento externo, fazendo com que a relação com pesquisadores de outras instituições seja enfraquecida nesse período, com isso poderíamos ter um decaimento no tamanho do maior componente conectado naquele período, enquanto o diâmetro poderia sofrer um aumento devido à inserção de novos autores na rede que não possuem interação com os pesquisadores já existentes.

7. Metodologia

7.1. Plano de atividades

A seguir descrevemos as atividades a serem seguidas neste trabalho. Na Figura 7 apresentamos um cronograma para as atividades propostas.

7.1.1. Revisão bibliográfica

Esta tarefa está em andamento, sendo que uma pesquisa inicial dos trabalhos relacionados ao estudo aqui proposto foi realizada. Os próximos passos para esta atividade são:

- Aprofundamento nos estudos de redes complexas e suas métricas de avaliação;
- Estudo sobre métricas temporais em grafos e outras áreas, como indexação de documentos.

7.1.2. Coleta dos dados

Existem diversas redes que podem ser usadas para o estudo aqui proposto. Inicialmente trabalharemos com dados da DBLP e do Twitter. A DBLP já disponibiliza os dados no formato XML para download, já o Twitter permite a coleta dos dados via API, desta forma será necessário realizar a coleta via API ou obter os dados de algum repositório que já realizou tal coleta.

7.1.3. Formalização das métricas a serem utilizadas

As métricas aqui descritas devem ser analisadas e expandidas, a fim de verificar sua viabilidade em grandes redes e aderência à questão temporal, objeto alvo deste trabalho.

7.1.4. Definição da arquitetura

Definição de uma arquitetura que suporte grandes grafos, bem como uma arquitetura que proporcione flexibilidade para análise de vários grafos, sem a necessidade de grandes esforços de implementação na variação de uma rede para outra.

7.1.5. Projeto

Nesta fase realizaremos o projeto dos algoritmos e métricas a serem implementados para o estudo com base na arquitetura proposta, nas métricas escolhidas e nas bases coletadas.

7.1.6. Implementação

Os algoritmos e métricas propostas serão implementadas nas bases escolhidas de acordo com o projeto realizado e a arquitetura definida.

7.1.7. Avaliação experimental

Projeto dos experimentos e avaliação experimental dos algoritmos e métricas utilizadas.

7.1.8. Escrita da dissertação

Redigir a dissertação e possíveis artigos para publicação do trabalho.

7.2. Cronograma

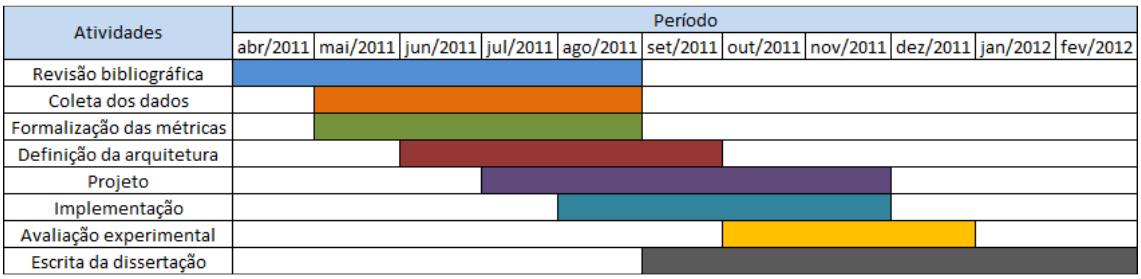

Figura 7. Cronograma para execução das atividades

8. Referências

[1] Albert, R., Barabási, A. L., 2002. Statistical mechanics of complex networks. Rev. Mod. Phys., 2002, 74.

[2] Dorogovtsev, S. N., Mendes, J. F. F., 2002. Evolution of networks. Adv. Phys., 2002, 51.

[3] Newman, M. E. J., 2003. The Structure and Function of Complex Networks. SIAM Rev., 2003, 167.

[4] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Huang, D. U., 2006. Complex networks: Structure and dynamics. Phys. Rep. 424, 2006, 175.

[5] Costa, L. F., Rodrigues, F. A., Travieso, G., Boas, P. R. U., 2007. Characterization of complex networks: A survey of measurements. Adv. Phys. 56, 2007, 167.

[6] Torkjazi, M., Rejaie, R., Willinger, W., 2009. Hot today, gone tomorrow: On the migration of myspace users. In ACM SIGCOMM Workshop on Online social networks (WOSN), 2009, 43-47.

[7] Liben-Nowell, D., Kleinberg, J., 2007. The link prediction problem for social networks. JASIST 58, 2007, 1019-1031.

[8] Wang, L., Hopcroft, J. E., 2010. Community Structure in Large Complex Networks. TAMC, 2010, 455-466.

[9] Shibata, N., Kajikawa, Y., Sakata, I., 2012. Link prediction in citation networks. Journal of the American Society for Information Science and Technology, 63, 2012, 78–85.

[10] Adamic, L., Buyukkokten, O., Adar, E. 2003. A social network caught in the web. First Monday, 2003, 8.

[11] Benevenuto, F., Duarte, F., Rodrigues, T., Almeida, V., Almeida, J., Ross, K., 2008. Understanding video interactions in YouTube. In ACM Conference on Multimedia (MM), 2008, 761–764.

[12] Benevenuto, F., Almeida, J., Silva, A., 2011. Coleta e Análise de Grandes Bases de Dados de Redes Sociais Online. Jornada de Atualizações em Informática (JAI), CSBC, 2011.

[13] Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 1998, 107–117.

[14] Procopio Jr., P. S., Laender, A. H. F., Moro, M. M., 2011. Análise da Rede de Coautoria do Simpósio Brasileiro de Bancos de Dados. Sessão de Pôsteres, Simpósio Brasileiro de Banco de Dados, 2011.