ANÁLISE DE REDES SOCIAIS CIENTÍFICAS

Victor Ströele de Andrade Menezes

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientador: Geraldo Zimbrão da Silva, D. Sc.

Rio de Janeiro
Abril de 2012

ANÁLISE DE REDES SOCIAIS CIENTÍFICAS

Victor Ströele de Andrade Menezes

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____

Prof. Geraldo Zimbrão da Silva, D. Sc.

_____

Prof. Jano Moreira de Souza, Ph. D.

_____

Prof. Geraldo Bonorino Xexéo, D. Sc.

_____

Prof. Jonice de Oliveira Sampaio, D. Sc.

_____

Prof. Renata de Matos Galante, D. Sc.

RIO DE JANEIRO, RJ – BRASIL

ABRIL DE 2012

Menezes, Victor Ströele de Andrade

Análise de Redes Sociais Científicas/Victor Ströele de Andrade Menezes. – Rio de Janeiro: UFRJ/COPPE, 2012.

XVI, 206 p.: il.; 29,7 cm.

Orientador: Geraldo Zimbrão da Silva

Tese (doutorado) – UFRJ / COPPE / Programa de Engenharia de Sistemas e Computação, 2012.

Referências bibliográficas: p. 190-204

1. Redes Sociais Científicas. 2. Fluxo em Redes 3. Mineração de Dados. 4. Sugestão de Relacionamentos. I. Silva, Geraldo Zimbrão da. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

A Deus, que me deu determinação para o
desenvolvimento deste trabalho.
À minha esposa por estar sempre ao meu lado e
pelo apoio incondicional.
À minha família pelo incentivo constante e pelos
conselhos nos momentos de dificuldades.

AGRADECIMENTOS

Agradeço a Deus pela benção da vida e por ter colocado em meu caminho pessoas tão especiais, que tive o prazer de conhecer e conviver durante essa jornada.

À minha querida esposa que teve o dom de me acalmar em momentos que parecia que nada iria se resolver. Ela soube ser exatamente o que eu precisava que ela fosse: esposa, companheira, amiga, psicóloga, colega de trabalho, co-autora, secretária, etc. Com seu jeito doce ela me guiou durante todos esses anos que estamos juntos. Sem você mozim eu não teria chegado até aqui.

Aos meus pais que dedicaram momentos preciosos das suas vidas, ajudando a trilhar o caminho correto, abdicando de coisas importantes de suas próprias vidas, estando ao meu lado nos momentos difíceis e ensinando-me a viver. Por eles serem até hoje o meu porto seguro e uma referência para minha vida.

Ao meu irmão e à Go que sempre me proporcionam momentos de muita alegria e estão sempre torcendo por mim.

À minha Avó que mesmo à distância está sempre torcendo e rezando por mim. Por ela nunca se esquecer desse neto que durante esse período esteve tão afastado fisicamente.

Agradeço aos professores Jano e Blaschek que em 2005 confiaram no meu potencial permitindo que eu trabalhasse nos projetos da Fundação Coppetec, mesmo eu não sendo da linha de pesquisa de Banco de Dados. Sem esse apoio, que me é dado até hoje pelo professor Jano, eu não conseguiria o suporte financeiro que eu tive e, com toda certeza, eu não teria conseguido concluir meus estudos.

Agradeço ao professor Jano novamente por ter me aceito para o desenvolvimento do meu doutorado na linha de pesquisa de Banco de Dados e por tantas vezes me dar sugestões criativas e úteis para o desenvolvimento do meu trabalho.

Agradeço ao professor Zimbrão por ele ter me aceito como seu aluno mesmo sem me conhecer. Agradeço por ele estar sempre me auxiliando, orientando e ainda assim permitindo que eu tivesse minhas próprias opiniões e idéias. Com sua visão precisa e seu entendimento rápido sobre os problemas que enfrentamos, ele soube guiar o desenvolvimento dos trabalhos realizados.

Agradeço à Jonice que no início do meu doutorado me colocou em contato com o assunto que se tornou tema deste trabalho. Com sua humildade e simplicidade ela me ensinou muito sobre esse tema.

Agradeço aos professores Jano, Xexéo, Jonice e Renata por aceitarem o convite de participação nesta banca examinadora. Agradeço por eles estarem dispostos a contribuír com seus conhecimentos para aperfeiçoar o trabalho desenvolvido.

À UFRJ, especialmente à COPPE/PESC, por oferecer a possibilidade de ter realizado um curso de excelência no nível de doutorado, avaliado com nota máxima na CAPES.

A CAPES por ter financiado minhas pesquisas.

A todos o meu MUITO OBRIGADO!!!

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

ANÁLISE DE REDES SOCIAIS CIENTÍFICAS

Victor Ströele de Andrade Menezes

Abril/2012

Orientador: Geraldo Zimbrão da Silva

Programa: Engenharia de Sistemas e Computação

As Redes Sociais são estruturas sociais dinâmicas formadas por indivíduos ou organizações. Geralmente, essas redes são representadas por nós ligados por um ou mais tipos de relacionamentos. Embora sejam estruturas extremamente complexas, analisá-las nos permite detectar diversos tipos de conexões entre as pessoas dentro e fora de suas instituições.

Este trabalho apresenta uma proposta baseada em técnicas de Mineração de Dados com o intuito de identificar ligações intra e inter organizacionais e permitir uma análise detalhada dessas ligações. Com o uso de técnicas de agrupamento são identificadas estruturas sociais e comunidades de pesquisas de forma que o fluxo de conhecimento na rede social pode ser avaliado. Além disso, novos relacionamentos podem ser sugeridos para melhorar o fluxo de informações na rede social. Assim, propomos um framework para suporte à análise das redes sociais multi-relacionais, que inclui módulos voltados para análise e previsão de novos relacionamentos, além de um módulo voltado para a análise visual da rede.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the Candidacy for Doctor of Science (D.Sc.)

SCIENTIFIC SOCIAL NETWORK ANALYSIS

Victor Ströele de Andrade Menezes

April/2010

Advisor: Geraldo Zimbrão da Silva

Department: Computer and Systems Engineering

Social Networks are dynamic social structures formed by individuals or organizations. Usually, these networks are represented by nodes connected by one or more relationships types. Although they are extremely complex structures, analyze it allows us to detect many types of connections between people within and outside their institutions.

This work presents a proposal based on data mining techniques in order to identify intra and inter organizational relationships and allow a detailed analysis of these connections. Using clustering techniques we can identify social structures and research communities, so that the flow of knowledge in the social network can be analyzed. In addition, new relationships can be suggested to improve the information flow in social networks. Therefore, we propose a framework to support the analysis of multi-relational social network, which includes modules focused on link prediction and relationships analysis, and a module devoted to visual analysis of the network.

Sumário

Índice de Figuras

Índice de Tabelas

Capítulo 1 – Introdução

1.1 O Problema

Atualmente, as pessoas estão se tornando cada vez mais dependente da tecnologia. Dispositivos eletrônicos, como laptops, *smartphones* e *tablets*, entre outros aparelhos, fizeram com que as pessoas também tivessem uma vida no mundo virtual. A tecnologia dá suporte a várias ferramentas sociais que permitem que as pessoas se comuniquem virtualmente e troquem informações em tempo real.

Os usuários dessas ferramentas sociais (e-mail, blogs, micro-blogs, *wikis*) enviam mensagens pessoais ou públicas, postam suas opiniões pessoais, contribuem com seus conhecimentos para alguma comunidade criando parcerias, etc. Os usuários que são devotos dessas ferramentas criam e compartilham documentos digitais, recomendam ou não sobre determinado assunto e também ajudam as pessoas que utilizam a mesma ferramenta social.

Milhares de pessoas criam bilhões de conexões através da mídia social a cada dia, mas poucos de nós compreendem como cada clique e cada tecla pressionada constroem relacionamentos que, como um todo, formam uma vasta rede social onde as pessoas compartilham o conhecimento em um ambiente virtual (Huang, Liben-Nowell and Kleinberg, 2003; Zhuge and Guo, 2007).

O resultado dessas interações é uma rede complexa de conexões que ligam as pessoas a outras pessoas, documentos, objetos, locais, conceitos, etc. Esse crescimento de pessoas interagindo virtualmente uns com os outros tem dado origem a um grande número de Redes Sociais. Novas ferramentas estão sendo desenvolvidas com o intuito de visualizar e analisar essas interações que são representadas através dessas grandes redes sociais.

Atualmente, existem muitos dados com conceitos diversos distribuídos pela Web. Alguns desses dados possuem características políticas, econômicas, ou sociais. Tais dados podem ser extraídos de páginas pessoais, ferramentas sociais e colaborativas, dentre outras formas de comunicação que produzam esses tipos de dados.

Além dos tipos citados anteriormente existem também os dados utilizados nos ambientes científicos, como currículos de pesquisadores, publicações científicas, livros, dentre outros. Muitos desses dados são utilizados como fonte de pesquisa por pesquisadores em diversas universidades para desenvolver seus trabalhos científicos. Pode-se dizer que, nos dias atuais, a Web é a principal interface usada pelos pesquisadores para buscar os trabalhos relacionados àqueles que estão sendo desenvolvidos por eles.

Muitas questões da humanidade são solucionadas através dos resultados obtidos pelas pesquisas realizadas nas universidades espalhadas por todo o mundo. Assim, a análise dos dados científicos irá nos permitir compreender como está o andamento das pesquisas nas diversas instituições de ensino e como esses pesquisadores e as universidades se relacionam.

Com a facilidade de acesso aos dados, a Web é a principal interface usada pelos pesquisadores na busca por trabalhos relacionados aos trabalhos que estão sendo desenvolvidos por eles. Muitas vezes um pesquisador usa a Web para encontrar artigos, livros, ou mesmo outro pesquisador que domina o assunto para auxiliar no desenvolvimento de suas pesquisas. Assim, os pesquisadores criam ligações entre eles pela colaboração direta ou indireta, citação e avaliação do trabalho desenvolvido em conjunto.

A crescente publicação de artigos nos permite identificar os pesquisadores que estão trabalhando em conjunto. A questão é que existem muitos grupos formados por poucos pesquisadores, ou seja, a maioria dos grupos desenvolve suas pesquisas com a colaboração de poucas pessoas (Barabasi, Jeong, Neda, Ravasz, Schubert and Vicsek, 2002). Com isso, dois grupos de pesquisas que não se relacionam podem estar aplicando seus esforços no estudo do mesmo assunto, sem nenhuma colaboração entre eles.

Esse fenômeno permite que os pesquisadores estudem como as conexões entre as pessoas são estabelecidas e como elas evoluem ao longo do tempo. No ambiente científico essa análise pode contribuir para que haja uma maior colaboração entre os grupos e, com isso, a qualidade e o volume de publicações também aumentem.

Vários esforços têm sido feitos para analisar as redes sociais a fim de ajudar a compreender as estruturas sociais (Freeman, 1979; Wasserman and Faust, 1994). Todas

essas pessoas interligadas por diferentes relações podem ser representadas como redes sociais.

Uma rede social é um conjunto de objetos, onde cada um deles está conectado a outro objeto. Uma rede social pode ser representada por um grafo no qual os nós ou vértices estão relacionados ou não por arestas. Uma rede social reflete uma estrutura social que pode ser representada por indivíduos ou organizações e suas relações. Em geral, as relações representam um ou mais tipos de interdependência (como idéia e religião) ou relacionamentos mais específicos (como troca de conhecimento, informação e amizade). Assim, com essa estrutura social, a troca de dados e informações entre os indivíduos ou organizações pode ser estudada e analisada em diferentes níveis de detalhes.

1.2 Motivação

O crescimento da Web em conjunto com a crescente disponibilidade de dados está permitindo que os pesquisadores estudem as maneiras pelas quais as conexões entre as pessoas são estabelecidas e como elas evoluem ao longo do tempo. Nesse sentido, vários esforços vêem sendo feitos para analisar essas redes, contribuindo para uma melhor compreensão das estruturas sociais (Freeman, 1979; Wasserman and Faust, 1994). Para auxiliar esses estudos, as redes sociais foram construídas de forma a representar as pessoas e as relações entre elas.

A perspectiva de rede enfatiza as relações estruturais como o seu princípio fundamental de orientação, ou seja, as estruturas sociais consistem de regularidades nos padrões das relações entre as entidades concretas, ou atores (Knoke and Song, 2008). Essas entidades concretas são pessoas, pequenos grupos, organizações, Estados ou Países, dentre outros. Os padrões regulares das relações são contextos macro-sociais que, de alguma forma, influenciam as percepções, crenças, decisões, e ações dessas entidades.

Os atores e seus relacionamentos são os elementos fundamentais na constituição de uma rede social. Os atores podem ser representados por uma pessoa individualmente ou por um grupo de pessoas, como grupos informais ou organizações formais. Um grupo de crianças brincando em um parquinho é um exemplo de uma rede social na qual os

atores são representados individualmente pelas crianças. Por outro lado, a competição entre equipes de crianças é um exemplo de rede social com atores coletivos.

Uma relação é definida como um tipo específico de contato, ou conexão entre um par de atores. As relações podem ser diretas ou indiretas. Nas relações diretas um ator fornece uma informação e o outro ator recebe essa informação diretamente. Já nas relações indiretas a informação chegará ao seu destino passando por atores intermediários, ou seja, existem atores que irão conectar os outros dois que estão trocando informações.

Existem vários tipos de relações, as crianças juntas em um parquinho podem ter uma relação de brincadeira, amizade, briga, ou confidência, por exemplo. No caso das redes sociais coletivas onde os atores representam equipes podem existir relacionamentos de competição, colaboração, comunicação, serviços, dentre outros. Assim, pode-se observar que as relações não são atributos dos atores, mas sim uma junção de interesses comuns em um assunto que só existirá enquanto ambos os atores tiverem interesse em manter essa associação.

Uma Rede Social reflete uma estrutura social através da qual podem ser estudadas as trocas de informações e de dados entre os indivíduos ou organizações que compõem essas redes. Esses estudos fazem parte da análise das redes sociais, cujos objetivos centrais são mensurar e representar essas estruturas relacionais e explicar o motivo pelo qual elas ocorrem e quais são as conseqüências de sua ocorrência.

A análise das redes sociais proporciona três resultados principais (Knoke and Song, 2008). Primeiro, as estruturas das relações são mais importantes para compreender um relacionamento entre duas entidades do que os atributos de perfil dessas entidades, tais como: idade, sexo, ideologia, etc. Assim, como as relações influenciam as entidades sociais, sem considerar seus atributos individuais, a análise da rede anseia por oferecer uma melhor compreensão das ações sociais produzidas por essas relações, avaliando a estrutura da rede como um todo.

Nas redes sociais científicas esse primeiro resultado indica como é o relacionamento entre os pesquisadores, independentemente de quais são as características pessoas de cada um deles. Com isso, será possível, por exemplo, localizar dois pesquisadores que estão bem relacionados, mesmo que eles sejam de áreas distintas.

Segundo, as redes sociais, como dito anteriormente, afetam percepções, crenças e ações através de uma série de estruturas que são socialmente construídas pelas relações entre as entidades. As relações sociais diretas facilitam a interação entre as entidades favorecendo a troca de informação e aumentando as chances da entidade influenciar e ser influenciada por outras entidades. As relações indiretas, embora com menor presença, também expõe as entidades às novas influências através de transitividade por uma entidade que esteja relacionada diretamente. Com a análise das redes sociais científicas deseja-se compreender como as relações afetam os pesquisadores, direta ou indiretamente.

Terceiro, os relacionamentos estruturais podem ser vistos como processos dinâmicos. As redes sociais não são estáticas, elas sofrem constantes alterações através das interações entre as pessoas, grupos, ou organizações. As entidades podem transformar as estruturas relacionais, às quais elas fazem parte direta ou indiretamente, de forma intencional ou não, através da influência dos seus conhecimentos. Assim, com a análise da rede social científica, podem ser localizados os principais influenciadores dessa rede e quais são as conseqüências dessa influência.

Em síntese, pode ser dito que analisar uma rede social científica permite identificar comunidades de pesquisas, pesquisadores que detêm maior influência (centralizadores), compreender a evolução social dos pesquisadores ao longo do tempo, sugerir novos relacionamentos para aprimorar as pesquisas dos pesquisadores ou melhorar o fluxo de conhecimento da rede, etc.

1.3 Objetivos

Do ponto de vista da Mineração de Dados, uma rede social é um conjunto de dados multi-relacional representado por um grafo (Han and Kamber, 2006). O grafo é tipicamente muito grande, com nós correspondendo a objetos e arestas às relações entre esses objetos. No mundo real as arestas geralmente representam tipos de relacionamentos diferentes.

Em mineração de dados, a área que estuda as redes sociais é chamada de mineração de ligações (*link mining*) ou análise de ligações (*link analysis*) (Agrawal,

Rajagopalan, Srikant and Xu, 2003; Han and Kamber, 2006) e um dos desafios dessa área é a detecção de grupo, que é a identificação de objetos que pertencem ao mesmo grupo ou cluster. Este trabalho mostra o uso e a avaliação da nossa abordagem na identificação de relações científicas, com base nos currículos dos pesquisadores disponíveis na web.

No mundo real, as redes sociais são em sua maioria multi-relacionais, ou seja, pessoas ou instituições estão relacionadas através de tipos de relacionamentos diferentes. Uma das propostas apresentadas neste trabalho é a construção de uma rede social multi-relacional que permita que as análises feitas através dela tenham a influência de todos os relacionamentos que cada indivíduo possui.

Com o intuito de permitir que a análise da rede social científica tenha a influência de vários tipos de relacionamentos, está sendo proposto neste trabalho um modelo para a construção de uma rede social científica multi-relacional. Nesse modelo são considerados vários fatores que influenciam a análise das redes sociais, tais como: tipos de relacionamentos diferentes, peso do relacionamento, idade do relacionamento, perda de informação na transferência de conhecimento entre os pesquisadores, etc. O modelo proposto, embora tenha sido construído para a análise de redes sociais científicas, pode ser adotado para qualquer tipo de rede social.

Baseado em uma rede social científica multi-relacional formulada através do modelo proposto, um dos temas deste trabalho é a busca por *comunidades de pesquisa*. O objetivo é identificar os grupos de pesquisadores que têm interesses comuns no desenvolvimento de suas pesquisas.

Com o objetivo de identificar os pontos levantados acima, um dos objetivos deste trabalho é agrupar as pessoas com relacionamentos comuns na rede social científica, por meio de técnicas de mineração de dados. No Capítulo 6, no qual é apresentado o estudo de caso, a rede social é analisada em vários níveis de detalhes e algumas conclusões são apresentadas, dentre elas a colaboração entre os pesquisadores e entre as universidades.

Neste trabalho foi analisado o fluxo de conhecimento em uma rede social científica multi-relacional. Para tal, foi desenvolvida uma técnica de agrupamento por fluxo máximo para encontrar as comunidades de pesquisa. Esse algoritmo se baseia no fluxo máximo de conhecimento entre os elementos fazendo com que os pesquisadores que possuem maior poder de comunicação entre si fiquem no mesmo grupo. A base do

algoritmo desenvolvido é o algoritmo de agrupamento *k-medóides*. Através desse algoritmo foi possível fazer uma análise detalhada das relações intra e interinstitucionais.

Identificando as comunidades científicas é possível avaliar como ocorre o fluxo de conhecimento entre as instituições de ensino e entre as próprias comunidades. O algoritmo de agrupamento por fluxo máximo foi aplicado a uma rede social científica formada por pesquisadores de cinco instituições de ensino brasileiras e os grupos identificados foram detalhadamente analisados.

Embora o estudo das redes sociais científicas multi-relacionais identifique pesquisadores que possuem o mesmo objetivo de pesquisa, também podem ser identificados grupos interdisciplinares. Esses grupos são construídos por pesquisadores pertencentes a áreas distintas.

Além disso, como dito anteriormente, um dos temas explorados neste trabalho são os centralizadores de conhecimento. Esses elementos, além de serem pontos críticos, pois centralizam o conhecimento, podem ser classificados como: pessoas que, em virtude de suas relações com as pessoas em diferentes organizações servem como *chaves de fronteira* (passando informações e contexto de um grupo para outro) ou *pontos de estrangulamento* (impedindo o fluxo de informações e contexto) (Jackson, 2010). Como as relações da rede social científica envolvem um conceito colaborativo, serão identificados neste trabalho alguns dos centralizadores de conhecimento que são chaves limite.

Muitos estudos sobre redes sociais procuram padrões em redes estáticas, identificando características de um “momento” da rede social. No entanto, tem sido difícil analisar tendências ao longo do tempo, dada a falta de informação sobre a evolução das redes sociais ao longo de períodos de tempo (Leskovec, Kleinberg and Faloutsos, 2005).

Assim, hoje em dia, são poucos os estudos que utilizam informações sobre o tempo nas análises de previsão de relacionamentos (Acar, Dunlavy and Kolda, 2009; Potgieter, April, Cooke and Osunmakinde, 2009). Nas redes sociais científicas é possível fazer uma análise das tendências dos relacionamentos utilizando informações temporais, uma vez que esses tipos de relacionamentos têm informações temporais.

No contexto de previsão de relacionamentos, o objetivo deste trabalho é sugerir novos relacionamentos para melhorar a qualidade da comunicação na rede social. Para sugerir/prever novos relacionamentos será proposto o uso de uma nova métrica composta por três outras métricas, que serão descritas no Capítulo 7. Essas métricas avaliam todas as relações da rede, considerando seu histórico evolutivo. Através da sugestão de relacionamentos um pesquisador pode criar, por exemplo, novos vínculos com pesquisadores de outras instituições para a qual ele está sendo transferido.

O fator que mais contribuiu para a construção da métrica composta foi a idade dos relacionamentos científicos. Através da análise do tempo foi possível identificar o comportamento evolutivo da rede social científica e, conseqüentemente, definir uma métrica que atendesse melhor a esse comportamento.

No final desse trabalho foi construída uma ferramenta de visualização que permite ao usuário fazer análises visuais e temporais da rede social científica. Essa ferramenta apresenta níveis de visualizações diferentes e exibe todos os relacionamentos ou apenas os relacionamentos mais fortes; o usuário pode fazer a análise de relacionamentos específicos ou de todos os relacionamentos; podem ser aplicados filtros para que apenas os relacionamentos de alguns pesquisadores sejam exibidos; o usuário pode analisar a evolução da rede social através de uma linha do tempo; dentre outras funcionalidades para análise visual da rede social científica multi-relacional.

Com o auxílio da ferramenta de visualização desenvolvida, das análises sobre as comunidades científicas definidas pelo método de agrupamento por fluxo máximo, e do módulo de sugestão de relacionamentos, a rede social científica multi-relacional como um todo foi analisada sobre vários aspectos. Assim, foi possível definir e visualizar o fluxo de informação científica em redes sociais no Brasil.

Em geral, estudando a formação dessas redes sociais científicas é possível identificar como os pesquisadores e organizações estão desenvolvendo seus trabalhos. A análise das redes sociais científicas indica o grau de envolvimento entre os pesquisadores, entre as áreas de pesquisas e, até mesmo, entre as instituições de ensino. Podem ser identificados também alguns padrões de colaboração inter e intra-universidades. Todas essas informações podem proporcionar uma melhora na comunicação da rede social e, conseqüentemente, melhorar a colaboração entre os pesquisadores.

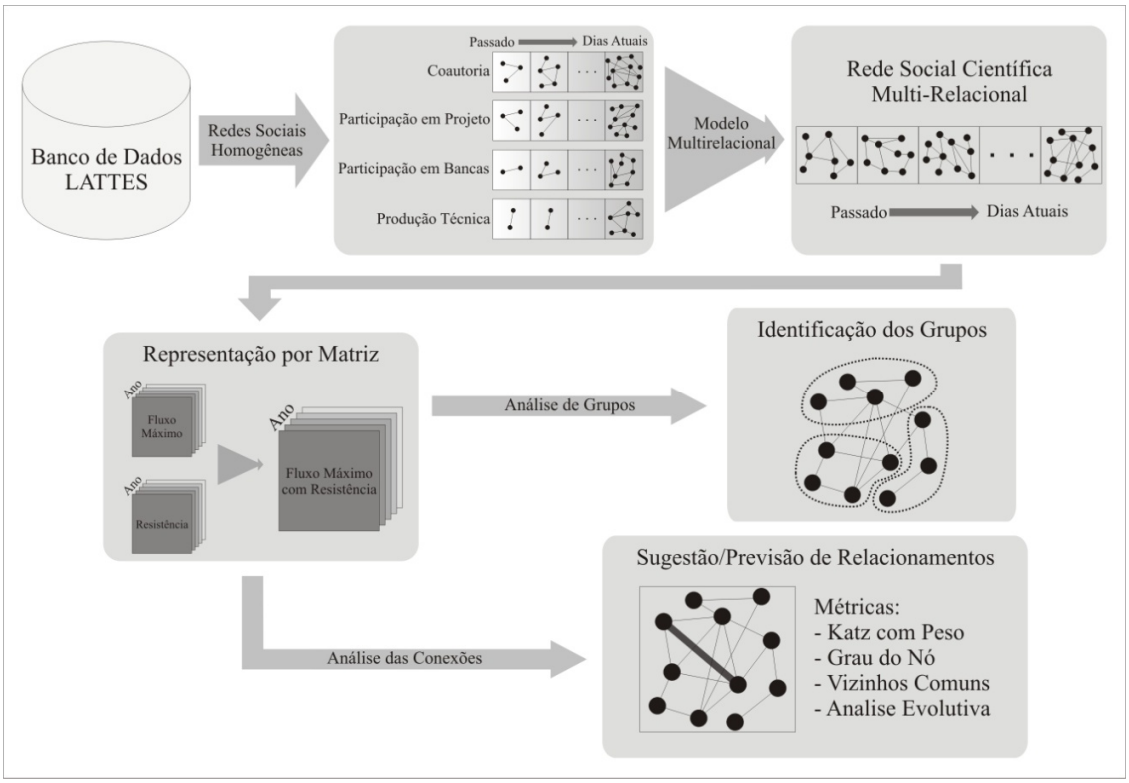


Figura 1 - Workflow da Proposta.

A Figura 1 ilustra todas as fases do desenvolvimento deste trabalho. Através dessa figura é fácil identificar cada um dos módulos descritos anteriormente. Além disso, os trabalhos futuros podem ser realizados independentemente em cada um dos módulos, basta que as saídas e entradas para cada módulo continuem as mesmas.

Os dados são extraídos do banco de dados e quatro redes sociais homogêneas são construídas de forma que a informação temporal de cada uma delas seja mantida. Cada uma dessas redes é formada por apenas um tipo de relacionamento científico.

As redes homogêneas são transformadas através da modelagem proposta em uma única rede com múltiplos relacionamentos científicos e os dados são armazenados em matrizes mantendo sempre a informação temporal dos mesmos.

Usando as matrizes resultantes do processo de modelagem da rede social científica é possível, finalmente, realizar tanto as análises dos agrupamentos, quanto as análises de previsão/sugestão de novos relacionamentos.

1.4 Trabalhos Relacionados

Neste tópico será feita uma síntese dos trabalhos que contribuíram com o desenvolvimento desta tese de doutorado. Entretanto, muitos outros trabalhos foram pesquisados e estão sendo citados no decorrer do desenvolvimento deste trabalho.

Como dito anteriormente, o crescimento acelerado das redes sociais se deve basicamente à evolução da Web. Esse crescimento tem atraído a atenção de muitos pesquisadores que pretendem analisar essas redes sociais. Muito trabalho tem sido feito na mineração de comunidades em páginas Web (Flake, Lawrence and Giles, 2000; Gibson, Kleinberg and Raghavan, 1998; Kumar, Raghavan, Rajagopalan and Tomkins, 1999) e em e-mails (Bird, Gourley, Devanbu, Gertz and Swaminathan, 2006; Schwartz and Wood, 1993; Tyler, Wilkinson and Huberman, 2003). Outros trabalhos incluem: mineração de grupos (Agrawal, Rajagopalan, Srikant and Xu, 2003; Han and Yan, 2011; Ichise, Takeda and Ueyama, 2005; Newman, 2004b) e previsão de relacionamentos (Acar, Dunlavy and Kolda, 2009; Clauset, Moore and Newman, 2008; Huang, Liben-Nowell and Kleinberg, 2003; Huang, Li and Chen, 2005; Huang and Lin, 2009; Liben-Nowell and Kleinberg, 2007; Lü and Zhou, 2009; Skillicorn, 2004); tecnologia das redes sociais e web semântica para compartilhar o conhecimento dentro da comunidade de engenharia de software (Dietrich, Dietrich and Wright, 2008); balanceamento das redes sociais, visando melhorar o fluxo de conhecimento na rede e, além disso, evitando que o fluxo seja interrompido caso algum elemento saia da rede social (Monclar, 2007).

A primeira etapa deste trabalho foi estudar os algoritmos de agrupamento. Newman (2004b), Han e Kamber (2006) examinaram vários algoritmos de agrupamento em redes sociais. Eles concluíram que ainda há muitas coisas a serem desenvolvidas, novos algoritmos para serem elaborados, e melhorias a serem feitas aos algoritmos que já existem.

A segunda etapa do trabalho foi construir e analisar a rede social multi-relacional. Existem muitos tipos de aplicações baseadas em análise de redes sociais, como redes escuras (Pioch, Barlos, Fournelle and Stephenson, 2005; Raab and Milward, 2003), sistemas de recomendação baseado no conteúdo (Golbeck, 2005; Huang, Li and Chen, 2005), previsão de relacionamentos (Farrel, Campbell and Mayagmar, 2005; Huang,

Liben-Nowell and Kleinberg, 2003; Lim, Negnevitsky and Hartnett, 2005; Zhu, 2003), as Redes Econômicas com interação cooperativa e não-cooperativa (Jackson, 2010), o uso da análise de redes sociais (SNA) para examinar as interações formais e informais em departamentos (Ryan and O'Connor, 2009), entre outros.

Nesta segunda etapa, foram construídas e analisadas as relações científicas dos pesquisadores brasileiros. Na etapa de modelagem da rede social, foram analisados diversos trabalhos que estudam redes sociais formadas por relações definidas por padrões de colaboração (Newman, 2000; Newman, 2001c; Newman, 2004a; Newman, 2004b). Algumas dessas obras analisam as redes sociais formadas por relações de co-autoria (Ichise, Takeda and Ueyama, 2005; Newman, 2000; Newman, 2001a; Newman, 2001b; Newman, 2004a), que é um tipo de relacionamento colaborativo.

Na etapa de modelagem da rede social científica foi importante o estudo do uso do tempo na análise dos relacionamentos. Neste contexto alguns trabalhos avaliam as mudanças estruturais da rede social (Leskovec, Kleinberg and Faloutsos, 2005; Potgieter, April, Cooke and Osunmakinde, 2009) e outros aplicam funções de penalização com base no ano do relacionamento para diminuir o peso do mesmo (Acar, Dunlavy and Kolda, 2009).

Além de examinar como os pesquisadores compartilham e trocam seus conhecimentos, na etapa de análise das comunidades científicas foram identificados alguns *centralizadores de conhecimento*, ou seja, pesquisadores que são extremamente importantes para manter o fluxo de conhecimento da rede (Cross, Parker and Cross, 2004).

Outro problema amplamente estudado no campo das redes sociais é a previsão de relacionamentos (Acar, Dunlavy and Kolda, 2009; Huang, Liben-Nowell and Kleinberg, 2003; Lü and Zhou, 2009; Potgieter, April, Cooke and Osunmakinde, 2009). Identificar as alterações nas redes sociais e analisar o comportamento futuro da rede através de informações sobre seu passado é chamado de *Previsão de Relacionamentos*. O problema da previsão de relacionamentos foi definido por Liben-Nowell & Kleinberg (2007) da seguinte forma: "Dado um retrato de uma rede social em um tempo *t*, procuramos prever com precisão as arestas que serão adicionadas à rede social durante o intervalo de tempo *t* para um tempo futuro *t*'" (Potgieter, April, Cooke and Osunmakinde, 2009).

A última etapa de análise deste trabalho é a sugestão/previsão de relacionamentos. A capacidade de prever mudanças nas relações antes que elas ocorram é altamente benéfica para uma organização. Alguns exemplos dos benefícios desta visão, entre outras coisas, são: identificar a estrutura de uma rede criminosa (Carpenter, Karakostas and Shallcross, 2002); superar o problema da dispersão dos dados em sistemas de recomendação usando filtragem colaborativa (Huang, Li and Chen, 2005); aumentar a velocidade de conexão entre as pessoas, pois um profissional ou pesquisador levaria mais tempo para formar a conexão por si mesmo (Farrel, Campbell and Mayagmar, 2005); ajudando a prever a propagação de uma entidade através de uma comunidade, tal como acontece com o vírus HIV (Zanette, 2002).

Em muitos problemas baseados nas redes sociais científicas apenas o relacionamento de co-autoria é analisado, mas não é difícil ver que essas redes sociais podem ter muitos outros tipos de relações científicas. Embora a análise de redes sociais multi-relacionais seja bastante interessante, poucos estudos têm sido feito nesta área (Cai, Shao, He, Yan and Han, 2005; Jung, Juszczyszyn and Nguyen, 2007; Ströele, Oliveira, Zimbrao and Souza, 2009).

A análise dos pontos levantados nesta seção (múltiplos relacionamentos, comunidades científicas, previsão e sugestão de relacionamentos, etc.) permite compreender como ocorre a comunicação e a colaboração entre os pesquisadores. Por isso, o estudo desses pontos é importante para a análise das redes sociais científicas.

1.5 Organização do Trabalho

Neste capítulo foram apresentados o objetivo, a motivação e as principais contribuições do desenvolvimento deste trabalho. No capítulo 2 são apresentados os principais conceitos envolvidos na análise das redes sociais. Nesse capítulo serão descritos as formas de representação das redes sociais utilizadas neste trabalho, bem como algumas de suas aplicações principais no mundo real.

O capítulo 3 descreve o processo de extração de conhecimento em grandes bases de dados. Esse processo é conhecido como KDD e a sua execução envolve várias etapas:

tais como: limpeza e transformação do conjunto de dados, técnica de mineração de dados, e formas de representação do conhecimento adquirido.

No capítulo 4 serão apresentadas as redes sociais científicas e a forma como as características dessas redes são modeladas por meio de grafos. Além disso, será apresentado o conceito de redes sociais multi-relacionais adotado no desenvolvimento deste trabalho. Esse capítulo contém a modelagem proposta para as redes sociais multi-relacionais e a etapa de pré-processamento do conjunto de dados. Nesse capítulo os dados são preparados para que a técnica de mineração de dados possa ser utilizada.

Os métodos de mineração de dados utilizados para a análise das redes sociais científicas multi-relacionais estão descritos no capítulo 5. Nesse capítulo são apresentados os dois métodos de agrupamento desenvolvidos neste trabalho, um baseia-se na árvore geradora mínima do grafo social e o outro utiliza informações sobre o fluxo máximo na rede social.

No capítulo 6 é apresentado um estudo de caso utilizado para validar e analisar os algoritmos desenvolvidos. Esse estudo de caso utiliza os dados de pesquisadores das instituições de ensino brasileiras.

O capítulo 7 faz uma análise dos estudos sobre previsão de relacionamentos em redes sociais. Serão apresentadas, nesse capítulo, diversas métricas adotadas na literatura para prever novos relacionamentos. Com base nessas métricas e na análise da evolução temporal da rede social científica será proposta uma nova métrica, composta por outras três métricas já conhecidas na literatura.

No capítulo 8 são apresentados os resultados obtidos para o estudo de caso deste trabalho. Nesse capítulo são feitas análises gráficas para avaliar o comportamento das métricas estudadas no capítulo 7.

Todas as análises visuais e temporais da rede social científica multi-relacional foram feitas com base em uma ferramenta desenvolvida neste trabalho. Essa ferramenta de visualização de redes sociais está descrita no capítulo 9.

Finalmente, no capítulo 10 são apresentadas as considerações finais e alguns trabalhos futuros.

Capítulo 2 – Análise das Redes Sociais

Neste capítulo serão apresentados os conceitos básicos sobre as redes sociais e a forma como as mesmas podem ser representadas. Além disso, serão enumeradas algumas das principais aplicações da análise das redes sociais.

2.1 Redes Sociais

O estudo de sócio-gramas, também conhecido por análise das redes sociais, envolve várias medidas computacionais de grafos, chamadas métricas, as quais provêem informações sociológicas úteis. Atualmente, as pesquisas em redes sociais combinam áreas diversas, tais como sociologia, antropologia, psicologia, geografia, estatísticas, e ciência da computação (McMahon, Miller and Drake, 2001).

O tipo de uma rede social é definido pelo contexto que essa rede representa. Geralmente, o tipo de uma rede social está diretamente ligado ao tipo de relacionamento envolvido nessa rede. Assim, diferentes tipos de relacionamentos produzem deferentes tipos de redes sociais, mesmo quando o conjunto de observações é mantido sem alteração. Por exemplo, uma rede na qual os atores são pessoas pode ter conceitos de amizade, co-autoria ou vínculo empregatício variando apenas o tipo do relacionamento da rede, mesmo que o conjunto de pessoas seja mantido o mesmo.

Embora seja importante analisar como os elementos das redes sociais se relacionam a análise das redes sociais não está interessada apenas em descrever como os conjuntos de atores estão ligados uns aos outros. Essa análise propõe que, como a estrutura da rede afeta tanto os indivíduos como o sistema como um todo, ela pode explicar as variações nas estruturas relacionais e as suas conseqüências.

Todo projeto de análise das redes sociais deve iniciar com o estudo de três elementos principais: entidade básica, forma e conteúdo dos relacionamentos, e nível de análise dos dados (Wasserman and Faust, 1994).

A definição das entidades básicas consiste em decidir quais entidades representam os atores da rede social. Essas entidades, como dito anteriormente, podem ser pessoas,

grupos formais ou informais, organizações formais complexas, comunidades, etc. Nas redes sociais científicas, que é o estudo de caso deste trabalho, os pesquisadores são as entidades básicas que representam a rede. Entretanto, os grupos de pesquisadores e as instituições de ensino também podem ser considerados entidades das redes sociais científicas.

Além das entidades, é importante definir a forma e o conteúdo dos relacionamentos. Esses dois elementos são inseparáveis e se distinguem apenas analiticamente. O *conteúdo* é o interesse, proposta, ou motivo da relação existir entre os indivíduos. A *forma* é a maneira pela qual o conteúdo se expressa.

Em uma rede social de empresas, por exemplo, pode haver um interesse de conteúdo econômico, sendo que esse conteúdo pode se expressar na forma de competição ou de colaboração. A rede social científica é formada por relacionamentos de troca de conhecimento científico e que, no presente trabalho, se dá sob a forma de colaboração entre os pesquisadores.

Após definir as entidades e seus relacionamentos é importante definir o nível de análise que será feita nos dados que foram coletados. Essa etapa que irá definir o nível de profundidade da análise da rede social.

A análise mais simples é a *análise egocêntrica* da rede. Nesse tipo de análise apenas um elemento é analisado de cada vez, de forma que caso existam **N** elementos na rede social serão feitas **N** unidades de análise. Durante a análise de cada ator são analisados o número, intensidade, e outras características de seus relacionamentos específicos.

O outro nível de análise é a *análise diádica* (*dyadic network*), que consiste na análise de pares de autores. A principal questão nesse tipo de análise é saber se um relacionamento específico existe entre um par de atores e, caso a ligação entre eles exista, o objetivo é identificar qual a intensidade, duração, ou força desse relacionamento.

O terceiro nível de análise das redes sociais é a *análise triádica* (*triadic relations*). Nesse tipo de análise é avaliada a intensidade das relações em um grupo de três entidades. Assim, supondo que A seja amigo de B e B seja amigo de C, é analisada a tendência de A se tornar amigo de C. Esse é um exemplo muito comum de análise

triádica e que se tornou base para o estudo de algumas métricas para previsão de novos relacionamentos.

As análises egocêntricas, diádicas e triádicas são aplicadas em estudos específicos de análise de redes sociais. Como essas análises são voltadas para alguns elementos pontuais elas são inviáveis se serem aplicadas para todos os elementos da rede social quando a mesma possui um número de elementos muito grande.

A análise mais importante é a *análise completa* da rede social. Todas as informações de todos os relacionamentos de todos os atores são utilizadas a fim de explicar a estrutura relacional da rede social como um todo. Assim, nessa análise todas as informações da rede social são usadas na análise, tornando a análise da rede social muito mais completa e também complexa.

É importante observar que cada tipo de análise implica em um tipo de inferência, sendo necessário saber qual o tipo de informação se deseja extrair para se definir a melhor estratégia de análise. Neste trabalho será feita uma análise completa sobre as redes sociais científicas, pois todas as informações dos relacionamentos são utilizadas nesse processo.

Existem várias referências padrões para o estudo das redes sociais, utilizadas por vários artigos dessa área, nas quais a análise está baseada em dados estáticos. A dificuldade de coletar dados sociais numerosos para permitir um estudo dinâmico das redes sociais sempre foi um empecilho.

Entretanto, com o contínuo crescimento do uso de computadores, coletar dados suficientes para criar grafos numerosos em intervalos de tempo tornou-se possível. Um exemplo é criar um grafo por semana de e-mails enviados, usando o remetente e o destinatário dos e-mails trocados para construir o grafo temporal (Campbell, Maglio, Cozzi and Dom, 2003). Essa seqüência de grafos pode ser usada para estudar a evolução da rede e as mudanças que ocorreram ao longo do tempo sobre a perspectiva de várias métricas.

A representação formal da rede social é um fator importante na análise da mesma. As análises que podem ser feitas sobre as redes sociais estão diretamente relacionadas à representação delas. Na próxima seção serão apresentadas as duas formas de representação das redes sociais que foram utilizadas neste trabalho.

2.2 Representação das Redes Sociais

Como a análise das redes sociais está diretamente relacionada à representação de seus dados é importante avaliar as opções existentes para representá-las. Existem basicamente dois conceitos de representação de uma rede social. Um bastante voltado para a visualização, que á a representação em grafos, e outro voltado para a manipulação dos dados, que é a representação dos dados por matrizes.

Essas duas formas de representação serão utilizadas neste trabalho. As matrizes serão usadas pelo algoritmo de agrupamento no processo de identificação das comunidades de pesquisa. Por outro lado, os grafos serão usados na etapa de visualização e análise dos resultados, já que a forma de representação em grafos permite uma visualização mais amigável da Rede Social Científica.

2.2.1 Grafos

Um grafo é uma estrutura $G = (X, U)$ onde X é um conjunto discreto (finito e enumerável) e U é um conjunto de linhas que representam qualquer tipo de ligação entre os elementos do conjunto X (Boaventura, 1996). Uma ligação l (onde $l \in U$) é um conjunto de dois nós do conjunto X. Por exemplo, considere os conjuntos $X = \{x_1, x_2, x_3, x_4, x_5\}$ e $U = \{l_1, l_2\}$ referentes ao grafo desconexo representado na Figura 2. O conjunto U também pode ser escrito como $U = \{\{x_1, x_5\}, \{x_3, x_5\}\}$, já que $l_1 = \{x_1, x_5\}$ e $l_2 = \{x_3, x_5\}$.

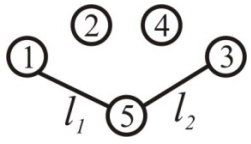


Figura 2 - Exemplo de Grafo

Em um grafo os elementos de X são representados por nomes, letras ou números. Já as linhas do conjunto U, que são as ligações entre os atores do grafo, são desenhadas apenas quando existe a ligação entre os elementos, caso a ligação não exista a linha não é representada.

Dois nós são ditos adjacentes se existir uma ligação direta entre eles. Outro conceito em grafo é que um nó é incidente em uma linha se ele for um dos nós definidos ou como origem ou como destino da linha.

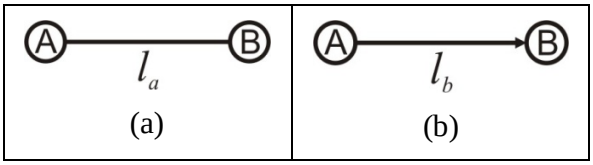


Figura 3 - Tipos de Arestas no Grafo.

Diz-se que os nós *A* e *B* são adjacentes e que esses nós são incidentes na ligação l_a, conforme pode ser observado na Figura 3(a). As ligações em um grafo podem ser não direcionadas ou direcionadas, como representado nas relações l_a e l_b, respectivamente. No caso das relações não direcionais os dois atores cooperam com aquele tipo de relação. Por exemplo, se, na Figura 3(a), a relação l_a é uma relação de confiança, então o elemento *A* confia no elemento *B* e o elemento *B* confia no elemento *A*.

Por outro lado, nos grafos direcionados apenas o elemento oposto à seta é o que exerce a relação. Assim, caso os dois elementos tenham a relação é necessário representar uma linha com duas setas. No exemplo da Figura 3(b), considerando que l_b também seja uma relação de confiança, então o elemento *A* confia no elemento *B*, mas o elemento *B* não confia no elemento *A*.

Nas redes sociais científicas os relacionamentos são todos não direcionados. Isso se deve ao fato dos trabalhos científicos serem desenvolvidos de maneira colaborativa, de forma que sempre ocorre a participação dos dois pesquisadores envolvidos no relacionamento.

Um grafo pode ser multi-relacional, indicando diferentes tipos de relacionamentos entre seus objetos. Esses relacionamentos podem ser representados através de desenhos diferentes. Por exemplo, em uma Rede Social Científica podem ser desenhadas linhas contínuas para identificar os relacionamentos de co-autoria entre os pesquisadores e usar linhas pontilhadas para representar os relacionamentos de participação no desenvolvimento do mesmo projeto.

Existem vários conceitos em grafos que ajudam na análise de Redes Sociais. Essas definições são facilmente representadas em pequenos grafos e podem ser visualizadas na Figura 4. Todas as definições que seguem foram baseadas nos conceitos de Boaventura (Boaventura, 1996).

Um percurso, conceito mais abrangente de um grafo, é uma família de ligações sucessivas adjacentes. O percurso será fechado se a última ligação da sucessão for adjacente à primeira, e aberto caso contrário. Na definição geral, despreza-se implicitamente a orientação das ligações, quando se trata de um grafo orientado, ou seja, na conceituação de percurso interessa apenas a adjacência sucessiva. A Figura 4 ilustra exemplos de percursos abertos (b, d, f) e fechados (a, c, e).

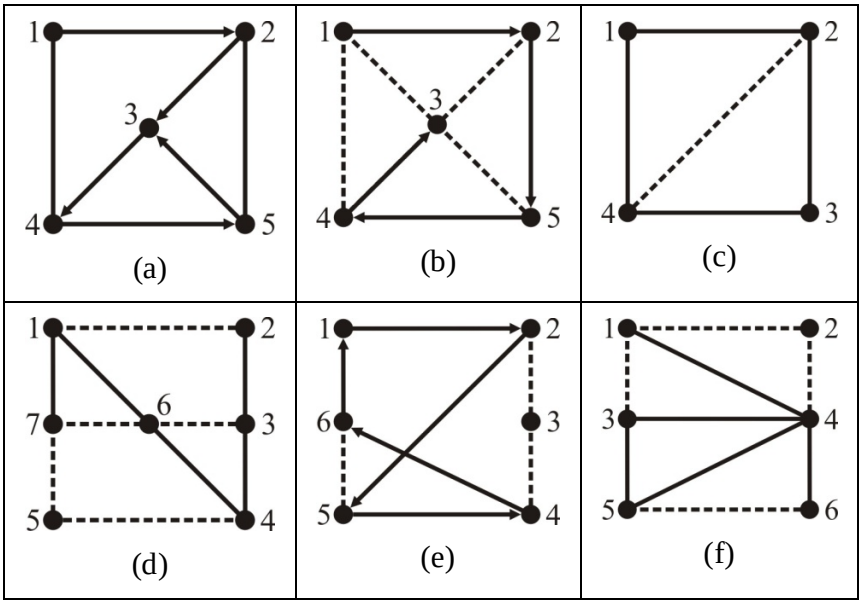


Figura 4 - Percursos (Boaventura, 1996).

Em um grafo não valorado, o comprimento de um percurso é o número de ligações por ele utilizadas, contando-se as repetições. No caso de grafos valorados, será necessária a generalização trazida pela noção de distância. O cálculo da distância entre dois nós é dependente da métrica utilizada em sua definição. Embora existam diferentes conceitos para o cálculo de distância em grafos, o mais comumente utilizado é a soma dos pesos das arestas.

Um ciclo é uma cadeia simples e fechada, representada pela Figura 4(c). No ciclo não é considerado a orientação das arestas. Já um caminho é uma cadeia em um grafo

orientado, na qual a orientação dos arcos é sempre a mesma, a partir do vértice principal (Figura 4(b)). O conceito de caminho em grafos será bastante utilizado na definição das métricas no estudo da previsão de relacionamentos.

Um circuito é um caminho simples e fechado em um grafo orientado, ou seja, um ciclo no qual a orientação dos arcos é sempre a mesma a partir de um vértice qualquer que seja utilizado como origem. Na Figura 4(a), tem-se como exemplo de circuito os elementos (3, 4, 5, 3).

Um grafo é dito fortemente conectado quando todo par de nós está conectado por um caminho direcionado em ambas as direções. O grafo é unilateralmente conectado quando todos os pares são ligados por um caminho em apenas uma direção. Finalmente, um grafo é dito não - fortemente conectado quando todos os pares estão unidos, mas não através de caminhos.

Dificilmente, um grafo definido por objetos do mundo real será fortemente conectado. Entretanto, existirão subgrafos nesses grafos que possuem, isoladamente, essa propriedade. Encontrar esses subgrafos significa encontrar comunidades de objetos afins, ou seja, que compartilham de um interesse comum. Se, por exemplo, os relacionamentos de um grafo representam amizades dos elementos, os subgrafos fortemente conectados iriam representar grupos de amigos comuns, pessoas que possuem maior grau de intimidade.

Um nó, de um grafo conectado, é um centralizador se a sua remoção implica na desconexão do grafo, ou seja, ao removê-lo são gerados dois ou mais subgrafos não conectados entre si (Opsahla, Agneessensb and Skvoretzc, 2010; Pozo, Manuel, González-Arangüena and Owen, 2011).

Da mesma forma, uma aresta é chamada de ponte se a sua retirada do grafo produzir o mesmo efeito no grafo, ou seja, se ao ser removida for gerado um grafo desconexo (Valente and Fujimoto, 2010).

As Redes Sociais multi-relacionais podem ser representadas através de grafos com múltiplas arestas, no caso das redes com um único tipo de objeto. Uma maneira é representar cada tipo de relacionamento em um grafo diferente. Assim, caso existam três tipos de relacionamentos na rede social serão desenhados três grafos, cada um deles representando um tipo de relacionamento.

As Redes Sociais Multi-Relacionais que possuem apenas um tipo de elemento também podem ser representadas por um único grafo, no qual todos os relacionamentos entre os objetos são representados em uma mesma imagem, conforme ilustrado na Figura 5. Embora seja mais simples a análise para redes sociais pequenas, a visualização torna-se bastante complexa na análise de redes sociais maiores.

A alternativa adotada neste trabalho para ilustrar a rede social multi-relacional foi utilizar cores nas arestas. Os conceitos adotados no desenvolvimento da ferramenta de visualização de redes sociais multi-relacionais estão descritos em detalhes no capítulo 9.

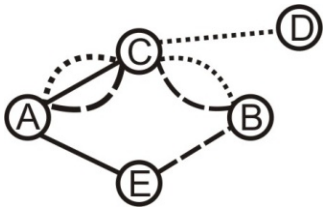


Figura 5 - Grafo com múltiplas arestas

Finalmente, para o caso das redes sociais com vários tipos de relacionamentos e objetos pode ser utilizado o conceito de grafo bipartido (Wasserman and Faust, 1994). Um grafo bipartido é aquele no qual os nós são divididos em dois grupos distintos, onde cada grupo representa um tipo diferente de entidade. Um modelo de grafo bipartido está ilustrado na Figura 6.

Por exemplo, pode ser criado um grafo bipartido no qual os dois conjuntos de elementos representam pesquisadores e artigos publicados. As arestas entre os dois grupos do grafo representam o relacionamento de autoria. Já as arestas entre os pesquisadores representam o relacionamento de co-autoria. Por fim, as arestas entre os artigos representam o relacionamento de citação.

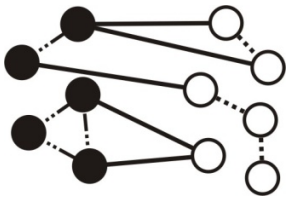


Figura 6 - Grafo Bipartido com múltiplas arestas

O principal conceito envolvido na representação em grafos das redes sociais é o layout das mesmas. O layout mais popular é o *force-directed* (força dirigida), no qual as arestas do grafo atraem os nós para que os elementos relacionados fiquem próximos uns dos outros (Spritzer, 2009).

Pode ser observado que a definição do layout para a visualização das redes sociais é de fundamental importância. Os conceitos envolvidos na definição do layout da rede social científica estão no capítulo 9.

2.2.2. Matrizes

Uma representação algébrica dos relacionamentos da rede social pode expressar todas as informações quantitativas incorporadas no sociograma. Além disso, permite um conjunto muito maior de análises do que é possível com a correspondente representação visual. A forma básica de representação de uma rede social para a análise matemática é a *matriz de adjacência*.

Geralmente, essas matrizes são representadas por **0** ou **1**, o **0** indica a inexistência de relacionamento e o **1** indica a existência de relacionamento. Uma alternativa de representação é utilizar **0** para indicar a ausência de relacionamento e valores positivos para indicar o peso do relacionamento entre dois elementos, equação (1). Assim, é possível identificar os relacionamentos mais fortes na rede. A Figura 7 ilustra as situações descritas acima.

A matriz de adjacência é definida matematicamente da seguinte forma:

$$A_{nxn} = \begin{cases} x_{ij} = 0, & se\ j \notin \Gamma^{+}(i) \\ x_{ij} > 0, & se\ j \in \Gamma^{+}(i) \end{cases} \quad (1)$$

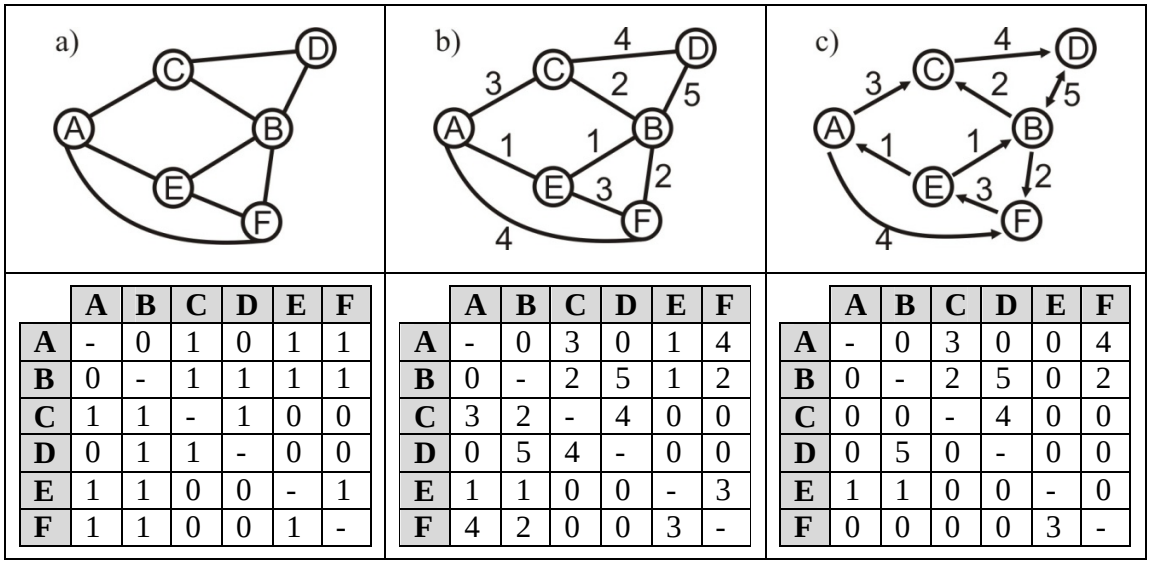
onde $\Gamma^{+}(i)$ representa o conjunto de elementos sucessores ao elemento *i*. Por exemplo, baseado na Figura 7(a) tem-se $\Gamma^{+}(A) = \{C, E, F\}$, por outro lado, baseado na Figura 7(c) $\Gamma^{+}(A) = \{C, F\}$, devido a orientação do grafo.

Na maioria dos estudos a diagonal principal não possui nenhum significado e por isso é ignorada, como representado em todas as tabelas na Figura 7. Uma pessoa, por exemplo, não será a melhor amiga dela mesma. Entretanto, alguns autores podem utilizar

a diagonal principal para indicar alguma informação pessoal do elemento. Por exemplo, a diagonal pode indicar o total de amigos de uma pessoa.

O valor numérico em uma célula da matriz de adjacência representa o peso de um relacionamento específico entre o par de atores associados pela linha e coluna correspondente. Por convenção, para relações direcionadas, os atores nas linhas são os que exercem o relacionamento e os atores das colunas são os receptores dessa relação.

No caso das redes sociais não direcionadas serão geradas matrizes simétricas, já que o elemento que exerce também recebe o mesmo relacionamento, Figura 7(a) e (b). Por outro lado, as redes sociais direcionadas, geralmente, possuem matrizes não simétricas, já que o elemento que exerce o relacionamento nem sempre o recebe e, caso receba, isso pode ocorrer com uma intensidade diferente, como representado na Figura 7(c).



	A	B	C	D	E	F
A	-	0	1	0	1	1
B	0	-	1	1	1	1
C	1	1	-	1	0	0
D	0	1	1	-	0	0
E	1	1	0	0	-	1
F	1	1	0	0	1	-

	A	B	C	D	E	F
A	-	0	3	0	1	4
B	0	-	2	5	1	2
C	3	2	-	4	0	0
D	0	5	4	-	0	0
E	1	1	0	0	-	3
F	4	2	0	0	3	-

	A	B	C	D	E	F
A	-	0	3	0	0	4
B	0	-	2	5	0	2
C	0	0	-	4	0	0
D	0	5	0	-	0	0
E	1	1	0	0	-	0
F	0	0	0	0	3	-

Figura 7 - Representações em Grafos e suas Matrizes Correspondentes.

2.3 Aplicações da Análise das Redes Sociais

A análise de Redes Sociais é uma aplicação do campo da ciência das redes sociais para estudar as conexões e relacionamentos humanos (Hansen, Shneiderman and Smith, 2010). Pode-se dizer que nos dias atuais as pessoas já nascem conectadas a outras através das redes sociais, quando seus pais criam para seus filhos perfis no Orkut, *Twiter*,

Facebook (Hansen, Shneiderman and Smith, 2010) e os relacionam com pessoas que eles irão conhecer futuramente.

A análise das redes sociais ajuda na descoberta de padrões em coleções de pessoas conectadas por diversos tipos de relacionamentos. Esses estudos estão voltados para a análise dos relacionamentos entre as pessoas e não para as características pessoais de cada indivíduo.

Assim, enquanto os métodos tradicionais de análise social avaliam os indivíduos e seus atributos, tais como: sexo, idade e renda, os estudos das redes sociais focam suas análises nas conexões que fazem com que os indivíduos se relacionem e troquem informações (Hansen, Shneiderman and Smith, 2010).

Nas próximas seções serão apresentadas algumas aplicações da análise de redes sociais. Embora exista uma infinidade de aplicações, neste trabalho, serão apresentados os principais problemas de análises de redes sociais.

2.3.1 Análise da estrutura organizacional

Através da perspectiva da análise das redes sociais, os organogramas que geralmente representam a estrutura hierárquica das organizações são extremamente simples e não possuem informações importantes sobre conexão que existem entre os departamentos e divisões (Cross, Parker and Cross, 2004).

Muitas vezes as pessoas estabelecem conexões que fogem da estrutura hierárquica da organização. Pode existir uma relação de amizade entre duas pessoas de setores diferentes da empresa; Uma pessoa, por não ter muita afinidade com seu chefe direto, se comunica diretamente com os chefes na hierarquia; Duas pessoas que deveriam estar em constante comunicação e definições não se relacionam bem, podendo impactar no processo da empresa; dentre várias outras possibilidades.

Esses relacionamentos extra-organizacionais não estão definidos na estrutura hierárquica da empresa, mas identificá-los pode ser muito útil para melhorar o fluxo dos processos da empresa. Assim, gestores que traçam estratégias para aprimorar a comunicação na rede social da empresa podem rapidamente melhorar a eficiência dentro de sua organização.

Um dos benefícios que são rapidamente identificados ao se analisar a rede social da organização vem da descoberta de relacionamentos excessivos, ou seja, pessoas que recebem uma demanda muito grande por parte de outros funcionários. Essa descoberta pode auxiliar os gestores a desenvolverem caminhos alternativos para aliviar as pessoas sobrecarregadas e reduzir o tempo de espera pela realização do trabalho que essas pessoas desempenham.

2.3.2 Busca em Redes Sociais

Adâmica, Lukose e Huberman investigaram a busca em redes de escala livre, tais como sistemas peer-to-peer ou redes sociais (Freeman, 1979). No processo de busca definidos por eles, para que um nó seja encontrado, todos os seus vizinhos também foram incluídos no processo de pesquisa.

Essa idéia é semelhante à forma como os participantes da experiência de Milgram (Milgram, 1974) tentaram passar uma mensagem por toda uma rede social, conhecendo apenas os seus vizinhos imediatos. O objetivo dessa experiência foi fazer com que uma mensagem chegasse até um nó de destino permitindo que o nó de origem passasse a mensagem apenas para as pessoas que ele conhecesse. Na segunda etapa cada nó transmite a mensagem para os seus vizinhos e assim sucessivamente.

Na busca em redes sociais a idéia é a mesma. Ao procurar por um determinado nó (um perito, ou um computador com um determinado arquivo) a mensagem de solicitação da busca deve ser transmitida de nó em nó. Assim, não é necessário transmitir a mensagem para a rede inteira e, na maioria dos casos, o elemento que está sendo procurado é encontrado rapidamente.

2.3.3 Redes Escuras/Brilhantes

Uma rede escura é uma rede que opera secretamente e ilegalmente. Exemplos desse tipo de rede incluem organizações políticas reprimidas em países governados por ditadores, terroristas, traficantes e as organizações de tráfico de drogas. Essas redes são

de grande interesse dos órgãos governamentais, pois o estudo da mesma pode, por exemplo, prevenir um ataque terrorista.

Analistas de inteligência do Governo usam previsão de relacionamentos, em conjunto com o passado das atividades criminosas, os potenciais locais de risco e técnicas de visualização para prever com o máximo de detalhes e precisão as futuras atividades criminosas de uma rede escura (Pioch, Barlos, Fournelle and Stephenson, 2005).

Por outro lado, uma rede brilhante é uma rede que opera abertamente e legalmente, porém com políticas de segurança das informações. Exemplos dessas redes incluem as organizações policias e militares, órgãos do governo e grandes empresas, que protegem suas informações dos concorrentes.

Raab e Milward (2003) encontraram nas análises em muitas instâncias de redes escuras que a topologia dessas redes é tão variada como a das redes brilhantes. Embora tanto as redes escuras quanto as brilhantes enfrentem problemas similares por operarem secretamente, elas competem em diversas formas (Raab and Milward, 2003).

2.3.4 Recomendação em Sistemas de Conteúdo

Huang, Li e Chen pesquisaram o uso da previsão de relacionamentos em sistemas de recomendação colaborativa (Huang, Li and Chen, 2005). Um exemplo de um sistema de recomendação é a *Amazon.com*, que recomenda livros semelhantes ao que a pessoa está comprando, os quais outros compradores do livro tenham comprado. A *Amazon.com* é também um exemplo de filtragem colaborativa, pois as compras de seus usuários vão mudando de forma colaborativa o ranking de preferências dos livros no sistema.

A Filtragem Colaborativa agrupa inicialmente os usuários em grupos similares, com base em suas preferências e, em seguida, recomenda a um usuário os itens preferidos dos seus vizinhos. Porém esses sistemas sofrem de dois problemas. O primeiro problema é que nada pode ser recomendado quando o sistema é usado pela primeira vez, já que ninguém optou por nenhum item ainda.

O segundo problema é que os dados são esparsos, pois poucos itens serão recomendados enquanto apenas alguns usuários tiverem escolhidos poucos itens. Essa questão é abordada por Huang realizando previsão de relacionamentos no grafo para

prever quais os itens poderiam ser selecionados pelos usuários e, assim, aumentar a densidade do grafo.

Golbeck também pesquisou sobre a recomendação de conteúdo (Golbeck, 2005). O sistema dela utiliza os usuários de confiança para relacioná-los e, dessa forma, reforçar as recomendações. Ela também tem usado as redes sociais para melhorar a filtragem e classificação dos e-mails, incluindo a filtragem colaborativa de spam (Golbeck and Hendler, 2004).

2.3.5 Marketing

Empresas como *Amazon* e *Yahoo* estão tentando descobrir formas de publicidade para clientes que incluem referências a partir de uma fonte confiável. Por exemplo, um anúncio seria exibido em uma página web que apóia um restaurante local, utilizando uma resenha escrita por alguém por perto na rede social do leitor.

O objetivo das grandes empresas é fazer a propaganda certa, na hora certa e para a pessoa certa. Dessa maneira, as empresas conseguiriam fazer propagandas personalizadas baseando-se no gosto e nas características de cada pessoa. Essa aplicação está ganhando ênfase à medida que a publicidade vem demonstrando sua importância e eficiência quando aplicada de maneira adequada.

2.3.6 Aplicações específicas em Previsão de Relacionamentos

A capacidade de prever as mudanças nos relacionamentos antes que eles ocorram é altamente benéfica para uma organização. Alguns exemplos das vantagens dessa visão social privilegiada estão listados a seguir.

Um dos problemas que ganhou grande destaque foi a *análise de redes terroristas*. Nesse tipo de rede social, como dito anteriormente, os relacionamentos não são apresentados de maneira explícita. Assim, identificar a estrutura de uma rede criminosa torna-se um problema complexo, pois o conjunto de dados desse tipo de rede social está sempre incompleto.

Prever as conexões que faltam nessas redes é de extrema importância para os governos que lutam contra o terrorismo. Através dessas descobertas eles podem se prevenir contra ataques futuros e prender os criminosos (Carpenter, Karakostas and Shallcross, 2002).

Além dos problemas de análise em redes criminosas, *prever as páginas* que os usuários da *web* irão visitar também tem recebido bastante atenção dos pesquisadores. Saber de antemão qual ou quais serão os próximos cliques do usuário pode melhorar a eficiência e eficácia da navegação de um site (Zhu, 2003).

Em casos de sites que recebem milhares de acessos por dia, esse tipo de previsão pode torná-lo mais rápido. Entretanto, no caso da navegação em páginas Web, o site pode não saber exatamente para onde o usuário deseja ir, mas caso ele tenha boas sugestões o usuário pode aceitá-las e, conseqüentemente, ele ficará mais satisfeito com os serviços prestados.

Existem problemas de análise de redes sociais voltados para a *prevenção da propagação de uma entidade* através da rede social. Esses problemas podem ser aplicados a cenários diferentes, tais como a prevenção da disseminação de doenças contagiosas, reter a propagação de boatos (Zanette, 2002), monitorar e controlar vírus de computadores que usam e-mails como meios de contaminação (Lim, Negnevitsky and Hartnett, 2005), dentre outros.

Superar o problema da *dispersão de dados* em sistemas de recomendação através de filtragem colaborativa (Huang, Li and Chen, 2005) e melhorar a *análise de hipertexto* para recuperação de informação e mecanismos de busca (Henziger, 2000), também são exemplos de problemas que podem ser solucionados através de técnicas de previsão de conexões.

Outra questão importante no estudo de muitos tipos de redes sociais é a *troca de conhecimento* na rede. Muitas vezes os novos relacionamentos entre as pessoas ou organizações que compõem a rede social demoram a serem estabelecidos. A previsão de relacionamentos pode ser utilizada para acelerar uma conexão mutuamente benéfica.

A previsão de relacionamentos com o intuito de sugerir novos relacionamentos pode ser utilizada tanto em redes de relacionamentos profissionais quanto em redes acadêmicas (Farrel, Campbell and Mayagmar, 2005).

Neste trabalho será desenvolvido um módulo de sugestão de relacionamentos para facilitar e melhorar a comunicação da rede social científica multi-relacional. Para tal, foi realizado um estudo da evolução temporal da rede social para que novos relacionamentos pudessem ser sugeridos.

Capítulo 3 – Processo de Extração de Conhecimento

Para que fosse possível analisar a rede social científica estudada neste trabalho foi necessário um estudo detalhado dos dados que compõem essa rede. Todas as análises feitas e as conclusões obtidas sobre as redes sociais científicas foram feitas com auxílio do processo de extração de conhecimento.

O crescimento da quantidade de informação disponível é considerado um fator de impacto no desenvolvimento de técnicas de descoberta de conhecimento. Como dito anteriormente, com o passar dos anos, milhares de informações vêem sendo armazenadas em diversos bancos de dados. Muitas vezes, essas informações não trazem nenhum benefício para quem as possui, devido à dificuldade em se extrair informações úteis em um conjunto de dados com volumes de informações tão grandes.

A Mineração de Dados, que é uma das etapas do processo de extração de conhecimento, surgiu com o intuito de auxiliar a análise de grandes conjuntos de dados. Pode-se dizer que a Mineração de Dados tem como objetivo principal *extrair novos conhecimentos que estão escondidos em grandes bases de dados* (Han and Kamber, 2006).

Neste capítulo serão apresentados os conceitos básicos do processo de extração de conhecimento e cada uma de suas etapas. No final deste capítulo será descrito em qual fase do desenvolvimento deste trabalho cada uma dessas etapas se encaixam.

3.1 Introdução

A mineração de dados pode ser vista como um resultado da evolução natural da tecnologia da informação. Os sistemas de bancos de dados estão cada vez mais evoluídos e continuam desenvolvendo novas funcionalidades que facilitam a manipulação de conjuntos de dados de volumes diversos.

Os sistemas de gerenciamento de banco de dados possuem componentes que auxiliam o gerenciamento de transações, a recuperação da informação, e a análise dos dados. Assim, os documentos que antes eram armazenados de maneira não muito

eficiente passaram a serem armazenados de forma organizada nessas mídias por meio dos sistemas de bancos de dados.

Embora os sistemas de bancos de dados tenham evoluído e ainda continuem melhorando a cada dia, o crescimento constante do volume de dados disponibilizados pelos meios de comunicação, como a Internet, é um fator impeditivo para que esses dados sejam analisados sem a utilização de outras técnicas de análise de dados. É difícil obter informações relevantes a respeito de um assunto específico nestes bancos de dados e essa é uma das motivações para o estudo de técnicas de extração de conhecimento ou, como são popularmente conhecidas, técnicas de mineração de dados.

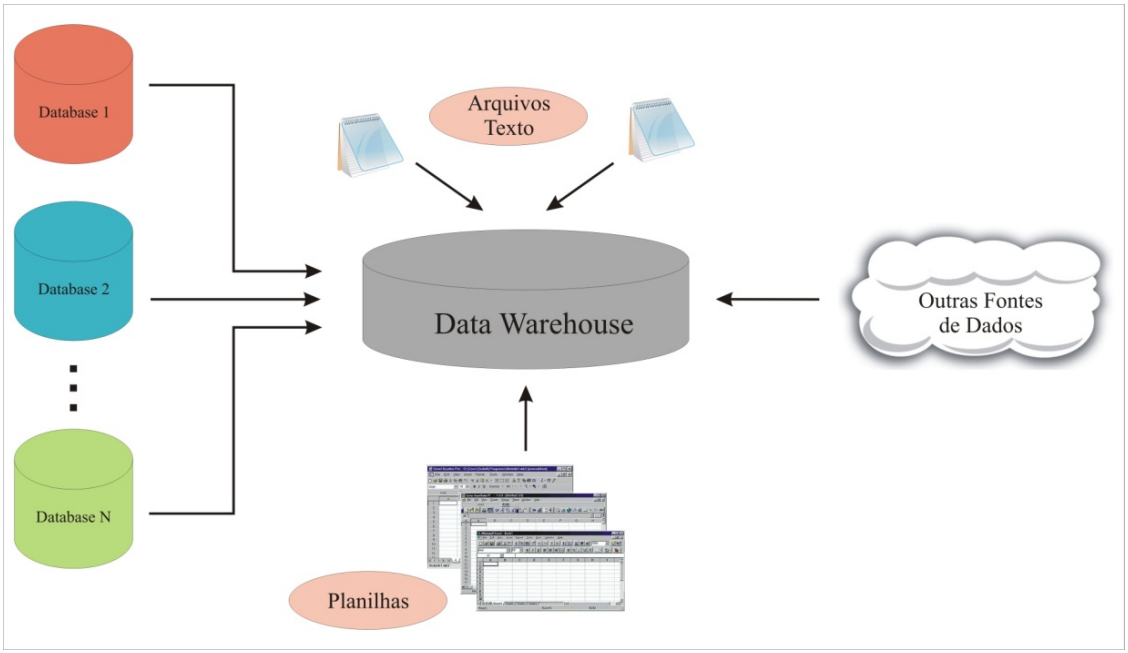


Figura 8 - *Data Warehouse.*

A abundância de dados, juntamente com a necessidade de poderosas ferramentas de análise de dados, tem sido descrito como conjunto de dados ricos, mas pobres de informação. Como resultado, os dados coletados em grandes repositórios de dados tornam-se “túmulos de dados”, ou seja, arquivos de dados que raramente são visitados (Han and Kamber, 2006).

Uma arquitetura de repositório de dados que surgiu para auxiliar as técnicas de mineração de dados é o *Data Warehouse* ou *Armazém de Dados*, que é um repositório de

múltiplas fontes de dados heterogêneos organizados no âmbito de um esquema unificado e consolidado em um único local. Esse repositório de dados está representado de maneira abstrata na Figura 8. Como o termo *Data Warehouse* é mais popular que a sua tradução para o português, neste trabalho será utilizada sua nomenclatura original em inglês.

O formato desses repositórios de dados favorece a extração de relatórios, a análise de grandes volumes de dados e a obtenção de informações estratégicas que podem facilitar a tomada de decisão gerencial.

O *Data Warehouse* possibilita a análise de grandes volumes de dados, coletados dos sistemas transacionais. São as chamadas séries históricas que possibilitam uma melhor análise de eventos passados, oferecendo suporte às tomadas de decisões no presente e a previsão de eventos para o futuro.

Por definição, os dados em um *Data Warehouse* não são voláteis, ou seja, eles não mudam, salvo quando é necessário fazer correções nos dados que foram previamente carregados. Os dados estão disponíveis no repositório somente para leitura e não podem ser alterados.

Geralmente, os *Data Warehouses* são construídos para facilitar a etapa de mineração de dados. Entretanto, neste trabalho, não houve a necessidade do desenvolvimento de um *Data Warehouse* para aplicar as técnicas de agrupamento, pois os dados utilizados neste trabalho já tinham sido carregados em um banco de dados único. Porém, para facilitar a análise evolutiva da rede foi construída uma tabela única que contém todos os dados importantes para o módulo de sugestão de relacionamentos. Os detalhes sobre a estrutura dessa tabela estão no capítulo sobre previsão de relacionamentos.

Assim como os outros tipos de dados, o volume de dados científicos também está aumentando rapidamente. O volume desses dados cresce anualmente surgindo a necessidade do desenvolvimento de uma ferramenta que extraia informações úteis, ou seja, que extraia conhecimento desses dados. As principais características do conjunto de dados utilizado no desenvolvimento deste trabalho, bem como o estudo de caso serão descritos em detalhes nos capítulos seguintes.

O processo de extração de conhecimento de grandes repositórios de dados independe do tipo de problema a ser solucionado ou da informação que se deseja

encontrar. Esse processo tem etapas bem definidas que levam à descoberta de conhecimento e se aplica em cenários variados. O processo de extração de conhecimento é conhecido como KDD, que é a “Descoberta de Conhecimento a partir dos Dados”, termo que vem do inglês *Knowledge Discovery from Data*.

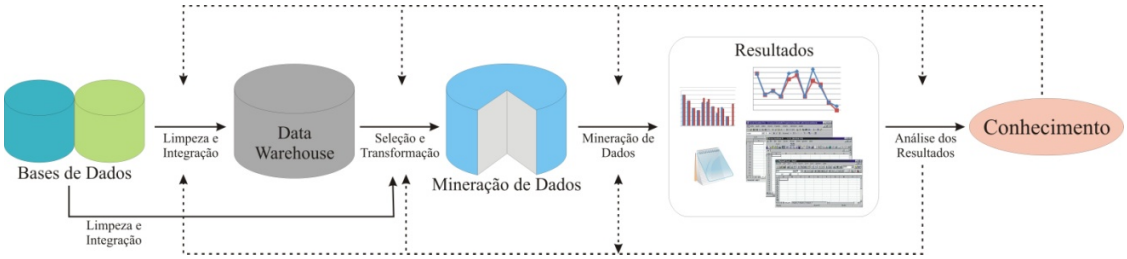


Figura 9 - Processo KDD (Han and Kamber, 2006).

KDD é um processo de extração de informações úteis em bases de dados, no qual a descoberta de conhecimento é a sua última etapa. O processo KDD está representado na Figura 9 e consiste de uma seqüência iterativa dos seguintes passos: limpeza, integração, seleção e transformação dos dados, mineração de dados, avaliação dos padrões e apresentação do conhecimento.

Cada um desses passos está representado por setas com linhas não pontilhadas. É importante observar nesta figura que a etapa de Limpeza e Integração pode ser seguida diretamente pela etapa de Seleção e Transformação dos dados sem a necessidade da criação de um *Data Warehouse*. Esse fluxo alternativo foi adotado neste trabalho já que, como dito anteriormente, as técnicas de mineração de dados foram aplicadas diretamente ao banco de dados.

As setas com linhas pontilhadas indicam iteração de retorno, ou seja, em cada etapa do processo KDD é possível voltar às etapas anteriores para possíveis ajustes. Uma iteração de retorno muito freqüente é entre a análise de resultados e a etapa de mineração de dados, pois muitos ajustes são necessários até que os resultados apresentem informações úteis.

De maneira resumida, esses passos do processo de descoberta de conhecimento a partir dos dados são definidos da seguinte maneira:

1. *Limpeza dos Dados*: remoção de ruídos e inconsistências nos dados.

2. *Integração dos Dados*: união de conjuntos de dados distintos.
3. *Seleção de Dados*: os dados relevantes para a tarefa de análise são obtidos a partir do banco de dados.
4. *Transformação dos Dados*: transformação e consolidação dos dados na forma adequada para a etapa de mineração de dados através de operações de agregação, por exemplo.
5. *Mineração dos Dados*: etapa essencial no processo de descoberta de conhecimento no qual métodos inteligentes são utilizados para extrair padrões dos dados.
6. *Avaliação dos Padrões*: identificação dos padrões verdadeiramente interessantes para a representação do conhecimento, ou seja, análise dos resultados obtidos na etapa de mineração de dados.
7. *Apresentação do Conhecimento*: onde técnicas de visualização e representação do conhecimento são utilizadas para apresentar o conhecimento extraído para o usuário.

As quatro primeiras etapas fazem parte do pré-processamento dos dados, onde, ao final dessas etapas, os dados estarão prontos para a mineração. Neste trabalho, todas essas etapas tiveram de ser aplicadas no conjunto de dados, com exceção da etapa ‘2’, pois como já foi dito anteriormente, os dados deste trabalho foram extraídos de um banco de dados único.

A etapa de mineração de dados é a etapa de extração do conhecimento propriamente dita, na qual as funcionalidades da mineração de dados são aplicadas ao conjunto de dados anteriormente preparado. Essas duas etapas (pré-processamento e mineração de dados) serão descritas com maiores detalhes nas próximas seções.

A etapa de mineração de dados pode interagir com o usuário ou com uma base de conhecimentos. Os padrões obtidos são apresentados ao usuário e podem ser armazenados como novos conhecimentos na base de conhecimentos. Note que a mineração de dados é apenas uma etapa no processo KDD, ainda que seja um fator essencial, pois é a etapa responsável por descobrir os padrões escondidos no banco de dados.

Alguns autores, embora entendam que a Mineração de Dados é uma etapa no processo de descoberta de conhecimento, preferem adotar esse termo ao invés de KDD. Eles alegam que o termo Mineração de Dados ganhou mais ênfase na mídia e vem sendo utilizado por grande parte das pessoas para representar todo o processo de extração de conhecimento. Neste trabalho o termo Mineração de Dados será utilizado para representar todo o processo de extração de conhecimento.

3.2 Pré-processamento dos Dados

No mundo real grande parte dos dados armazenados em Bancos de Dados ou em *Data Warehouses* possui algum tipo de "sujeira". Nesta seção serão apresentados os possíveis problemas em um conjunto de dados e quais são as soluções para resolver tais problemas.

Geralmente, os objetos que são armazenados em um repositório de dados não possuem todas as informações necessárias no momento em que são aplicadas as técnicas de mineração de dados. Embora pareça estranho esse fato não é nem um pouco raro no mundo real. Por exemplo, ao registrar informações sobre a venda de um produto o usuário não informou quantos itens foram vendidos e assim não tem como saber se o valor indica o preço unitário ou o valor total da venda.

Assim, o repositório que irá prover os dados para as técnicas de mineração de dados pode conter *dados incompletos*, que são dados nos quais faltam todos ou parte dos atributos que estão sendo analisados; *dados com ruídos* ou *outliers*, que são dados que possuem erros nos atributos ou são dados que fogem do padrão de distribuição do conjunto de dados; e *dados inconsistentes*, aqueles que possuem atributos com valores que não estão contidos no domínio do atributo.

O pré-processamento dos dados é definido pelas etapas de limpeza, integração, seleção e transformação. Essas etapas estão ilustradas na Figura 9, que representa todo o processo de extração de conhecimento.

As rotinas de *limpeza de dados*, como o próprio nome já diz, trabalham para "limpar" os dados através do preenchimento de valores ausentes, suavizando os dados

com ruídos, identificando ou removendo *outliers*. O objetivo dessas rotinas é resolver as inconsistências do conjunto de dados.

A confiabilidade dos resultados obtidos pelas técnicas de mineração de dados está diretamente relacionada à boa qualidade do conjunto de dados. Os dados sujos podem causar inconsistências no processo de mineração de dados, resultando na produção de informações não confiáveis. Por isso, o processo de pré-processamento dos dados é crucial para se ter um resultado consistente.

Outra rotina adotada no pré-processamento dos dados é a *integração dos dados*. Suponha que um gerente regional deseja obter informações sobre as vendas em seu estado. Para isso ele terá de unificar os dados que estão distribuídos em repositórios em cidades distintas. Esse processo deve ser feito de maneira muito cautelosa para que não sejam introduzidas inconsistências na nova base de dados.

A integração dos dados é uma etapa muito utilizada nos problemas de mineração de dados. Na maioria das vezes os dados estão espalhados e, em alguns casos, estão armazenados em repositórios com estruturas diferentes (arquivos texto, planilhas, bases de dados, etc.). Uma vez unidos em um mesmo repositório, esses dados podem ser trabalhados com maior facilidade e com uma menor possibilidade de erros.

Outra técnica de pré-processamento é a *transformação dos dados*. Suponha que se deseja usar um algoritmo de mineração que tenha como base o cálculo da distância entre os objetos no seu processo de análise, tais como redes neurais, classificadores por vizinho mais próximo, ou técnicas de agrupamento. Tais métodos fornecem melhores resultados se os dados a serem analisados estiverem normalizados, ou seja, estiverem em uma faixa específica, como de 0.0 a 1.0. A transformação de cada atributo para essa escala é um exemplo de transformação dos dados.

Neste trabalho foi necessário o uso de técnicas de limpeza e de transformação dos dados. Essas duas técnicas estão descritas nas próximas seções.

3.2.1 Limpeza dos Dados

Como dito anteriormente, a maioria dos conjuntos de dados apresenta alguns dados incompletos, com algum tipo de ruído, e inconsistentes. A limpeza dos dados é

uma etapa do processo KDD e tem como objetivo identificar e tratar os dados que podem causar impactos negativos na etapa de mineração dos dados.

Na limpeza no conjunto de dados procura-se identificar os dados que possuem informações incompletas, mas que precisam ser corrigidos para serem utilizados no processo de mineração. As possíveis ações para amenizar o impacto dos dados incompletos são (Han and Kamber, 2006):

1. *Ignorar o Elemento*: Geralmente, essa ação é tomada quando o rótulo de classe do elemento está faltando (assumindo que a tarefa de mineração envolve a classificação). Esse método não é muito eficaz, a não ser que o elemento tenha vários atributos com informações incompletas ou faltando.
2. *Preencher os valores em falta manualmente*: Essa abordagem pode ser muito demorada, tornando-se inviável de ser aplicada em grandes conjuntos de dados com muitos valores em falta.
3. *Usar uma constante global para preencher os valores em falta*: Essa técnica busca substituir todos os valores de um atributo que estão faltando pela mesma constante, com rótulos como "Desconhecido" ou $-\infty$. Se os valores em falta passam a ter o rótulo "Desconhecido", então o método de mineração de dados pode pensar equivocadamente que eles formam um conceito interessante, pois todos eles têm um valor em comum: o de "Desconhecido". Assim, embora esse método seja bastante simples, ele não é infalível.
4. *Usar a média do atributo para preencher o valor que está faltando*: Por exemplo, suponha que o rendimento médio dos clientes de uma empresa seja de R$ 3.000,00. Nesta técnica, esse valor poderia ser utilizado para substituir a falta de informação sobre a renda de alguns clientes.
5. *Usar a média do atributo para todas as amostras pertencentes à mesma classe*: Por exemplo, se uma empresa está classificando os seus clientes de acordo com o risco de crédito, então, essa técnica irá substituir o valor em falta com o valor de rendimento médio para os clientes da mesma categoria de risco de crédito.
6. *Usar o valor mais provável para preencher o valor em falta*: Isto pode ser determinado com regressão, ferramentas de inferência baseadas em um formalismo bayesiano, ou árvore de decisão. Por exemplo, usando os atributos

dos outros clientes do conjunto de dados, a empresa pode construir uma árvore de decisão para prever os valores que estão faltando para a renda.

Os métodos que utilizam valores constantes para a substituição dos dados que estão faltando correm grande risco de atribuírem valores incorretos aos dados. O sexto método, no entanto, é a estratégia mais popular, mesmo sendo mais complexo. Em comparação com os outros métodos, ele utiliza mais informações a partir dos dados atuais para prever os valores que estão faltando. Ao considerar os valores dos outros atributos, na estimativa do valor da renda, há uma chance maior de que a relação entre a renda e os outros atributos seja preservada.

Outro fator importante na limpeza do conjunto de dados é encontrar os *outliers*. Geralmente, esses elementos são encontrados através da análise da distribuição dos atributos. Os *outliers* são dados que estão fora do comportamento padrão do conjunto de dados e podem confundir as técnicas de mineração de dados, levando a resultados insatisfatórios.

Neste trabalho foi utilizado o diagrama de caixa ou *boxplot* (Frigge, Hoaglin and Iglewicz, 1989) para auxiliar a análise dos *outliers*. A Figura 10 ilustra um exemplo de análise da distribuição de atributos em um conjunto de dados usando o *boxplot*. Nessa figura as duas distribuições possuem *outliers* que, nesse caso, estão representados por elementos que possuem valor igual a zero (elementos definidos por asterisco).

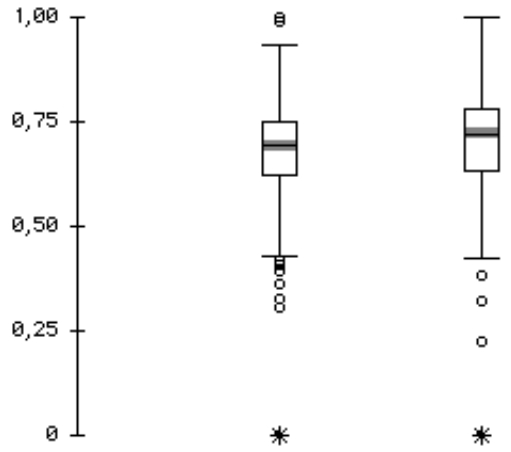


Figura 10 - Exemplo de outliers.

Os *outliers* podem ser removidos ou permanecer no conjunto de dados, essa ação irá depender da influência que esses elementos terão na etapa de mineração de dados. Observe que os *outliers* podem ser erros no conjunto de dados, mas podem também ser elementos especiais, ou seja, elementos que possuem características particulares. No caso de erro, esses elementos precisam ser corrigidos, mas no caso de serem particularidades do conjunto de dados, duas medidas podem ser tomadas: manter os elementos na mineração de dados, ou retirar os elementos da etapa de mineração e voltar com eles na etapa de avaliação dos resultados da mineração de dados.

3.2.2 Transformação dos dados

Na transformação dos dados, os dados são transformados ou consolidados em formas adequadas para a etapa de mineração. A transformação dos dados pode envolver os seguintes processos:

1. *Agregação*, onde operações de agregação são aplicadas aos dados. Por exemplo, os dados de vendas diárias podem ser agregados, de modo a computar totais mensais e anuais.
2. *Generalização dos dados*, onde dados em baixo nível ou "primitivos" (brutos) são substituídos por conceitos de alto nível através do uso de hierarquias de conceito. Por exemplo, atributos categóricos, como endereço, podem ser generalizados para os conceitos de nível superior, como a cidade ou país. Da mesma forma, o valor de atributos numéricos, como a idade, pode ser mapeado para conceitos em mais alto nível, como jovens, meia-idade e idosos.
3. *Normalização*, onde os atributos dos dados são dimensionados de forma a ficarem dentro de um pequeno intervalo, como de -1.0 a 1.0, ou de 0.0 a 1.0, que são os intervalos de normalização mais freqüentes.
4. *Construção de Atributos*, onde novos atributos são construídos e adicionados ao conjunto de atributos para ajudar no processo de mineração.

Para os métodos baseados em distância, a normalização ajuda a prevenir atributos inicialmente com intervalos grandes (por exemplo, renda) e atributos com escalas inicialmente menores (por exemplo, os atributos binários). Existem muitos métodos de

normalização de dados. Alguns deles são: normalização min-max, normalização z-score, e normalização por escala decimal (Han and Kamber, 2006).

A *normalização min-max* produz uma transformação linear nos dados originais. Suponha que $\min_A$ e $\max_A$ sejam, respectivamente, os valores mínimo e máximo de um atributo do elemento *A*. A normalização min-max mapeia um valor *v*, de *A*, para um valor $\overline{v}$ que estará no intervalo $[\text{novo\_min}_A, \text{novo\_max}_A]$. Essa normalização é calculada da seguinte forma:

$$\overline{v} = \frac{v - \min_A}{\max_A - \min_A}(\text{novo\_max}_A - \text{novo\_min}_A) + \text{novo\_min}_A. \quad (2)$$

A vantagem da normalização min-max é que ela preserva as relações que existem entre os dados na forma originais. Esta será a transformação utilizada no conjunto de dados deste trabalho.

3.3 Funcionalidades de Mineração de Dados

Um problema muito comum nos dias atuais está relacionado à tomada de decisão em ambientes cercados de incertezas e de imprecisões. O ser humano, a se ver cercado de alternativas e opções, não é capaz de tomar uma decisão com a certeza de que ela seja a mais correta. Isso ocorre por ele não ser capaz de analisar todos os dados que estão ao seu redor. Para evitar enganos que podem até mesmo serem fatais em alguns casos, esses problemas têm sido resolvidos utilizando-se técnicas que empregam, sobretudo, o conceito de aprendizado.

O aprendizado se dá a partir de dados experimentais ou da experiência do agente com o ambiente no qual o problema está inserido. Assim, é desenvolvido um sistema, a partir de um conjunto de dados, denominado conjunto de treinamento, capaz de extrair informações mais precisas sobre o problema.

Existem várias funcionalidades da mineração de dados que podem ser utilizadas para auxiliar esse processo de tomada de decisão e análise do conjunto de dados: Descrição Conceitual (Caracterização e Discriminação); Mineração de Padrões Freqüentes, Associações e Correlações; Classificação e Predição; e Análise de Agrupamentos.

3.3.1 Descrição Conceitual: Caracterização e discriminação

Na *descrição conceitual* os dados podem ser separados em classes ou conceitos. Por exemplo, em uma loja, classes de itens à venda podem conter computadores e impressoras, e os conceitos de clientes podem ser “grandes gastadores” e “pequenos gastadores”. Isso pode ser útil para descrever de forma resumida e precisa as classes e conceitos individuais. Essas descrições de uma classe ou conceito são chamadas de descrição de classe/conceito.

Essas descrições podem ser feitas através da caracterização dos dados, resumindo os dados das classes que está sendo estudada (muitas vezes chamada de classe principal) em termos genéricos; ou através da discriminação dos dados, pela comparação da classe principal com uma ou um conjunto de classes comparativas (muitas vezes chamadas de classes contrastantes); ou ainda através tanto da caracterização quando da discriminação dos dados.

A *Caracterização de Dados* é um resumo das características gerais ou características de uma classe de destino de dados. Os dados correspondentes à classe especificada pelo usuário normalmente são selecionados por meio de uma consulta ao banco de dados. Por exemplo, para estudar as características de produtos de software, cujas vendas aumentaram 10% no ano passado, os dados relacionados a esses produtos podem ser selecionados por meio da execução de uma consulta SQL.

Os resultados da caracterização de dados podem ser apresentados de várias formas. Exemplos incluem gráficos de pizza, gráficos de barras, curvas, cubos de dados multidimensionais e tabelas. As descrições resultantes podem também ser apresentadas como relações generalizadas ou em forma de regras, sendo chamadas de regras de características.

A *Discriminação dos Dados* é uma comparação entre as características gerais dos dados dos objetos da classe principal com as características gerais dos objetos de uma ou um conjunto de classes contrastantes. O destino e as classes contrastantes podem ser especificados pelo usuário e os objetos podem ser recuperados através de consultas ao banco de dados. Por exemplo, o usuário pode querer comparar as características gerais dos produtos de software, cujas vendas aumentaram 10% no ano passado com os

produtos cujas vendas diminuíram em pelo menos 30% durante o mesmo período. Os métodos utilizados para a discriminação de dados são semelhantes aos utilizados para a caracterização de dados.

As formas de apresentação dos resultados são semelhantes às descrições de características, embora as descrições de discriminação devam incluir medidas comparativas que ajudam a distinguir entre o objeto da classe principal e objetos das classes contrastantes. As descrições de discriminação representadas na forma de regras são definidas como *regras discriminantes*.

3.3.2 Padrões Freqüentes

Os Padrões Freqüentes, como o nome sugere, são os padrões que ocorrem com freqüência no conjunto de dados. Existem muitos tipos de padrões freqüentes, incluindo *itemsets*, subseqüências e subestruturas. Um *itemset freqüente* se refere a um conjunto de itens que freqüentemente aparecem juntos em um conjunto de dados transacional.

Por exemplo, no conjunto de dados dos consumidores de uma padaria, provavelmente, será encontrado o padrão de compra do item leite e do item pão. Outro padrão muito curioso é o de compra de fraldas e cervejas em supermercados. Acredita-se que os pais que possuem filhos pequenos e não podem sair compram suas bebidas para ficarem em casa e cuidar das crianças.

A *subseqüência freqüente* representa uma seqüência de ações que se repetem com freqüência no conjunto de dados. Como exemplo, no conjunto de dados de uma loja de informática, uma seqüência de ações freqüentes dos clientes poderia ser em primeiro lugar comprar um computador, seguido por uma câmera digital, e então um cartão de memória.

A *subestrutura* pode se referir a diferentes formas estruturais, tais como grafos e árvores, que podem ser combinadas com *itemsets* ou subseqüências. Se uma subestrutura ocorre com freqüência, é chamado de padrão (freqüente) estruturado. Mineração de padrões freqüentes leva à descoberta de associações e correlações interessantes que estão escondidas dentro de grandes bancos de dados.

Análise de Associação. Suponha que o gerente de marketing de uma empresa de informática gostaria de determinar quais os itens são freqüentemente comprados juntos nas mesmas operações. Um exemplo de regra, que pode ser extraída do seu banco de dados transacional, é:

$$\text{Compra(X,"computador")} \rightarrow \text{Compra(X,"software")} \quad [\text{suporte=1\%, confiança=50\%}],$$

onde X é uma variável que representa um cliente.

A **confiança** indica a probabilidade do segundo evento ocorrer, ou seja, no exemplo acima, a confiança de 50% significa que se um cliente comprar um computador, há 50% de chance que ele também irá comprar um software. O **suporte** de 1% significa que 1% de todas as operações em análise mostrou que o computador e o software foram comprados juntos. Normalmente, as regras de associação são descartadas por não serem consideradas interessantes, se não satisfazerem suporte e confiança mínimos.

É importante ressaltar que as técnicas de mineração de dados podem produzir centenas e até milhares de padrões ou regras, porém a grande maioria não é interessante. Geralmente, apenas uma pequena parcela das regras e padrões encontrados tem um valor interessante para um determinado usuário. Pode-se afirmar que para que um padrão seja interessante ele precisa ser de fácil compreensão para os humanos; ser válido quando aplicado a novos dados ou dados de testes, ou seja, precisa ter certo grau de confiança em novos dados; precisa ser potencialmente útil; e, por último, precisa ser novo, ou seja, padrões que já são conhecidos não são interessantes.

3.3.3 Classificação

A *classificação* é o processo de encontrar um modelo (ou função) que descreve e distingue classes de dados. A finalidade dessa técnica é que o modelo seja capaz de prever a classe de objetos cujo rótulo da classe é desconhecido. O modelo é derivado com base na análise de um conjunto de dados de treinamento, ou seja, objetos de dados cuja classe é conhecida.

O modelo utilizado para a classificação e predição de novos objetos pode ser representado de diversas formas, como: regras de classificação *se-então*, árvores de decisão, redes neurais, máquinas de vetores suporte, dentre outros.

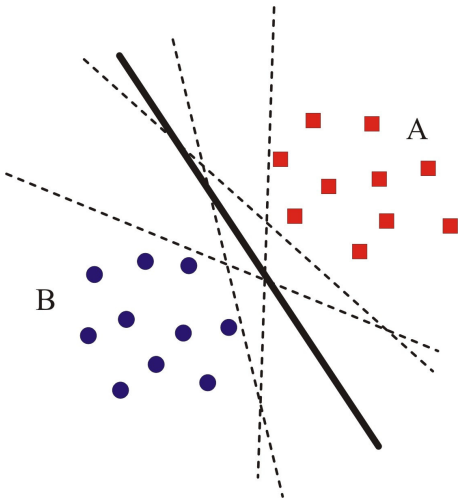


Figura 11 - Hiperplanos separadores das classes A e B.

Nos problemas de classificação todos os elementos do conjunto de treinamento estão rotulados indicando a qual classe eles pertencem. Assim, na etapa de treinamento do classificador é verificado se o mesmo está acertando ou errando a classe do dado. Como há esse acompanhamento dos erros e acertos do classificador na fase de treinamento, essa fase é chamada de treinamento supervisionado.

A Figura 11 ilustra um exemplo de um conjunto de hiperplanos separadores de duas classes A e B. Os hiperplanos representados por linhas pontilhadas separam as classes sem erros no treinamento, ou seja, nenhum elemento de uma classe é classificado como sendo da outra classe. Entretanto, o hiperplano representado pela linha mais grossa separa melhor os dados das classes, pois ele está exatamente no meio delas.

Assim, embora existam infinitos hiperplanos separadores, as técnicas de classificação procuram aqueles que separam as classes com a maior margem possível. Esses classificadores possuem um poder de generalização melhor e, por isso, produzem resultados melhores quando são aplicados a dados onde não se conhece a classe a qual eles pertencem.

3.3.4 Análise de Agrupamento

Análises de agrupamentos são técnicas que visam identificar subconjuntos de dados com base na similaridade entre eles (Han and Kamber, 2006). Objetos do mesmo subconjunto são mais semelhantes entre si do que entre os objetos de subconjuntos diferentes.

A diferença principal entre classificação e agrupamento está na forma como os dados são fornecidos. Nos problemas de classificação cada elemento do conjunto de treinamento está rotulado com alguma informação que especifica a qual classe ele pertence. Já nos problemas de agrupamento os dados não são rotulados.

Na análise de agrupamentos o grau de semelhança entre cada par de elementos do conjunto de treinamento é determinado através do uso de métricas. Assim, os dados mais semelhantes são agrupados em grupos comuns, de forma que os dados com menor similaridade ficam em grupos diferentes.

Em geral, os rótulos das classes dos objetos não estão presentes nos dados do conjunto de treinamento, simplesmente porque essa informação não é conhecida. Os objetos são agrupados com base no princípio da maximização da similaridade intra-classes e minimização da similaridade inter-classes. Em outras palavras, conjuntos de objetos são formados para que os objetos dentro de um mesmo grupo tenham alta similaridade entre si, mas sejam muito diferentes dos objetos dos outros grupos.

Existem basicamente duas técnicas de agrupamento: *agrupamento hierárquico* e *agrupamento via particionamento* (Jain and Dubes, 1988). As técnicas de agrupamento hierárquico são bastante simples, intuitivas e úteis para pequenos conjuntos de dados, onde é possível analisar visualmente o processo de agrupamento passo a passo. Nesses algoritmos, um elemento não muda de um grupo para outro, pois uma vez definido o grupo ao qual ele pertence não é mais possível alterá-lo.

A técnica de agrupamento hierárquico é um procedimento que constrói uma “árvore” onde cada nível dessa árvore representa uma alteração na estrutura dos grupos. Existem dois tipos de agrupamento hierárquico. O primeiro é chamado de agrupamento hierárquico *aglomerativo* que adiciona os dados aos grupos em cada nível da árvore. O

segundo é denominado agrupamento hierárquico *particionado* que realiza sucessivas divisões nos grupos.

Para um melhor entendimento do funcionamento dos algoritmos hierárquicos, considere *N* como sendo o número de elementos de um problema de agrupamento. Os métodos hierárquicos aglomerativos começam com *N* grupos, onde cada um dos grupos possui uma única observação, essa etapa está ilustrada no quadro 1 da Figura 12. O número de grupos é então reduzido para *N-1* pela junção de dois elementos em um único grupo baseado em algum critério de similaridade entre eles (quadro 2 da Figura 12). O processo continua até que o número desejado de grupos seja atingido ou até que seja obtido um único grupo que contenha todas as observações (quadro 8 da Figura 12).

Nessa técnica de agrupamento aglomerativo a árvore é construída das folhas para a raiz, onde cada folha representa uma observação. Esse processo é ilustrado na Figura 12, começando do primeiro até o oitavo quadro.

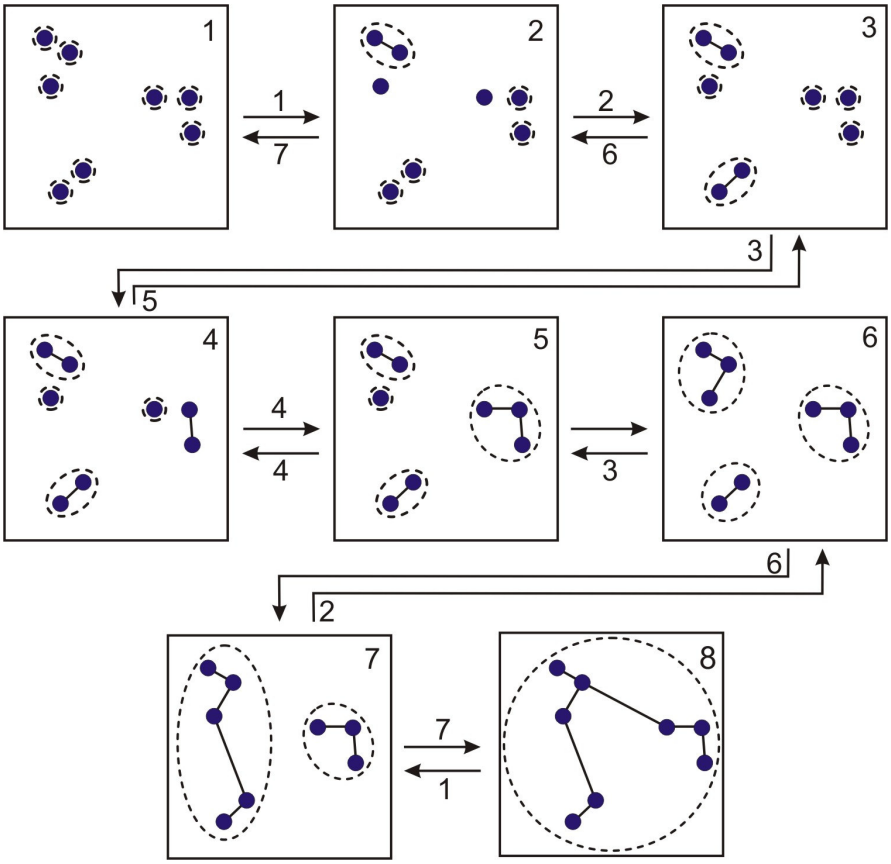


Figura 12 - Agrupamento Hierárquico Aglomerativo e Particionado.

Já os métodos hierárquicos particionados trabalham no sentido oposto dos métodos aglomerativos. Esses métodos começam com um único grupo de *N* elementos (quadro 8 da Figura 12) e divide esse grupo em grupos menores usando um critério de dissimilaridade entre os dados. Esse método de agrupamento está ilustrado na Figura 12, sendo que a primeira etapa desse algoritmo está ilustrada no quadro 8 e a última no quadro 1 dessa figura.

Um critério bastante utilizado de dissimilaridade é a distância entre os elementos. Várias funções de distância podem ser utilizadas, sendo a mais tradicional a distância Euclidiana. O processo de particionamento continua até que algum critério seja atingido ou até que sejam obtidos *N* grupos com um elemento em cada grupo. O processo dos métodos hierárquicos particionados também é ilustrado na Figura 12, pela ordem inversa, começando do oitavo até o primeiro quadro.

Os *Algoritmos de Particionamento*, que utilizam técnicas de agrupamento não hierárquico, buscam minimizar a distância intra-grupos, ou seja, minimizar a distância entre as observações pertencentes ao mesmo grupo.

Diferentemente dos algoritmos de agrupamento hierárquicos, os algoritmos de agrupamento via particionamento exigem que o número *K* de grupos seja definido no começo do processo. Após ser definido o número de grupos as observações são separadas em *K* grupos.

Nessas técnicas de agrupamento é definido um centro ou um elemento que melhor represente cada grupo. Assim, em cada iteração, a distância entre as observações do conjunto de treinamento e os centros dos grupos é calculada. Cada observação será então associada ao grupo do centro mais próximo. No início de cada iteração os centros dos grupos são recalculados e as distâncias são novamente obtidas. Assim, a observação pode permanecer no mesmo grupo ou ser associada a outro grupo que ela esteja mais próxima. As iterações devem continuar até que não haja nenhuma troca de elementos entre os grupos.

Os Algoritmos de Particionamento podem ser aplicados a qualquer conjunto de dados, independentemente do tamanho desse conjunto. Esses métodos geram uma única partição dos dados em uma tentativa de recuperar grupos naturais presentes nos dados. Além disso, todos os grupos têm um elemento central, e todos os elementos de um grupo

são mais parecidos com o elemento central do seu próprio grupo do que com os elementos centrais dos outros grupos.

Nos algoritmos hierárquicos, não é necessário definir o número de grupos. Já os algoritmos de particionamento requerem que o número de grupos seja definido no início do processo de agrupamento. Embora a definição do número de grupos não seja trivial, foi utilizado neste trabalho um algoritmo de agrupamento não hierárquico, uma vez que não existem restrições com relação ao tamanho do conjunto de dados e, além disso, também é possível identificar os elementos que melhor representam os grupos (elementos centrais).

Nos algoritmos hierárquicos a definição do número de grupos ideal é feita através da análise dos grupos formados em cada iteração. Assim, quanto maior o conjunto de dados maior será o número de etapas do algoritmo de agrupamento e, conseqüentemente, maior será a dificuldade em definir o número de grupos. Por isso, para grandes conjuntos de dados essa metodologia se torna inviável.

Como dito anteriormente, nessa metodologia a semelhança entre os objetos é extraída de suas estruturas. Geralmente, os métodos de agrupamento usam uma matriz que representa a similaridade entre os objetos. Neste trabalho, será utilizada uma matriz que representa o grau do relacionamento entre os pesquisadores, ou seja, os relacionamentos serão utilizados como uma métrica de similaridade de forma que os pesquisadores mais fortemente relacionados serão mais “semelhantes” entre si que os menos fortemente relacionados. Essa matriz que representam o grau de similaridade será construída no capítulo 4.

3.4 Etapas do KDD

Nesta seção serão descritas as etapas do processo KDD aplicadas no desenvolvimento deste trabalho. A primeira etapa do trabalho foi a limpeza e a integração dos dados que foram utilizados na modelagem da rede social científica. Essa etapa será descrita em detalhes no próximo capítulo.

Embora a construção de um *Data Warehouse* seja uma etapa crucial para muitos problemas de mineração de dados, neste trabalho não foi necessário o desenvolvimento

dessa arquitetura. Todos os dados utilizados neste trabalho já estão consolidados em uma base única.

Entretanto, os dados tiveram que ser trabalhados e unificados de maneira a facilitar a sua manipulação. A modelagem da rede social científica multi-relacional exigiu que os dados estivessem consolidados de uma forma mais simples do que eles estavam no banco de dados.

Assim, para viabilizar a modelagem da rede social científica multi-relacional foram criadas novas tabelas no banco de dados. Essas tabelas foram criadas de forma a consolidar os relacionamentos dos pesquisadores distribuídos por ano, para que seja possível a análise temporal da rede social.

Como o objetivo deste trabalho com a funcionalidade de mineração de dados é encontrar comunidades científicas, a melhor técnica a ser utilizada é a técnica de agrupamento. Os algoritmos de agrupamento desenvolvidos estão descritos em detalhes no capítulo 5.

Nos próximos capítulos serão apresentadas as etapas do processo KDD que tiveram de ser aplicadas para se obter os resultados. As principais etapas do processo abordadas foram: limpeza e transformação dos dados, mineração dos dados, análise e apresentação dos resultados.

Capítulo 4 – Rede Social Científica Multi-relacional

As redes sociais Científicas são tipos específicos de redes sociais que representam as interações sociais oriundas do meio acadêmico. No desenvolvimento deste trabalho as redes sociais científicas serão utilizadas como fontes de dados para testes e avaliação dos resultados obtidos.

Além dos conceitos envolvidos na definição das redes sociais científicas, neste capítulo será apresentada a primeira etapa do desenvolvimento deste trabalho. Nesta etapa serão apresentados todos os passos da modelagem da rede social científica multi-relacional.

Esta etapa é formada por três passos principais, que são: extração das redes sociais homogêneas da base de dados, modelagem da rede social científica multi-relacional e representação dessa rede na forma de matrizes. Essas etapas podem ser visualizadas de maneira abstrata no primeiro capítulo deste trabalho na Figura 1.

4.1 Rede Social Científica Brasileira

Anualmente, a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), uma instituição que recebe suporte do Ministério da Educação do Governo Federal Brasileiro, avalia os programas de pós-graduação e atribui a eles notas que variam de 3 a 7.

Os critérios para essa avaliação incluem a qualificação do corpo docente, as pesquisas desenvolvidas, formação, produção intelectual, qualidade e quantidade de teses e dissertações, dentre outros aspectos. A avaliação das instituições de ensino é feita por uma equipe formada pelos profissionais mais experientes da área que está sendo avaliada. As instituições avaliadas com nível sete são consideradas de grande excelência.

Os pesquisadores utilizados para a construção da rede social científica modelada nesse trabalho foram selecionados das instituições nível seis e nível sete dos programas de Ciência da Computação. As instituições nível sete são: COPPE/UFRJ, PUC-RIO e UFMG. Já as instituições nível seis são: UFPE e UFRGS. Os dados foram restritos a

essas instituições de ensino para que fosse analisado como as instituições com níveis máximos de excelência se relacionam entre si.

Assim, os nós do grafo são representados por pesquisadores e as arestas são os relacionamentos entre cada par de pesquisadores. Existem várias maneiras de identificar um relacionamento científico entre dois pesquisadores. Em geral, essas relações podem ser: participações em Projetos; relacionamentos de co-autoria; orientações em dissertações e teses; participação em banca de defesa de dissertações e teses; produções técnicas; participação em comissões examinadoras; dentre outros tipos de relações científicas.

O relacionamento de participação em projetos só existe quando dois ou mais pesquisadores trabalharam em conjunto no desenvolvimento do mesmo projeto. Os pesquisadores que trabalham no mesmo projeto, além de executarem atividades comuns, estão empenhados na resolução do mesmo problema.

O relacionamento de co-autoria é uma das mais importantes e com maior expressão nas redes sociais científicas. Isto se deve ao fato dos pesquisadores estarem envolvidos nos estudos e nas publicações de um mesmo assunto. Portanto, há um interesse comum entre eles sobre o assunto que está sendo pesquisado, por isso, é assumido que esses pesquisadores estão mais diretamente relacionados. A maioria dos estudos em redes sociais científicas considera o relacionamento de co-autoria na construção da rede social.

A co-orientação ocorre quando dois pesquisadores orientam o mesmo aluno na mesma obra (dissertação e tese). Então, assim como as relações de co-autoria, os pesquisadores também desenvolvem as suas pesquisas sobre o mesmo assunto ou assuntos que estejam relacionados ou sejam complementares.

O relacionamento de orientação representa a conexão entre o pesquisador (orientador) e o aluno. A identificação dessa relação é importante para analisar a evolução do relacionamento professor-aluno no decorrer do tempo. O acompanhamento desse relacionamento pode revelar a evolução de um aluno até ele também se tornar pesquisador.

O relacionamento de participação em banca ocorre quando dois pesquisadores participam da mesma banca de defesa da mesma obra, devido à conclusão de um trabalho

de dissertação ou tese. Por exemplo, quando dois pesquisadores participam da mesma banca examinadora de uma defesa de tese de doutorado, significa que eles têm conhecimentos em comum sobre o tema que está sendo apresentado. Apesar de ser uma ligação mais fraca, é considerado um tipo de relacionamento importante, pois os relacionamentos com pesquisadores externos podem começar por meio dos relacionamentos de participação em banca.

O relacionamento de produção técnica ocorre quando dois pesquisadores trabalharam em conjunto no desenvolvimento de algum produto técnico, tais como: Software com ou sem registro de patente, produtos tecnológicos, trabalhos técnicos.

As relações existentes entre membros de comissões julgadoras ou comissões de premiação surgem de maneira semelhante à relação de participação em banca, descrita anteriormente. Geralmente, dois pesquisadores terão esse tipo de relacionamento quando eles participarem da mesma comissão, seja para julgar ou para premiar um trabalho científico.

Os relacionamentos mais importantes no estudo das redes sociais científicas são aqueles que melhor representam os interesses comuns dos pesquisadores. Assim, mesmo que todos os tipos de relacionamentos possam ser considerados importantes, os relacionamentos de co-autoria, produções técnicas, participação conjunta no desenvolvimento de projetos, e co-orientação de trabalhos são mais interessantes, pois representam o interesse dos pesquisadores pelo mesmo assunto.

Por outro lado, embora o relacionamento de participação em banca nem sempre represente o interesse dos pesquisadores pelo mesmo tema, esse relacionamento tem um grande potencial de criar novos vínculos. Algumas instituições, como o caso da UFRJ, exigem que ao menos um membro da banca seja de uma universidade externa. Assim, os pesquisadores internos criam novos vínculos com pesquisadores de outras instituições de ensino. Por ter essa característica o relacionamento de participação em banca tem grande importância na análise de redes sociais científicas.

Além dos relacionamentos, cada pesquisador possui um perfil individual. O perfil de um pesquisador é definido por um conjunto de atributos pessoais, tais como: formação acadêmica, área de pesquisa e área de atuação, número de publicações em revistas,

número de publicações em congressos; número de publicações em relatórios técnicos, número de participações em projetos; dentre outros.

O atributo de formação acadêmica define a qualificação de um professor, por exemplo, M.Sc., D.Sc, e assim por diante. As áreas de pesquisa e atuação indicam a qual área de atividade o pesquisador faz parte. Exemplos de áreas de pesquisa e atuação são: bancos de dados, inteligência artificial, mineração de dados, engenharia de software, dentre outros.

Os atributos quantitativos das publicações indicam o número total de cada tipo de publicação do investigador. Assim, o atributo de publicação em revistas indica o número de artigos que o pesquisador publicou em revistas. Da mesma forma, o atributo de participações em projetos indica o número de projetos que o pesquisador participou no decorrer de sua vida científica.

Como dito anteriormente, os pesquisadores estão ligados uns aos outros, direta ou indiretamente. Essa associação pode ser mais forte ou mais fraca, segundo o grau do relacionamento entre eles. Pesquisadores que, por exemplo, têm publicações em comum, que trabalham em áreas semelhantes e que participaram no desenvolvimento de um mesmo projeto, podem ser considerados como tendo um relacionamento forte. Por outro lado, se dois pesquisadores participaram de apenas uma banca examinadora, o relacionamento entre eles é considerado fraco. Além disso, há casos em que os pesquisadores não estão diretamente conectados, nesse caso, a ligação entre eles se dará por intermédio de outros pesquisadores.

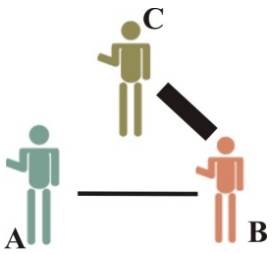


Figura 13 - Relacionamento Direto e Indireto

Tabela 1 - Matriz de pesos correspondente a Rede Social da Figura 8

	A	**B**	**C**
A	-	1	0
B	1	-	2
C	0	2	-

A pequena rede social apresentada na Figura 13 ilustra a situação descrita anteriormente. O pesquisador **B** está diretamente relacionado aos pesquisadores **A** e **C**.

Os pesquisadores **A** e **C** não estão diretamente relacionados, porém eles estão ligados indiretamente, por transitividade, pelo pesquisador **B**.

Além disso, o grau do relacionamento entre os pesquisadores **B** e **C** é maior que o grau do relacionamento entre os pesquisadores **A** e **B**. O grau do relacionamento é representado pela aresta mais grossa entre os elementos mais fortemente relacionados e por arestas mais finas para os elementos menos fortemente relacionados. Esse fato pode ser confirmado através da matriz de pesos representada na Tabela 1, na qual pode ser observado que o peso do relacionamento entre **B** e **C** é o dobro do peso do relacionamento entre **A** e **B**.

Os dados para o desenvolvimento da rede social científica multi-relacional foram selecionados inicialmente da plataforma Lattes (LATTES, 2008). Essa é uma plataforma Web mantida pelo Governo brasileiro, onde pesquisadores e estudantes devem fornecer suas informações em um currículo acadêmico público.

Essa plataforma foi desenvolvida para registrar de maneira evolutiva a vida acadêmica dos pesquisadores brasileiros. Todos os relacionamentos e atributos mencionados anteriormente são fornecidos pelos pesquisadores e pelos estudantes nessa plataforma. Entretanto, esses dados só estão disponíveis na web, sendo necessário o uso de uma ferramenta para extraí-los e armazená-los em um banco de dados. Para isso foi utilizado o GCC (Gestão do Conhecimento Científico) (Oliveira, Souza, Miranda and Rodrigues, 2006).

O GCC é um ambiente Web desenvolvido originalmente pela COPPE/UFRJ, cujo objetivo é permitir a gestão do conhecimento nas instituições de pesquisa e melhorar a colaboração entre pesquisadores, estimulando o desenvolvimento de novas idéias. Através dos serviços do GCC as informações sobre os currículos dos pesquisadores, obtidas a partir da plataforma Lattes, são armazenas em um banco de dados facilitando o acesso e a manipulação dessas informações.

Os dados para a modelagem da rede social multi-relacional foram extraídos do currículo Lattes dos pesquisadores de instituições avaliadas com níveis 6 e 7 de acordo com os critérios da CAPES (a pontuação máxima é 7). Ao todo foram analisados 169 pesquisadores da área de Ciência da Computação a partir de cinco universidades brasileiras.

Após a análise dos dados disponíveis, foram selecionados os atributos que fossem sempre informados pelos pesquisadores em seus currículos e que, além disso, foram considerados mais importantes no contexto científico. Em um primeiro momento, foram selecionados apenas alguns dos atributos perfil, tais como: áreas de pesquisa e de atuação do pesquisador, número de publicações em revistas, número de publicações em congressos e número de orientações, e apenas os relacionamentos de co-autoria (Ströele, Silva, Oliveira, Souza and Zimbrao, 2009; Ströele, Silva, Souza, Oliveira, Mello, Souza and Zimbrão, 2008; Ströele, Silva, Souza, Mello, Souza, Zimbrao and Oliveira, 2011).

Em um segundo momento, foram selecionados vários tipos de relacionamentos científicos para modelar o grafo social, gerando assim, uma rede social científica multi-relacional (Ströele, Oliveira, Zimbrao and Souza, 2009; Ströele, Zimbrão and Souza, 2011; Ströele, Zimbrão and Souza, 2012).

4.2 Conceitos

Existem dois tipos de redes sociais: homogêneas e heterogêneas (Cai, Shao, He, Yan and Han, 2005). As Redes Sociais Homogêneas são aquelas onde há apenas um tipo de objeto e um tipo de relacionamento entre esses objetos.

Já as Redes Sociais Heterogêneas podem representar tipos de objetos diferentes conectados por vários tipos de relacionamentos. As redes sociais heterogêneas que possuem apenas um tipo de objeto e vários tipos de relacionamentos diferentes são conhecidas como Redes Sociais Multi-relacionais.

Em geral, a maioria dos métodos de mineração em redes sociais considera apenas as Redes Sociais Homogêneas. No entanto, no mundo real, quase todas as redes sociais têm vários tipos de relacionamentos entre os objetos.

Um tipo particular de problema com relação às redes sociais multi-relacionais encontra-se em extrair os diferentes tipos de relacionamentos que existem na rede. Neste tipo de problema, cada tipo de relacionamento pode ser modelado como um grafo. Dependendo das informações que se deseja extrair da rede social, analisar um tipo de relacionamento será mais importante do que analisar outro tipo. Assim, para uma melhor análise da rede social multi-relacional é necessário selecionar as relações que têm um

efeito positivo para o problema que está sendo solucionado (Cai, Shao, He, Yan and Han, 2005).

A Figura 14 é um exemplo de uma rede social com três tipos diferentes de relacionamentos. Dependendo da informação que se deseja encontrar será necessário analisar a rede social do ponto de vista de um relacionamento específico. Assim, se um usuário quer que os objetos A, B, C e D pertençam a uma mesma comunidade, então o relacionamento representado na Figura 14(a) terá um efeito positivo sobre as informações, enquanto que o relacionamento mostrado na Figura 14(c) terá um efeito negativo sobre a análise. Conseqüentemente, o usuário precisa saber o que ele quer descobrir na rede social, e, em seguida, definir qual o tipo de relacionamento deve ser utilizado para que ele tenha uma análise correta da rede social.

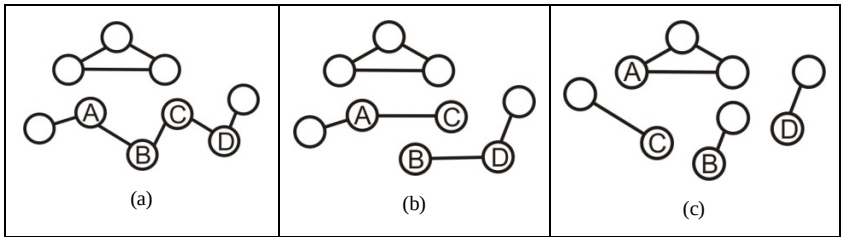


Figura 14 - Rede Social Multi-relacional com três tipos de relacionamentos.

As Redes Sociais Científicas são tipos específicos de redes sociais, onde dois pesquisadores são considerados conectados se eles foram co-autores de ao menos um artigo (Ichise, Takeda and Ueyama, 2005; Newman, 2000; Newman, 2001a; Newman, 2001b; Newman, 2001c). Entretanto, as redes sociais em ambientes acadêmicos são mais complexas já que as suas conexões envolvem diferentes tipos de colaborações ou interações científicas. Assim, as redes sociais científicas podem ser consideradas um tipo específico de Rede Social Multi-relacional.

Nas redes sociais homogêneas, onde há apenas um tipo de relacionamento, o fluxo de conhecimento na rede se dá através desse relacionamento específico. Assim, a Análise da Rede Social irá considerar apenas um tipo de troca de conhecimento entre os elementos da rede.

Por outro lado, nas redes multi-relacionais a troca de conhecimento se dá através de diferentes tipos de relacionamentos. Assim, a análise da rede social multi-relacional

assume que os elementos estão trocando conhecimentos de conteúdos diferentes, dependendo dos tipos de relacionamentos que ligam os objetos.

A análise de múltiplos relacionamentos também permite que os elementos relacionados por conexões secundárias tenham seus relacionamentos refletidos na rede social. Por exemplo, dois pesquisadores que não possuem o relacionamento de co-autoria, mas que são parte do mesmo projeto terão um relacionamento explícito na rede social científica multi-relacional.

Neste trabalho, foram utilizados quatro tipos de relacionamentos diferentes para modelar a Rede Social Científica Multi-relacional: Participação em Projeto; Publicações como co-autor; Participação em Bancas de tese de doutorado e dissertação de mestrado; e Produção Técnica. A definição de cada um deles encontra-se na seção anterior. Esses relacionamentos foram selecionados, pois eles estavam com uma boa taxa de preenchimento no currículo Lattes. Pouquíssimos pesquisadores possuíam os outros tipos de relacionamentos e, por isso, esses relacionamentos não foram considerados.

Na próxima seção será apresentada a etapa de pré-processamento de dados do processo KDD aplicada aos dados utilizados neste trabalho.

4.3 Pré-Processamento dos Dados

Como mencionado anteriormente, foram adotadas duas metodologias distintas de mineração de dados para analisar os dados durante o desenvolvimento deste trabalho. Essas técnicas, embora não precisem explicitamente de um *Data Warehouse*, elas necessitam que os dados sejam preparados antes de serem utilizados para que os resultados da mineração seja satisfatório.

Portanto, foram aplicadas algumas etapas do processo de pré-processamento de dados a fim de torná-los adequados para as técnicas de mineração de dados. As etapas de pré-processamento utilizadas foram limpeza, análise exploratória e normalização dos dados.

4.3.1 Limpeza e Análise Exploratória dos Dados

Na etapa de pré-processamento foi verificada a consistência dos dados extraídos do banco de dados do GCC. Nessa fase foram encontradas algumas inconsistências nos relacionamentos entre pesquisadores e suas produções bibliográficas: alguns pesquisadores estavam associados à produção bibliográfica mais de uma vez. Tais inconsistências prejudicam a análise dos dados e, como solução, os relacionamentos duplicados foram removidos da base de dados.

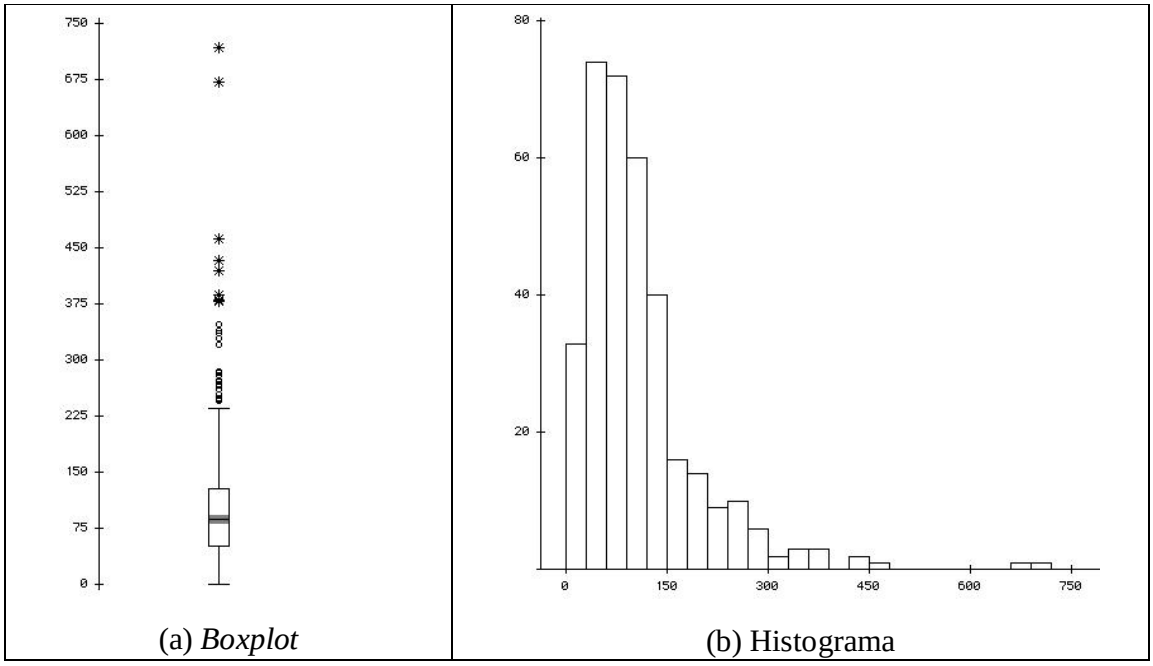
Além das informações duplicadas, a análise dos dados de relacionamento permitiu que fossem identificados se dois pesquisadores possuíam um forte relacionamento entre si, mas não se relacionavam com nenhum outro pesquisador. Esses dois pesquisadores foram retirados do conjunto de dados, já que eles tornam o grafo desconexo e as técnicas de mineração de dados não funcionam com esse tipo de grafo. Todos os outros pesquisadores e seus relacionamentos foram mantidos.

A análise exploratória dos dados consiste em analisar as distribuições dos atributos, tanto de perfil quanto de relacionamento, identificar *outliers*, e redundâncias nas variáveis. Para auxiliar a análise, foram usadas basicamente ferramentas gráficas, tais como: histogramas e *boxplot*. Esses gráficos facilitam a visualização da distribuição das variáveis do conjunto de dados e, conseqüentemente, facilitam a análise desses dados.

A Figura 15 ilustra esses dois gráficos para o atributo de produções bibliográficas. Esse atributo indica quantas publicações de artigos científicos cada pesquisador possui. No *boxplot*, ilustrado na Figura 15(a), é possível ver de maneira mais clara que a maioria dos pesquisadores possui máximo de publicações em torno de 225. Analisando ainda o *boxplot* é possível ver que existem alguns pesquisadores que possuem valores muito discrepantes comparados com o restante do conjunto de dados. Esses pesquisadores estão marcados por asteriscos e são considerados possíveis *outliers*, ou seja, são elementos que precisam ser analisados cuidadosamente para ver se existem erros nesses dados.

O histograma, ilustrado na Figura 15(b), também representa o comportamento do conjunto de dados. Através desse gráfico é fácil visualizar se os dados estão seguindo uma distribuição normal ou não. A transformação dos dados de maneira que eles estejam normalizados e bem distribuídos faz com que as medidas utilizadas nos métodos de

mineração de dados tenham a mesma escala. Assim, as variáveis envolvidas nos métodos de agrupamento, técnica de mineração de dados adotada neste trabalho, possuem o mesmo peso, evitando que uma variável tenha uma influência maior que as outras no processo de análise.



(a) *Boxplot* (b) Histograma

Figura 15 - *Boxplot* e Histograma do atributo de produções bibliográficas

A análise desses gráficos orientou o trabalho de limpeza do conjunto de dados. Através deles foi possível identificar os pesquisadores que mereciam maior atenção durante essa etapa. Alguns pesquisadores tiveram de ser removido do conjunto de dados por não ter sido possível corrigir os seus dados. Entretanto, alguns outros tiveram seus dados atualizados e foram mantidos.

Após essa primeira limpeza e análise dos dados foi feita a etapa de transformação dos dados do processo KDD. Na etapa de transformação os atributos de perfil e os relacionamentos foram normalizados para que pudessem ser utilizados com maior segurança na etapa de mineração de dados. A transformação desses dados está descrita nas próximas seções.

4.3.2 Atributos de Perfil

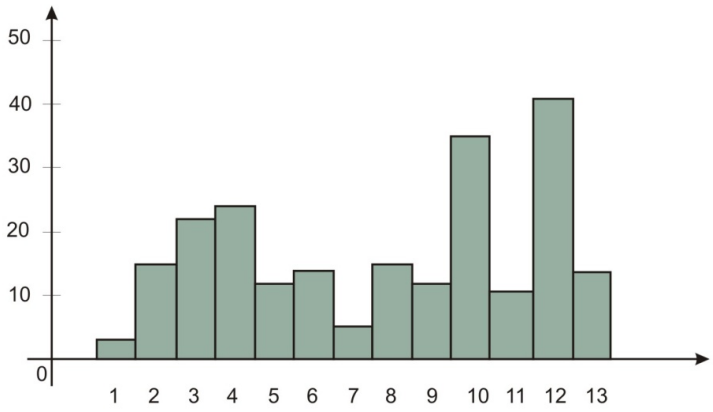
Os atributos de perfil são informações particulares de cada pesquisador. A idade, sexo, área de pesquisa e número de produções bibliográficas são exemplos de atributos de perfil de um pesquisador.

Durante a análise dos perfis dos pesquisadores, os perfis de número de participações em banca, número de produções tecnológicas, número de participações em projetos, número de publicações em congressos e número de publicações em revistas foram removidos da análise, pois muitos desses atributos eram iguais a zero no currículo de muitos pesquisadores. Geralmente, esse problema é causado por pesquisadores que não preenchem completamente as informações em seus currículos.

Neste trabalho foram utilizados dois métodos de agrupamento. Um deles utiliza as informações dos atributos de perfil dos pesquisadores e o outro não. Esses métodos estão descritos em detalhes no capítulo 5. Apesar da remoção desses atributos de perfil, a influência dos relacionamentos dos pesquisadores ajuda a manter os grupos juntos no método que utiliza as informações dos atributos de perfil.

Durante a análise dos atributos que foram mantidos no conjunto de dados, observou-se que os pesquisadores que têm um número maior de alunos têm um grande número de publicações. Entretanto, em alguns casos há pesquisadores que orientam vários alunos, mas que não publicam com muita freqüência. Portanto, não se pode generalizar essa regra.

A CAPES define as áreas de trabalho no momento em que um professor registra o seu currículo na plataforma escolhendo uma ou mais áreas de pesquisa. No entanto, 45 dos 169 pesquisadores não preencheram nenhuma área de trabalho e um grande número de áreas possui apenas um pesquisador associado. Para tornar a análise mais prática, optou-se por analisar as áreas com mais de três pesquisadores. A Figura 16 mostra a distribuição dos pesquisadores nessas áreas de atuação.

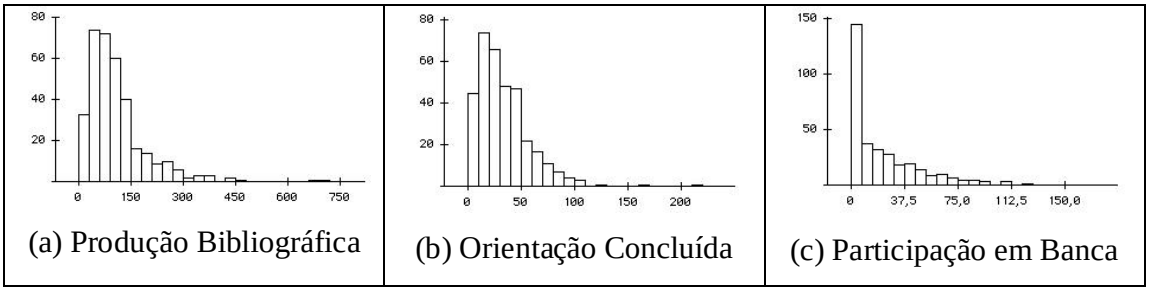


1 - Mineração de Dados
2 - Hardware
3 - Arquitetura de Computadores
4 - Banco de Dados
5 - Inteligência Artificial
6 - Análise e Complexidade de Algoritmos
7 - Redes de Computadores
8 - Teleinformática
9 - Simulação e Modelos Analíticos
10 - Sistema de Informação
11 - Processamento Gráfico
12 - Engenharia de Software
13 - Linguagem de Programação

Figura 16 - Distribuição dos Pesquisadores por Áreas da CAPES

A Figura 16 mostra que as áreas de maior interesse de pesquisa são Engenharia de Software, Sistemas de Informação e Banco de Dados. Isso demonstra que, possivelmente, o Brasil tem uma carência de pesquisas em áreas como Redes de Computadores, por exemplo.

Ao final do processo de análise dos atributos de perfil foram selecionados três atributos: número de produções bibliográficas, número de participação em banca e número de orientações concluídas. Os histogramas obtidos na primeira análise desses atributos estão ilustrados na Figura 17.



(a) Produção Bibliográfica (b) Orientação Concluída (c) Participação em Banca

Figura 17 - Histograma dos atributos de perfil não normalizados

Após a escolha dos atributos, iniciou-se a etapa de transformação dos dados. A primeira transformação aplicada foi a normalização dos dados para que todas as variáveis estivessem na mesma escala. Os dados foram normalizados para que seus valores estejam no intervalo $[0,1]$. Utilizando a função de normalização min-max definida na equação (2) e assumindo que $\text{novo\_min}_A = 0$ e $\text{novo\_max}_A = 1$, é obtida uma equação de normalização reduzida definida por:

$$\overline{v} = \frac{v - \min_A}{\max_A - \min_A}, \tag{3}$$

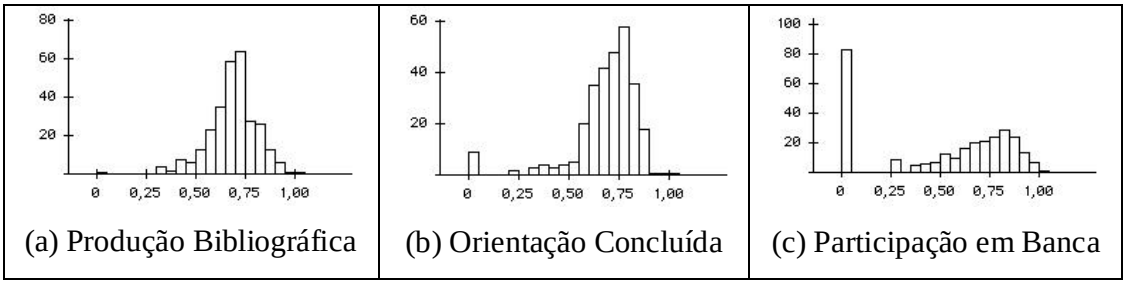
onde v é o valor original do atributo, $\min_A$ é o menor valor do atributo A em todo o conjunto de dados e $\max_A$ é o maior valor do atributo A em todo o conjunto de dados.

Como grande parte dos valores dos atributos está próximo de zero, como pode ser visto nos histogramas da Figura 17, após a normalização dos dados, a quantidade de valores tendendo a zero foi alta.

Assim, optou-se pela aplicação do logaritmo natural, definido na equação (4), antes da normalização dos atributos. O objetivo é fazer com que os dados tenham uma distribuição mais próxima de uma distribuição normal, reduzindo a quantidade de valores tendendo à zero.

$$\tilde{v} = \ln(v) \tag{4}$$

A Figura 18 mostra a distribuição dos atributos de produção bibliográfica, orientação concluída e participação em banca após a aplicação do logaritmo natural e posterior normalização dos mesmos. Observando esses histogramas é possível verificar que os atributos apresentam uma distribuição muito mais próxima de uma distribuição normal do que apresentavam anteriormente sem o uso do logaritmo natural.



(a) Produção Bibliográfica | (b) Orientação Concluída | (c) Participação em Banca

Figura 18 - Histograma dos atributos após transformação dos dados

Após a transformação dos dados, os mesmos foram analisados novamente para determinar possíveis *outliers* que ainda estavam no conjunto de dados. Como o atributo de participação em banca apresentou muitos valores iguais à zero, optou-se por não analisá-lo na busca por *outliers*.

Analisando os relacionamentos de orientações concluídas e de co-autoria dos pesquisadores foi possível corrigir os valores desses atributos que estavam iguais à zero. Os pesquisadores que não tiveram seus atributos corrigidos foram removidos do conjunto de dados.

Após os ajustes finais nos atributos e a exclusão de alguns poucos pesquisadores do conjunto de dados foi obtida uma normalização mais suave para os atributos de produções bibliográficas e de orientações concluídas, conforme ilustrado pelos histogramas e pelo *boxplot* dessas variáveis na Figura 19.

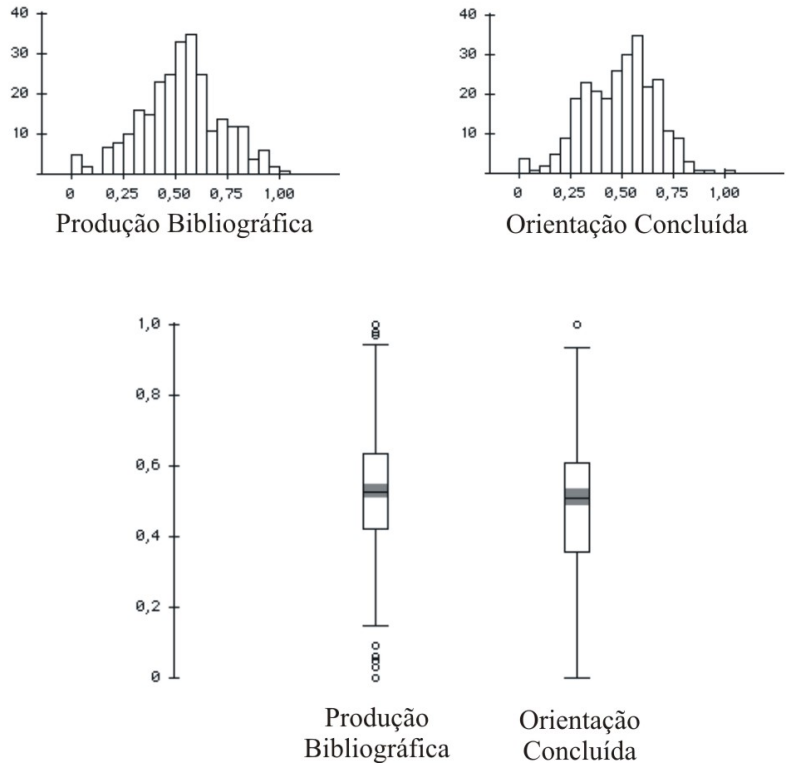


Figura 19 - Transformação final dos atributos de perfil.

4.3.3 Atributos de Relacionamento

A análise dos atributos de relacionamento depende de se estabelecer uma medida para diferenciar os relacionamentos mais fracos dos relacionamentos mais fortes, ou seja, depende da definição de um peso para os relacionamentos. A medida inicialmente adotada foi a contagem do número de relacionamentos existentes entre cada par de pesquisadores. Após essa definição dos pesos as mesmas análises feitas para os atributos de perfil foram feitas para os relacionamentos dos pesquisadores.

O problema é que a grande maioria dos professores possuía um grau de relacionamento muito fraco (cerca de 90%), com apenas uma produção bibliográfica, por exemplo. O gráfico que ilustra o resultado da contagem dos relacionamentos está ilustrado no histograma da Figura 20.

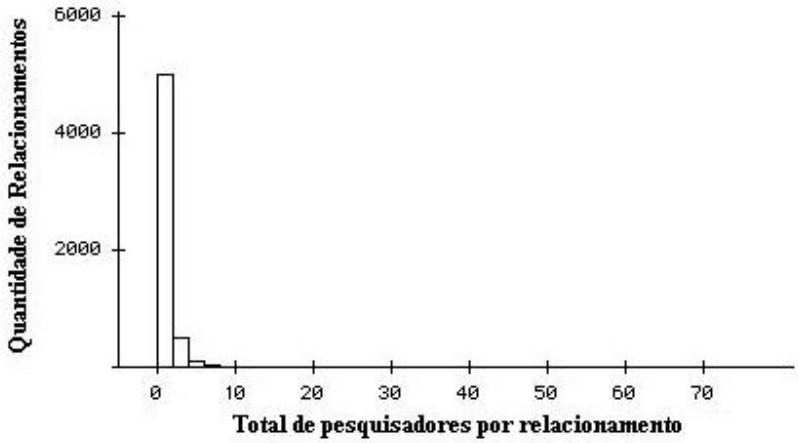


Figura 20 - Histograma dos relacionamentos antes da transformação dos dados

Pode ser observado na Figura 20 que grande parte dos pesquisadores possui poucos relacionamentos entre si. Assim, ao normalizar esses dados, muitos relacionamentos sumiram, pois seus valores tenderam a zero, formando um grafo desconexo que não poderia ser utilizado pelos algoritmos de mineração de dados.

A análise inicial dos dados de relacionamento permitiu identificar pesquisadores que não estavam relacionados a nenhum outro pesquisador na rede. Esses pesquisadores foram removidos do conjunto de dados para que a rede social científica produzisse um grafo conexo.

Outros critérios foram adicionados para a definição do peso dos relacionamentos. Os critérios para definir o peso final dos relacionamentos estão descritos em detalhes na próxima seção que descreve a modelagem da rede social científica multi-relacional.

Todos os critérios de transformação dos dados utilizados nos atributos de perfil foram utilizados nos atributos de relacionamentos. Assim, após aplicar o logaritmo natural e a transformação min-max sobre os pesos dos relacionamentos os dados apresentaram uma melhor distribuição e os pesos dos relacionamentos apresentaram valores mais significativos, conforme na Figura 21.

Analisando a Figura 21 é possível observar que embora a maioria dos relacionamentos continue sendo fracos, com essa normalização, eles passaram a ter um valor representativo comparados aos relacionamentos mais fortes. Com isso, foi obtido um grafo conexo no qual até mesmo os relacionamentos mais fracos continuaram sendo representados.

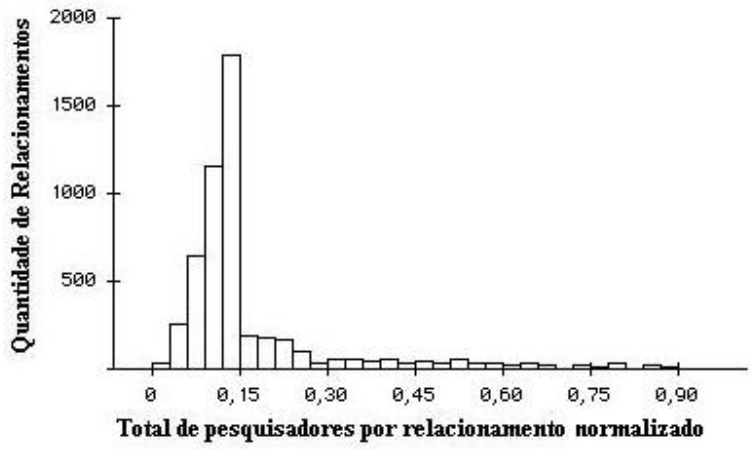


Figura 21 - Histograma dos relacionamentos após transformação dos dados

Na próxima seção será apresentado o processo de construção da Rede Social Científica Multi-relacional com peso, ou seja, a rede social científica será modelada de forma que cada conexão possua um peso diferente, representando os quão conectados dois pesquisadores estão.

4.4 Modelagem da Rede Social Científica Multi-Relacional

Como dito anteriormente, todos os dados utilizados nos experimentos iniciais foram retirados da plataforma Lattes, com o auxílio do GCC. Um exemplo de parte desse currículo pode ser visto na Figura 22.

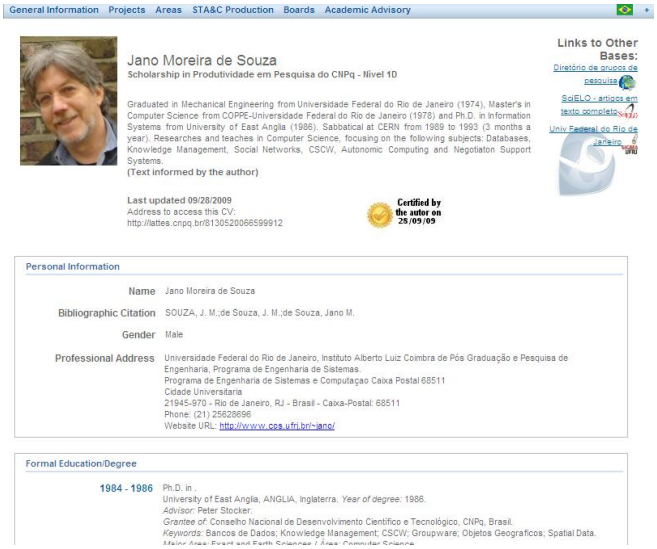


Figura 22 - Exemplo de um Corrículo Lattes

Na modelagem do relacionamento de uma rede social o peso do mesmo representa como o quão fortemente dois elementos estão conectados. O peso de uma ligação em uma rede social multi-relacional deve considerar o peso de todos os tipos de relacionamentos existentes entre os dois elementos.

O processo da modelagem da rede social multi-relacional foi dividido em três etapas: número de relacionamentos comuns entre os pesquisadores; a idade dos relacionamentos que ligam esses pesquisadores e perda de conhecimento quando o relacionamento entre os pesquisadores é indireto, ou seja, quando a conexão entre eles ocorre por intermédio de outros pesquisadores.

4.4.1 Número de Relacionamentos em Comum

Como dito na seção anterior, foi necessário estabelecer uma medida que pudesse diferenciar os relacionamentos mais fracos dos mais fortes. A primeira opção foi contar

os totais de participação em projetos, publicações de co-autoria, participação em banca de trabalhos científicos, e produções técnicas entre os pesquisadores. Através desse critério, definiu-se que pesquisadores com maior número de interações possuem relacionamentos mais fortes.

A fim de tentar aumentar relativamente o grau do relacionamento entre os pesquisadores e tentando refletir a realidade de cada um deles foi utilizada a fórmula representada na equação (5):

$$R_i = \frac{CR_i}{P1 + P2}, \quad i = 1, ..., t. \tag{5}$$

Onde R_i representa o grau de relacionamento 'i', CR_i é o número de relacionamentos comuns do tipo 'i' entre os pesquisadores 1 e 2, *P1* representa o total de relacionamentos do tipo 'i' do pesquisador 1, *P2* é o total de relacionamentos do tipo 'i' do pesquisador 2, e 't' é o total de tipos de relacionamento.

O número de relacionamentos comuns foi dividido pela soma total dos relacionamentos de cada pesquisador para que a força do relacionamento fosse relativa ao total de relacionamentos. O objetivo dessa medida é evitar que os relacionamentos com a mesma freqüência tenham a mesma força.

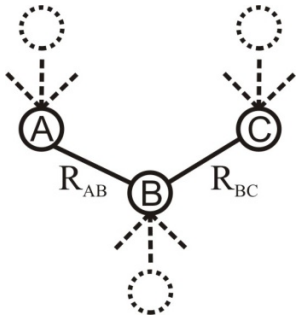


Figura 23 - Análise dos relacionamentos.

Para ilustrar a situação descrita anteriormente, foi representada na Figura 23 uma pequena parte de uma Rede Social. Nessa figura estão representados apenas dois relacionamentos R_{AB} e R_{BC} e três elementos **A**, **B** e **C**; todos os outros elementos e seus relacionamentos foram removidos para facilitar a ilustração.

Nesse exemplo, é assumido que o pesquisador **B** possui 5 relacionamentos em comum com o pesquisador **A** e 5 relacionamentos em comum com o pesquisador **C**, ou

seja, $R_{AB} = R_{BC} = 5$. Admita também que o número total de relacionamentos do elemento **A** é dado por $P_A = 10$, do elemento **B** é $P_B = 10$ e do elemento **C** é dado por $P_C = 50$. Embora o número de relacionamentos comuns seja igual, os pesos dos relacionamentos **AB** e **BC** não podem ser considerados iguais.

Como **A** e **B** possuem a maioria dos seus relacionamentos em comum pode-se considerar que o relacionamento entre eles é mais forte que o relacionamento entre **B** e **C**, já que o elemento **C** possui vários relacionamentos com outros pesquisadores. Partindo da equação (5) tem-se que $R_{AB} = 5/10 + 10 = 1/4$ enquanto que $R_{BC} = 5/10 + 50 = 1/12$. Assim, tem-se que o relacionamento entre os elementos **A** e **B** é mais forte que entre os elementos **B** e **C**, mesmo que em termos quantitativos eles sejam iguais.

Depois de aplicar a equação (5) para os quatro tipos de relacionamentos, foram somados todos os graus desses relacionamentos (R_i) e atribuídos pesos para cada um deles. A equação (6) representa essa etapa:

$$TR_{AB} = \sum_{i=1}^{t} \alpha_i R_i, \tag{6}$$

onde TR_{AB} significa o número total de relacionamentos entre os pesquisadores *A* e *B*, α_i é o peso dado ao relacionamento 'i'.

Neste trabalho foi adotado $\alpha = 1$ para todos os tipos de relacionamentos. Entretanto, a modelagem da rede social científica multi-relacional foi feita de forma que ela possa ser utilizada em qualquer problema de análise de rede social.

A fim de representar o conceito apresentado anteriormente de problemas com relacionamentos com pesos diferentes, considere a rede social formada por empresas e seus relacionamentos, ilustrada na Figura 24. Nesta rede social empresarial foram ilustradas seis empresas conectadas por três tipos de relacionamentos: concorrência, colaboração por parceria, e colaboração por interesse financeiro.

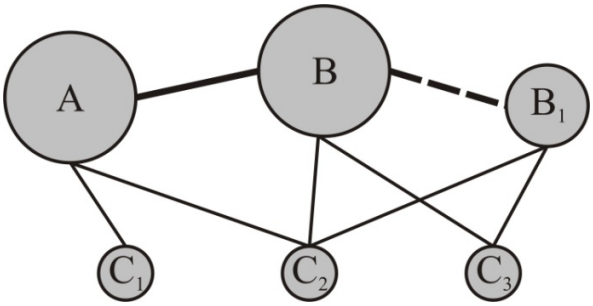


Figura 24 - Relacionamentos Colaborativos e Não-colaborativos.

O *relacionamento de concorrência*, representado pela conexão entre as empresas A e B, são relacionamentos existentes entre as empresas que concorrem entre si na prestação de serviços comuns.

O *relacionamento de colaboração por parceria* ocorre entre as empresas B e B1. Esse tipo de relacionamento ocorre entre empresas do mesmo grupo, como se B1 fosse uma filial de B.

Finalmente, a *colaboração por interesse financeiro* ocorre entre empresas que prestam serviços a outras empresas. Esses relacionamentos são ilustrados na Figura 24 pelas conexões entre as empresas C1, C2 e C3.

Nesse cenário podem ser considerados pesos diferentes para cada tipo de relacionamento dependendo da análise que se deseja realizar. Para analisar as empresas que trabalham juntas, deve ser adotado $\alpha < 0$ para os relacionamentos de competição, $\alpha > 1$ para os relacionamentos de colaboração por parceria, e $0 < \alpha < 1$ para os relacionamentos colaborativos com interesse financeiro.

Assim, se um novo problema tiver que ser modelado com pesos diferentes para cada tipo de relacionamento ($\alpha \neq 1$), a modelagem da rede social multi-relacional proposta neste trabalho ainda pode ser utilizada.

O resultado da equação (6) foi normalizado aplicando-se o logaritmo natural e normalização Min-Max, conforme apresentado na seção anterior. No final dessa etapa da modelagem um grafo conexo foi obtido.

4.4.2 Idade do Relacionamento

Outro fator importante a ser considerado na definição do grau de relacionamento é a idade do mesmo, ou seja, é importante saber o ano em que o relacionamento foi criado. A idade do relacionamento é útil para indicar se o relacionamento reflete uma conexão atual, ou se é apenas uma conexão que existia no passado e que talvez nem exista nos dias atuais.

Existem dois tipos de relacionamentos a serem considerados quando se olha para o ano em que a conexão ocorreu. O primeiro tipo é o *relacionamento exato* que ocorre em um determinado momento e a conexão entre os elementos não irá necessariamente continuar ao longo do tempo, como é o caso dos relacionamentos de co-autoria. O outro tipo é o *relacionamento contínuo* que é aquele que tem uma duração definida, ou seja, ele reflete uma coexistência dos elementos durante um intervalo de tempo, como é o caso dos relacionamentos de participação em projetos.

Para ilustrar a importância de analisar a idade de relacionamento, suponha que dois pesquisadores A e B tenham publicado três trabalhos há vinte anos, e dois outros pesquisadores C e D publicaram um artigo em conjunto no ano passado. Se for considerado apenas o número de publicações em comum, será concluído que A e B têm uma relação mais forte do que os pesquisadores C e D. No entanto, as relações entre A e B são muito antigas e, provavelmente, esses pesquisadores podem não estar trabalhando juntos nos dias atuais. Por outro lado, as conexões entre C e D são recentes, o que indica que eles atualmente têm interesses comuns.

Para considerar a idade dos relacionamentos na modelagem da rede social científica foi adicionado um peso do ano para os relacionamentos na equação (6), obtendo a seguinte equação:

$$TR_{AB} = \sum_{i=1}^{t} \sum_{j=1}^{d} \rho_j \alpha_i R_i, \tag{7}$$

onde *d* é a duração do relacionamento em anos e ρ_j é uma função de penalização com relação ao ano do relacionamento. Assume-se que os relacionamentos exatos possuem duração de um ano e os relacionamentos contínuos têm duração igual ao número de anos

que a conexão existiu, sendo que essa informação está disponível no currículo Lattes do pesquisador.

A definição da função de penalização com base no ano do relacionamento foi feita a partir da análise dos resultados gerados por três tipos de funções diferentes, são elas: função potência (Acar, Dunlavy and Kolda, 2009), função exponencial e função sigmóide.

Neste capítulo essas funções serão formalmente definidas para compor o processo de modelagem da rede social científica. A análise dos resultados produzidos por cada uma delas será apresentado em detalhes nos próximos capítulos. Com base nos resultados obtidos foi possível definir qual a função de penalização reflete melhor o comportamento das redes sociais científicas multi-relacionais.

A *função potência* foi sugerida por Acar em (Acar, Dunlavy and Kolda, 2009) e a sua definição matemática é dada por:

$$\rho_j = (1-\theta)^{BY-RY} \tag{8}$$

onde *BY* é o ano-base utilizado neste trabalho, sendo igual a 2011 (ano do cálculo dos experimentos), *RY* é o ano do relacionamento e o parâmetro $\theta \in (0,1)$ pode ser definido pelo usuário ou segundo experimentos realizados em um conjunto de treinamento. Essa função descreve uma curva representada graficamente pela Figura 25.

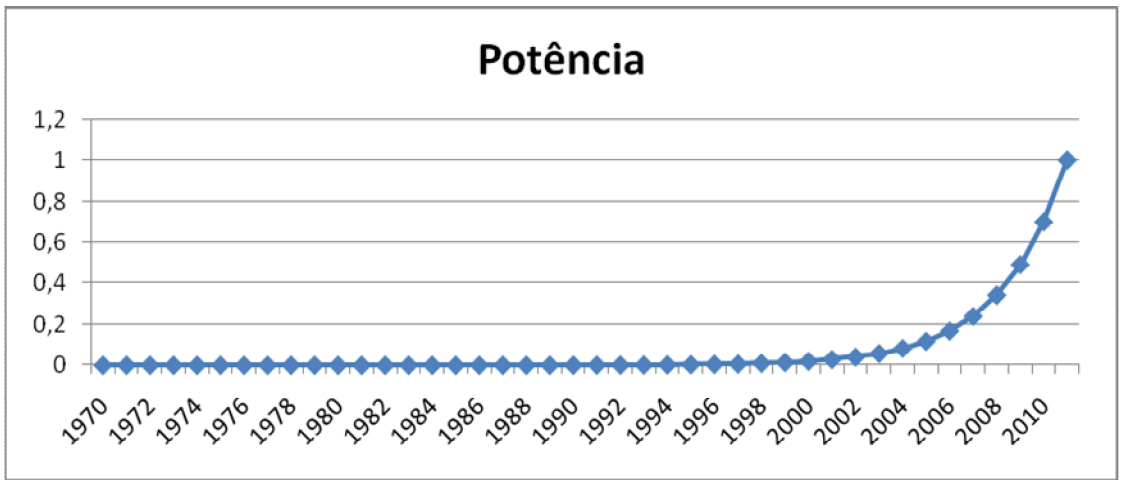


Figura 25 - Representação gráfica da função de penalização Potência

Valores de θ próximos a zero produzem uma função peso que penaliza rapidamente os relacionamentos mais antigos, enquanto que valores próximos a 1 penaliza pouco os relacionamentos antigos. As funções geradas pela variação do parâmetro θ estão representadas na Figura 26.

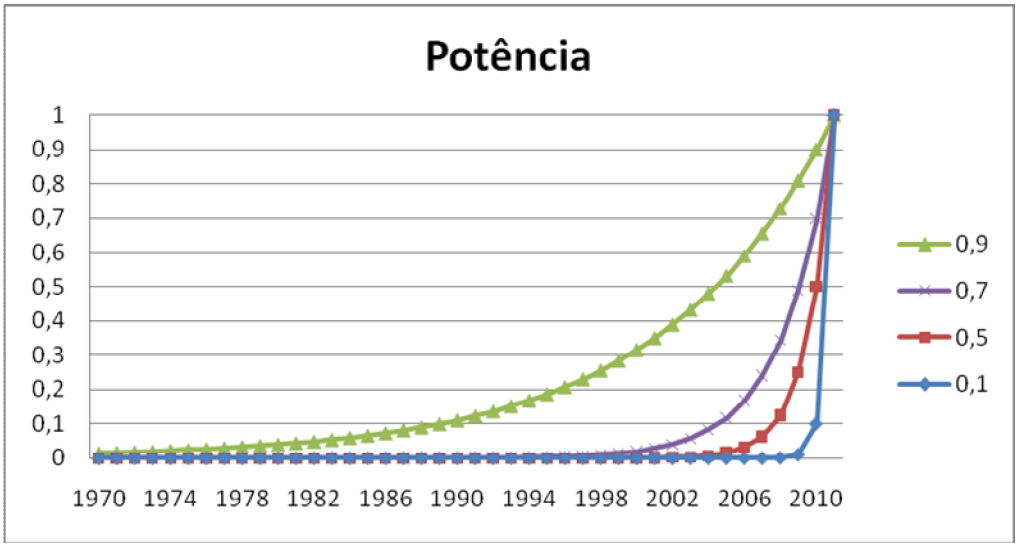


Figura 26 - Variação do parâmetro θ da função de penalização Potência

A definição do parâmetro θ não foi trivial. Essa definição foi feita a partir da análise dos resultados produzidos pelo método de agrupamento utilizando valores diferentes para esse parâmetro. A análise dos grupos formados mostrou que os melhores resultados foram obtidos para $\theta = 0,7$.

A *função exponencial* é definida como segue:

$$\rho_j = e^{1/(BY-RY)}, \tag{9}$$

onde *BY* e *RY* seguem a mesma definição da equação (8).

A função definida na equação (9) descreve a curva mostrada na Figura 27. Pode ser observado que, assim como na função potência, quanto mais recente for o relacionamento, maior será o peso dado a ele. Assim, tem-se a garantia de que os relacionamentos mais recentes possuem um peso maior na análise da rede social que os relacionamentos mais velhos.

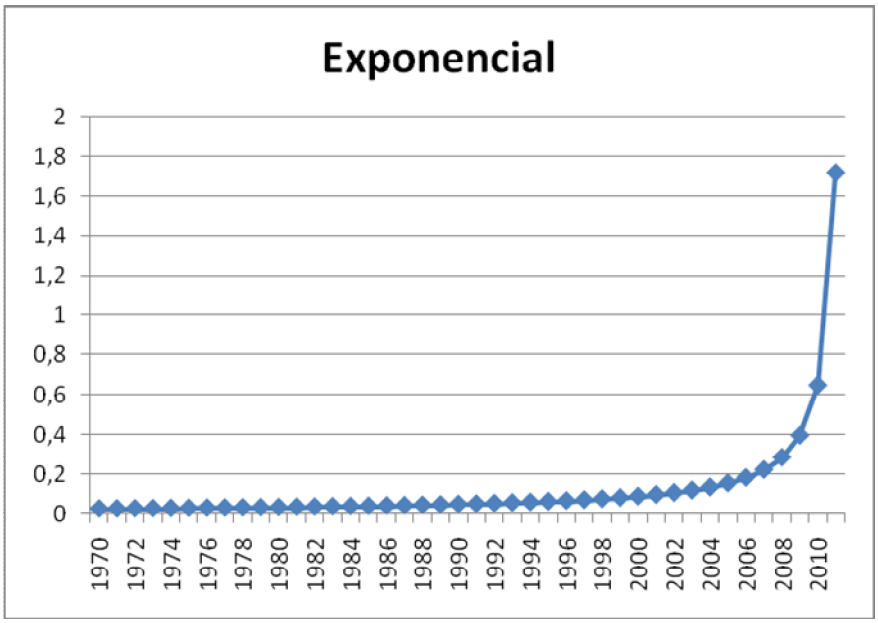


Figura 27 - Representação gráfica da função de penalização Exponencial

Durante a análise evolutiva da rede social científica, realizada durante o desenvolvimento do módulo de sugestão de relacionamentos (capítulo 7), foi verificado que alguns pesquisadores não se relacionavam com uma freqüência anual, ou seja, existem pesquisadores que se relacionam de dois em dois anos e até mesmo com freqüências maiores.

Esse comportamento foi identificado através da análise manual do conjunto de dados. Foi identificado, por exemplo, que dois pesquisadores publicam novos artigos com uma freqüência de dois em dois anos. Em alguns casos essa freqüência de publicação chega a ser de três anos. Assim, embora existam muitos pesquisadores que desenvolvem novos trabalhos anualmente, existem outros que não possuem uma freqüência tão alta.

Com o intuito de representar o comportamento descrito anteriormente foi definida a *função de penalização sigmóide*. Essa função de penalização possui valores bem próximos de 1 quando o ano do relacionamento é até 3 anos menor que o ano corrente. Os relacionamentos só começam a ser realmente penalizados quando a diferença entre o ano corrente e o ano do relacionamento é maior que 3. Assim, os pesquisadores que se relacionam com uma freqüência de dois em dois anos ou de três em três não seriam penalizados.

Como dito anteriormente, a freqüência de publicações foi identificada através da análise manual do conjunto de dados. Cerca de 30% do total de relacionamento não ocorre com uma freqüência anual, sendo que, desses 30%, cerca de 90% deles ocorrem com uma freqüência máxima de 3 anos. Embora existam relacionamentos que ocorreram com uma freqüência mair que 3 anos, a maioria dos pesquisadores publica novos trabalhos em no máximo 3 anos.

A *função sigmóide* é definida por:

$$\rho_j = \frac{1}{1+e^{(1-\theta)(RY-(BY-10))}} + 0{,}01 \qquad (10)$$

Onde $\theta \in (0,1)$ é o parâmetro que define a intensidade da curva da função, *BY* é o ano corrente e *RY* é o ano do relacionamento. O valor de 0,01 somado na função impede que a penalização seja nula para os relacionamentos muito antigos. Já o valor 10, no denominador da equação, indica o intervalo para que a curva da função decresça com maior intensidade, ou seja, a curva irá decrescer com maior intensidade entre os anos de 1998 e 2008. Esses valores foram obtidos através da análise dos resultados gerados para configurações diferentes. A curva descrita pela função sigmóide está ilustrada na Figura 28.

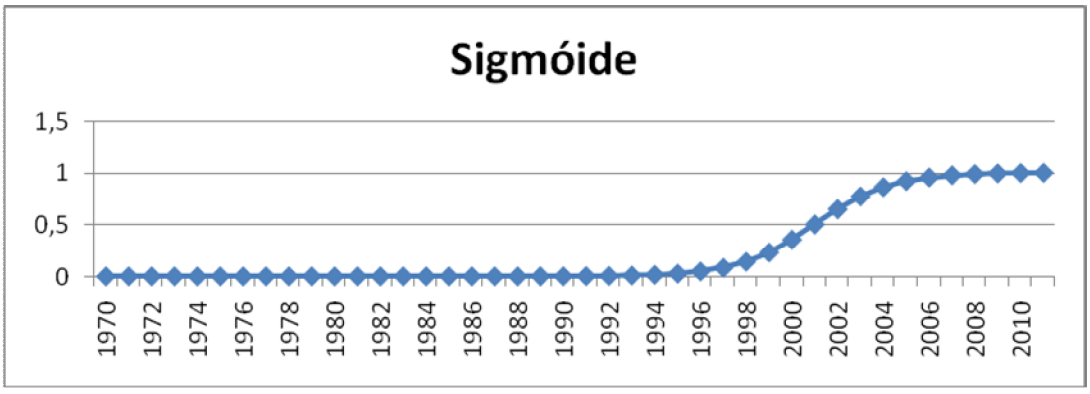


Figura 28 - Representação gráfica da função de penalização Sigmóide

Assim como na função potência, o valor de θ é definido pelo usuário ou através de experimentos feitos no conjunto de treinamento. Os experimentos realizados mostraram que $\theta = 0,4$ produz uma curva um pouco mais suave, gerando resultados consistentes na análise dos grupos formados pelo método de mineração de dados.

O objetivo das funções de penalização, como o próprio nome já diz, é penalizar os relacionamentos muito antigos de tal forma que, caso não surjam novos relacionamentos entre os pesquisadores a ligação entre eles receberá uma penalização maior a cada ano até ser completamente eliminada da rede social científica.

Embora tenham sido sugeridas três funções de penalização com base no ano do relacionamento, a função exponencial foi a que apresentou melhores resultados. A análise dos resultados de cada uma dessas funções está no capítulo 5 e também no capítulo 8.

Após aplicar a equação (7) em todos os relacionamentos foi construída uma matriz de pesos *MxM* que representa o grau de relacionamento entre cada par de pesquisadores de uma rede social científica multi-relacional, onde *M* é o número de pesquisadores do conjunto de dados. Como o grau do relacionamento representa o peso entre os pesquisadores essa matriz é chamada de *matriz de pesos* e está representada na equação (11).

$$MP = \begin{cases} TR_{AB} & \text{se } A \text{ se relaciona com } B \\ 0 & \text{caso contrário.} \end{cases} \tag{11}$$

4.4.3 Perda de Informação em Relacionamentos Longos

Outro conceito introduzido na modelagem da rede social científica multi-relacional é a perda de informações quando o caminho entre o pesquisador de origem e o pesquisador de destino é muito grande. Acredita-se que todo conhecimento que é passado de um indivíduo para outro tem alguma perda de conteúdo. O objetivo desta etapa na modelagem da rede social científica é reproduzir o efeito “telefone sem fio”, no qual a informação sempre chega com alguma distorção em seu destino final.

Há muitas razões para explicar a perda de conhecimento durante a transferência da informação, tais como: erros nas informações transferidas; transferência incompleta de informações; interpretação errada do conhecimento passado; desejo de reter parte do conhecimento adquirido para autoproteção; disputa por conhecimento, etc.

Tentando refletir a perda de conteúdo durante a troca de informações, foi considerado que o nó receptor recebe a informação com uma perda de *N%* do total do

conhecimento de que ele poderia receber, onde *N* é o número de nós intermediários entre a fonte e o receptor.

A idéia é adicionar uma resistência ao fluxo de conhecimento, quando esse conhecimento está passando por pelo menos um elemento intermediário, ou seja, o caminho entre a fonte e o receptor é maior do que um.

O objetivo do uso da resistência na modelagem da rede social é tornar esse modelo da rede multi-relacional o mais próximo possível da realidade. Assim, assumindo que o conhecimento máximo (grau do relacionamento), que pode ser transmitido entre dois pesquisadores *A* e *B*, seja dado por $MaxFlow_{AB}$, então o novo grau do relacionamento entre eles será dado como segue:

$$\overline{TR} = MaxFlow_{AB} - \frac{N * MaxFlow_{AB}}{100} \tag{12}$$

onde *N* é o número de elementos intermediários entre *A* e *B*, e $MaxFlow_{AB}$ é o fluxo máximo calculado utilizando a matriz de pesos (equação (11)).

No final do processo de modelagem da rede social científica multi-relacional é obtida uma *matriz de fluxo máximo* com resistência, a qual será utilizada pelo método de mineração de dados. O algoritmo que calcula o fluxo máximo entre os pesquisadores será apresentado no próximo capítulo.

Capítulo 5 – Algoritmos de Agrupamento

Nesta seção serão apresentadas as duas técnicas de agrupamentos utilizadas durante o desenvolvimento deste trabalho. A primeira delas busca por grupos de pesquisadores considerando tanto os atributos de perfil quanto os atributos de relacionamentos no processo de mineração dos dados. Já a segunda técnica utiliza as informações sobre o fluxo máximo na rede social para identificar as comunidades científicas.

As vantagens e desvantagens de cada uma delas serão enumeradas para justificar a escolha de apenas um desses dois algoritmos para a análise das redes sociais científicas multi-relacionais.

5.1 Algoritmo de Agrupamento – Árvore Geradora Mínima

Nesta seção, será apresentado o primeiro método desenvolvido neste trabalho para a detecção de grupos em Redes Sociais. Esse método visa identificar grupos de pessoas em redes sociais, que têm perfis semelhantes, e um forte relacionamento entre elas. Esse método utiliza uma abordagem em grafo que reduz o problema de agrupamento a um problema de particionamento em grafo.

As redes sociais utilizadas neste trabalho foram modeladas como um grafo chamado de grafo social. Em um grafo social cada nó representa uma pessoa e cada aresta mantém uma relação social existente entre duas pessoas. A fim de identificar grupos de pessoas, o método segue uma estratégia baseada na poda das arestas do grafo social.

O algoritmo de agrupamento utilizado é baseado no método de aglomeração espacial descrito em (Assunção, Neves, Câmara and Freitas, 2006). A topologia espacial pode ser entendida como uma rede social de objetos espaciais, na qual os objetos espaciais são os nós e seus relacionamentos são as arestas da rede social. Assim, é possível usar as etapas do método de agrupamento espacial para identificar os grupos em redes sociais.

A principal diferença entre o método utilizado neste trabalho e o método original é a modelagem do gráfico de entrada. O peso das arestas usado no algoritmo de agrupamento espacial é uma medida de dissimilaridade entre os nós ligados por essas arestas. No método utilizado, o peso da aresta é definido pelos atributos de perfil, conforme apresentado no capítulo anterior. O algoritmo é composto de três etapas, que serão apresentadas nas próximas subseções.

5.1.1 Modelagem do Grafo Social

Os dados nas redes sociais podem ser divididos em dois tipos: perfil de cada pessoa e os relacionamentos entre essas pessoas. O perfil de cada pessoa é formado por um vetor que armazena as características ou atributos das pessoas, tais como idade, peso, altura, etc. Assim, as características e os atributos descrevem uma pessoa, já os relacionamentos representam uma relação social existente entre duas pessoas.

A etapa de modelagem consiste em transformar a rede social em um grafo chamado grafo social. Essa modelagem diz respeito aos dois tipos de dados apresentados anteriormente. No grafo social, cada nó representa uma pessoa e cada aresta indica uma conexão entre duas pessoas.

A força do relacionamento social é representada por pesos nas arestas. Portanto, o grafo social tem pesos em todas as suas arestas, os quais devem pertencer ao intervalo (0,1), onde os pesos perto de 1 indicam uma forte relação e o oposto, ou seja, o peso próximo de 0, indica um relacionamento fraco. A regra utilizada para o cálculo dos pesos foi descrita no Capítulo 4.

No desenvolvimento desse algoritmo foram utilizados os seguintes atributos de perfil: número de orientações, número de participações em bancas, e número de publicações (congressos e revistas). Foi utilizado inicialmente apenas o número de publicações em comum dos pesquisadores como critério para dar pesos aos relacionamentos, em um segundo momento foram usados, além do relacionamento de co-autoria, mais 3 tipos de relacionamentos diferentes. Os atributos de perfil e os relacionamentos também foram definidos detalhadamente no Capítulo 4.

No algoritmo de agrupamento via árvore geradora mínima, o peso final das arestas do grafo social será definido pela distância euclidiana entre os elementos e pelos pesos dos relacionamentos. O peso final de uma aresta *l* do grafo social é dado como segue:

$$custo(l) = d(x_i, x_j).(2 - peso(x_i, x_j)) \quad (13)$$

onde x_i e x_j são os elementos incidentes na aresta *l*; $d(x_i, x_j)$ é a distância euclidiana entre esses elementos; e $peso(x_i, x_j)$ é o peso do relacionamento desses elementos.

Como $peso(x_i, x_j) \subset [0,1]$, faz-se dois menos esse valor para que a distância dos elementos seja mantida constante, caso o peso do relacionamento seja muito forte (igual a 1), ou para que a distância dobre, caso o peso do relacionamento seja muito fraco (igual a 0). Dessa maneira, o custo da aresta será sempre maior ou igual à distância euclidiana entre os elementos.

Vale observar que neste algoritmo o peso final da aresta recebe uma forte influência dos atributos de perfil dos elementos, pois o cálculo da distância euclidiana é obtido através desses dados.

5.1.2 Árvore Geradora Mínima

Nesta etapa, a Árvore Geradora Mínima (AGM) é construída. O agrupamento de pessoas em Redes Sociais pode ser definido assim como o problema de identificar subgrafos em um grafo. Considerando que o particionamento em grafo é um problema NP – difícil, para reduzir a complexidade, a AGM foi gerada utilizando-se os pesos das arestas do grafo.

Existem vários algoritmos para a geração da AGM. O algoritmo PRIM é um deles (Cormen, Leiserson, Rivest and Stein, 2001), e consiste em remover as arestas de acordo com seus pesos. No grafo social, a AGM representa a rede social com apenas os relacionamentos mais fortes. Com a criação da árvore geradora mínima, ao ir removendo as arestas, o grafo vai se tornando desconexo e os grupos vão sendo formados. Na próxima seção o processo de remoção das arestas é explicado mais detalhadamente.

5.1.3 Poda da Árvore Geradora Mínima

O objetivo nessa etapa é remover algumas arestas da AGM para encontrar os grupos de pessoas que estejam fortemente relacionadas e que possuam perfis semelhantes. Para isso, calcula-se o custo das arestas relacionadas à homogeneidade do grupo. Cada vez que uma aresta é removida é formado um novo grupo de pessoas.

A dificuldade nessa etapa é estabelecer os critérios para selecionar as arestas que devem ser eliminadas. Neste trabalho, esse critério foi definido com a seguinte fórmula:

$$custo(l) = SSD_T - SSD_l, \tag{14}$$

onde SSD_T é a soma dos desvios quadráticos dos atributos de perfil na árvore T e l é uma aresta dessa árvore. Já SSD_l é obtido com a seguinte fórmula:

$$SQD_T = \sum_{j=1}^{m}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2, \tag{15}$$

onde n é o número de nós (pessoas) na árvore T, x_{ij} é o j^{th} atributo da pessoa i^{th}, m é o número total de atributos de perfil considerados no agrupamento, e $\bar{x}_j$ é o valor médio do atributo j^{th} entre todos os indivíduos da árvore.

SSD_l é a soma dos desvios quadráticos das duas sub-árvores que estão conectadas pela aresta l. Esse cálculo é feito como segue:

$$SSD_l = SSD_{T_a} - SSD_{T_b} \tag{16}$$

onde SSD_{T_a} é a soma dos desvios quadrados da árvore T_a e SSD_{T_b} é a soma dos desvios quadrados da árvore T_b, como mostrado na Figura 29.

Considerando o que foi exposto anteriormente, pode-se dizer que o custo de uma aresta representa uma medida de homogeneidade. Dessa forma, as arestas com os custos mais elevados são as candidatas a serem podas. Depois que uma aresta é removida, os cálculos dos custos devem ser refeitos para todas as arestas do grafo social podado, pois a ausência da aresta removida afeta os resultados do cálculo.

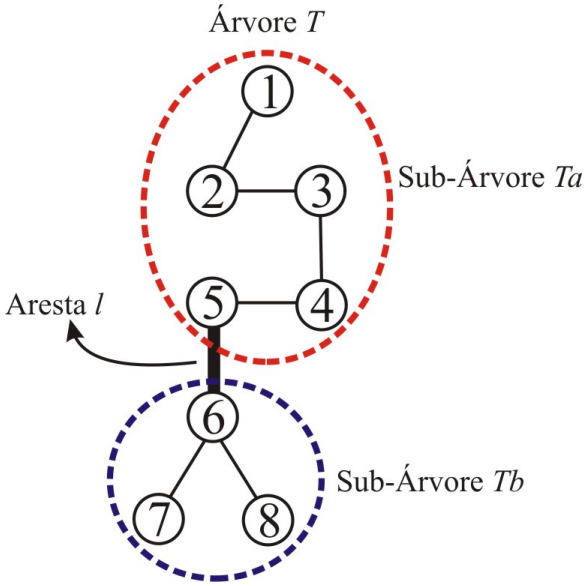


Figura 29 - Cálculo da aresta *l*.

O custo da aresta foi definido baseando-se na função objetivo do algoritmo *k-medoids* (Newman, 2001c). Portanto, a poda da aresta com o custo mais alto aumenta a homogeneidade das sub-árvores resultantes, ou seja, ele gera grupos de pessoas mais homogêneos.

5.2 Algoritmo Agrupamento: Fluxo Máximo

Nesta seção será apresentado o segundo algoritmo desenvolvido para solucionar o problema de identificação de comunidades em Redes Sociais Científicas. O objetivo principal desse algoritmo é considerar no processo de definição dos grupos apenas as informações sobre os relacionamentos entre os pesquisadores.

Esse algoritmo visa agrupar os pesquisadores avaliando apenas os relacionamentos entre eles. Assim, os atributos de perfil serão analisados apenas em uma segunda etapa, não influenciando o processo de agrupamento. O objetivo principal desse algoritmo é manter no mesmo grupo os pesquisadores com maior fluxo de troca de informações entre si.

5.2.1 Métrica de Similaridade: Relacionamentos

O objetivo do método é identificar grupos de pessoas no grafo social que têm uma forte relação entre si. A fim de identificar esses grupos o método tem a estratégia de analisar o fluxo das informações na rede social. Assim, as pessoas que possuem um grande fluxo de informações entre si tendem a pertencer ao mesmo grupo.

O método proposto é baseado no problema de *Fluxo Máximo* em redes (Boaventura, 1996). O peso de cada aresta do grafo social é uma medida de similaridade entre os nós conectados pelas arestas. Neste método, o peso das arestas é definido pelos atributos dos relacionamentos entre os pesquisadores. A definição dos pesos pode ser analisada com mais detalhe no capítulo 4.

Assim como no algoritmo de agrupamento via árvore geradora mínima, durante o desenvolvimento desse algoritmo foi considerado que a Rede Social Científica Multi-relacional é um grafo, onde cada nó representa um pesquisador e cada aresta representa um tipo de relacionamento científico entre dois pesquisadores. O relacionamento representa uma relação social científica existente entre duas pessoas. Essa relação social pode ser forte, entre duas pessoas, ou fraca entre duas outras. A força do relacionamento depende da medida de avaliação escolhida para o mesmo.

Como o objetivo do método de mineração de dados é identificar grupos de pessoas com uma relação forte (bom fluxo conhecimento) entre elas, esse método analisa o fluxo de conhecimento na rede social, de forma que as pessoas com um grande fluxo de informações entre elas fiquem no mesmo grupo.

5.2.2 Descrição do Algoritmo

O problema de Fluxo em Redes pode ser interpretado como um problema de Fluxo em Grafo. O Grafo Social com o Fluxo será representado por $\mathbf{G} = (X, U, \mathbf{f})$, no qual **f** é um vetor de dimensão *m+1* e pode ser escrito como segue:

$$\mathbf{f} = (f_0, f_1, \ldots, f_m) \tag{17}$$

O vetor **f**, representado na equação (17) é o fluxo no grafo **G** e cada uma das componentes indicam o valor do fluxo entre os elementos de **G**. O grafo social é

representado nesse trabalho por um grafo não-orientado. Assim, o fluxo máximo saindo de x_i para x_j é igual ao fluxo máximo de x_j para x_i, para todo $x_i, x_j \in X$.

O algoritmo de agrupamento desenvolvido neste trabalho utiliza o fluxo máximo entre dois pesquisadores como uma métrica de similaridade. Considere **G** como sendo o grafo social, que representa a Rede Social Científica Multi-relacional, **X** é o conjunto de pesquisadores, **U** é o conjunto de relacionamentos entre esses pesquisadores, e **f** é o conjunto de fluxos máximos entre cada par de pesquisadores. Assim, para todo $x_i, x_j \in X$ o fluxo máximo entre esses dois pesquisadores será igual a f_w, onde $0 \leq w \leq m$.

O cálculo do fluxo máximo entre os elementos do conjunto de dados foi feito com auxílio do algoritmo de *Edmonds-Karp* (Cormen, Leiserson, Rivest and Stein, 2001; Edmonds and Karp, 1972). Esse algoritmo é uma variação do algoritmo de fluxo máximo de *Ford-Fulkerson* (Boaventura, 1996; Cormen, Leiserson, Rivest and Stein, 2001; Ford and Fulkerson, 1956), descrito na Figura 30.

A principal diferença entre essas duas abordagens é que o algoritmo de Edmonds-Karp (Zadeh, 1972) visa o fluxo máximo entre dois elementos para o caminho mais curto. Assim, é garantido que o algoritmo irá convergir em um número finito de iterações, mesmo para os grafos não-orientados.

Algoritmo de Fluxo Máximo: Ford–Fulkerson

Entrada: Grafo **G** com fluxo máximo *c*, um nó origem x_s, e um nó sorvedouro x_t, onde $x_s, x_t \in X$.

Saída: Um fluxo f_i de x_s a x_t, o qual é máximo, onde $f_i \in \mathbf{f}$.

1. $f(x_u, x_v) \leftarrow 0$ para todas as arestas (x_u, x_v)
2. Enquanto existirem caminhos p de x_s até x_t em G_f, tal que $c_f(x_u, x_v) > 0$ para toda aresta $(x_u, x_v) \in p$:
   a. Encontre $c_f(p) = \max\{c_f(x_u, x_v) \mid (x_u, x_v) \in p\}$
   b. Para cada aresta $(x_u, x_v) \in p$
      i. $f(x_u, x_v) \leftarrow f(x_u, x_v) + c_f(p)$ (*Envia o fluxo ao longo do caminho*)
      ii. $f(x_v, x_u) \leftarrow f(x_v, x_u) - c_f(p)$ (*O fluxo pode ser 'devolvido' mais tarde*)

Figura 30 - Algoritmo Ford-Fulkerson.

O problema do fluxo máximo é definido como encontrar o máximo fluxo possível de algum nó fonte a um nó de destino. Para saber o fluxo máximo do grafo ilustrado na Figura 31, é necessário analisar quantas unidades de fluxo cada nó pode passar para outro nó.

Para facilitar a visualização e a análise da transferência do fluxo cada aresta foi colocada de uma cor. Com base na Figura 31, pode ser observado que o nó 1 pode passar até dez unidades para os nós 2, 3 e 4, esse fluxo está representado pelas arestas vermelha, azul e verde, respectivamente.

O nó 2 pode passar até 5 unidades para ambos os nós 6 e 7. Como o fluxo que o nó 2 está repassando é o mesmo fluxo enviado pelo nó 1, as arestas também foram coloridas da mesma cor.

O nó 3 pode passar até seis unidades de fluxo para o nó 6, e quatro unidades de fluxo para o nó 7. Da mesma maneira que no nó 2, as arestas desse nó foram coloridas da mesma cor da aresta que originou o fluxo (azul).

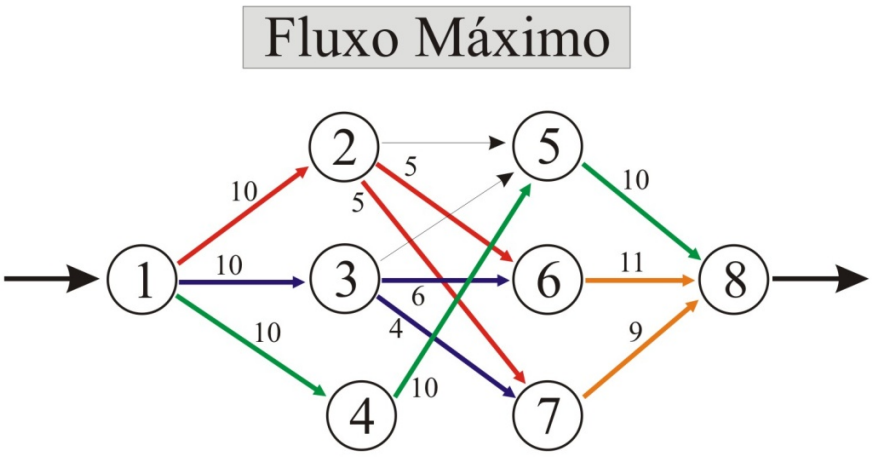


Figura 31 - Exemplo de Fluxo Máximo.

O nó 4 pode passar até dez unidades de fluxo para o nó 5, ou seja, todo o fluxo recebido por ele pode ser repassado para o nó 5. Como o nó 5 também pode repassar todo o fluxo recebido para o nó receptor 8, a sua aresta foi colorida de verde, que é a mesma cor da aresta que originou o fluxo no nó 1.

Os nós 6 e 7 podem passar até onze e nove unidades de fluxo cada um para o nó receptor 8, respectivamente. Como o fluxo que está sendo repassado por esses elementos é uma mistura dos fluxos enviados pelos nós 2 e 3 as arestas que representam a saída de fluxo desses nós foram coloridas de laranja.

Assim, somando o fluxo que cada nó pode transferir, pode-se concluir que o fluxo máximo da rede social da Figura 31, representado pelo fluxo entre os nós 1 e 8, é igual a 30 unidades.

Vale observar que os nós 2 e 3 poderiam enviar algum fluxo para o nó 5. Entretanto, o nó 5 só consegue repassar 10 unidades de fluxo para o nó 8, que já estão sendo enviadas pelo nó 4. Assim, se o nó 2 ou o nó 3 passarem algum fluxo para o nó 5, o nó 4 não será capaz de passar toda a sua capacidade de fluxo e, conseqüentemente, o fluxo total da rede irá diminuir.

A fim de juntar o fluxo máximo de conhecimento da rede social e o algoritmo de agrupamento, foi calculado o fluxo máximo entre todos os pares de pesquisadores da rede social científica multi-relacional. O cálculo do fluxo máximo foi feito usando a matriz de pesos construída no Capítulo 4. Como dito anteriormente, essa matriz representa o grau de similaridade entre os pesquisadores.

O algoritmo de fluxo máximo foi aplicado na matriz de pesos que representa a rede social científica multi-relacional. Como resultado do algoritmo, foi obtido uma matriz de fluxo máximo que representa o fluxo entre todos os pares de pesquisadores da rede social. Essa matriz de fluxos máximos será utilizada pelo algoritmo de agrupamento a fim de identificar as comunidades científicas.

O objetivo do algoritmo desenvolvido é agrupar os pesquisadores que possuam o maior fluxo de informação entre si. A fim de auxiliar o desenvolvimento do algoritmo, foi utilizado o algoritmo *k-Medoids* (Han and Kamber, 2006) como base para o desenvolvimento.

No algoritmo desenvolvido, assim como no algoritmo K-Medoids, são definidos aleatoriamente, na primeira etapa, os ***k*** medóides e cada um deles é associado a um grupo. Na segunda etapa, cada elemento do conjunto de dados será associado ao grupo com o qual esse elemento possui a melhor comunicação, ou seja, ele é associado com o

grupo que ele possui o maior fluxo de informação. Na terceira etapa, mais uma vez são definidos os medóides de cada grupo.

A definição dos medóides nesse algoritmo é baseada nas habilidades de comunicação de cada pesquisador dos grupos. Todos os fluxos de conhecimento de cada pesquisador interno ao grupo são somados, o pesquisador que possuir a maior soma é considerado o medóide do grupo. Após a definição dos novos medóides, volta-se para a segunda etapa. Como no algoritmo K-medoids, esse processo continua até que não ocorram alterações na estrutura dos grupos.

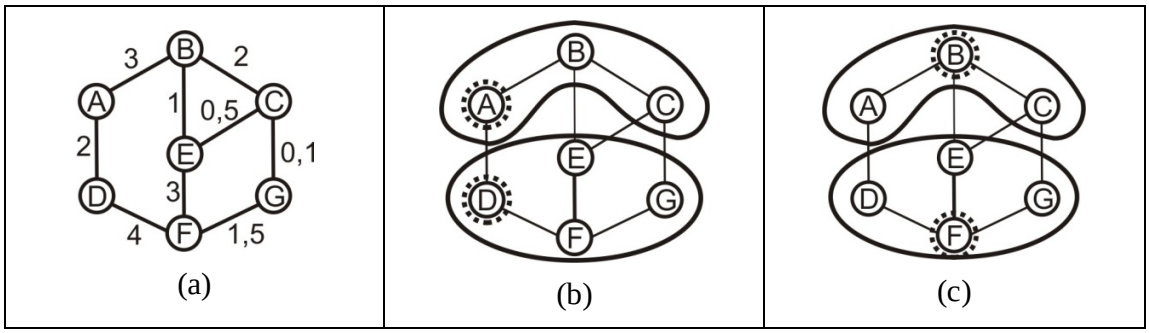


Figura 32 - Etapas para a definição dos medóides.

Para um melhor entendimento do processo de definição dos medóides considere a rede social representada pelo grafo da Figura 32 e sua respectiva matriz de adjacência dada pela Tabela 2.

Tabela 2 - Matriz de Adjacência

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	-	3	0	2	0	0	0
B	3	-	2	0	1	0	0
C	0	2	-	0	0,5	0	0,1
D	2	0	0	-	0	4	0
E	0	1	0,5	0	-	3	0
F	0	0	0	4	3	-	1,5
G	0	0	0,1	0	0	1,5	-

Com o auxílio da matriz de adjacência é definida uma matriz com os fluxos máximos de cada elemento calculados segundo o algoritmo apresentado anteriormente. A Tabela 3 representa essa matriz de fluxos máximos e o tamanho do caminho entre os elementos. O tamanho do caminho é utilizado como critério de desempate na definição

do medóide mais próximo ao elemento. Assim, caso o fluxo máximo entre um elemento e os medóides forem iguais, o elemento será associado ao grupo do medóide que estiver mais próximo a ele.

Nesse exemplo está sendo considerado $k = 2$. No primeiro passo do algoritmo, como descrito anteriormente, os medóides são definidos de forma aleatória e o elemento com maior fluxo de comunicação com um determinado medóide é associado ao grupo do mesmo. Esse passo é demonstrado na Figura 32(b).

Tabela 3 - Matriz de Fluxo Máximo/Caminho Mínimo

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	-	4,6/1	2,6/2	3,6/1	3,6/2	3,6/2	1,6/3
B	4,6/1	-	2,6/1	3,6/2	3,6/1	3,6/2	1,6/2
C	2,6/2	2,6/1	-	2,6/3	2,6/1	2,6/2	1,6/1
D	3,6/1	3,6/2	2,6/3	-	4,5/2	5,6/1	1,6/2
E	3,6/2	3,6/1	2,6/1	4,5/2	-	4,5/1	1,6/2
F	3,6/2	3,6/2	2,6/2	5,6/1	4,5/1	-	1,6/1
G	1,6/3	1,6/2	1,6/1	1,6/2	1,6/2	1,6/1	-
Fluxo Interno	7,2/3	**7,2/2**	5,2/3	11,7/5	10,6/5	**11,7/3**	4,8/5

Nesse exemplo foram definidos os elementos **A** e **D** como medóides iniciais. Analisando a matriz de fluxos máximos da Tabela 3 observa-se que o fluxo máximo de comunicação do elemento **B** é com o elemento **A** e, por isso, ele foi associado ao grupo do elemento **A**. Já o elemento **C** possui o mesmo fluxo com os elementos **A** e **D**. Entretanto o tamanho do caminho entre os elementos **C** e **A** é menor que entre os elementos **C** e **D**. Por isso, o elemento **C** foi associado ao grupo do elemento **A**.

Os elementos **E** e **F** foram associados ao grupo do elemento **D**, pois eles possuem um fluxo de informação melhor com o elemento **D** do que com o elemento **A**. O elemento **G**, assim como o elemento **C**, possui fluxo máximo igual com os elementos **A** e **D**, mas o tamanho do caminho entre **G** e **D** é menor e, por isso, ele foi associado ao grupo do elemento **D**.

No próximo passo do algoritmo os medóides são redefinidos baseando-se no seguinte critério: o elemento com o maior poder de comunicação no grupo é o novo medóide. Para encontrar os novos medóides, o fluxo e o tamanho do caminho entre todos os elementos do mesmo grupo são somados e aquele elemento que possuir o melhor fluxo

interno com o menor caminho é considerado o novo medóide. Assim, embora os elementos **A** e **B** tenham o mesmo fluxo de comunicação analisando a última linha da Tabela 3, o elemento **B** será o novo medóide já que ele tem a melhor comunicação com o menor caminho. O mesmo caso ocorre com os elementos **D** e **F**.

Após serem definidos os novos medóides os elementos são redistribuídos nos grupos, conforme representado na Figura 32 (c). Os medóides são calculados novamente, como não houve alterações nos grupos o algoritmo termina.

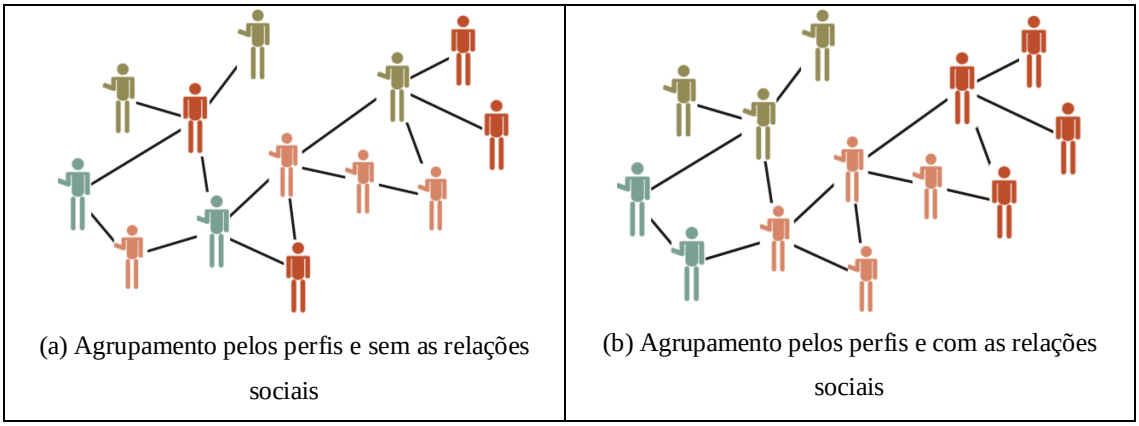
5.3 Árvore Geradora Mínima X Fluxo Máximo

Como apresentado anteriormente, o algoritmo de agrupamento por árvore geradora mínima utiliza tanto os atributos de perfil dos pesquisadores quanto o grau do relacionamento entre eles para identificar as comunidades científicas. Por outro lado, o algoritmo de agrupamento por fluxo máximo utiliza apenas as informações sobre o fluxo de conhecimento entre os pesquisadores.

O uso do perfil dos objetos é uma vantagem do método de agrupamento por árvore geradora mínima dependendo do tipo de problema que está sendo solucionado. Essa técnica deve ser utilizada quando se deseja agrupar objetos com perfil semelhante, mantendo no mesmo grupo apenas objetos que estejam relacionados de alguma maneira.

Por exemplo, considere um trabalho de propaganda para um estado do Brasil gastando o mínimo possível. Essa propaganda será diferenciada dependendo do perfil de cada pessoa (idade, sexo e salário). Caso as pessoas sejam agrupadas considerando apenas os atributos de perfil de cada uma delas pode acontecer de pessoas de regiões diferentes do estado pertençam ao mesmo grupo e assim, não é possível traçar uma estratégia de marketing por grupo, essa situação está ilustrada na Figura 33(a).

Entretanto, se for considerado, na etapa de agrupamento, o relacionamento entre essas pessoas, tem-se a garantia de que todas as pessoas do mesmo grupo estão relacionadas, conforme ilustrado na Figura 33(b). Assim, bastaria investir em propagandas para algumas pessoas que sejam influentes no grupo e elas, naturalmente, distribuiriam a informação para o restante do grupo, já que todas as pessoas estão relacionadas.



(a) Agrupamento pelos perfis e sem as relações sociais

(b) Agrupamento pelos perfis e com as relações sociais

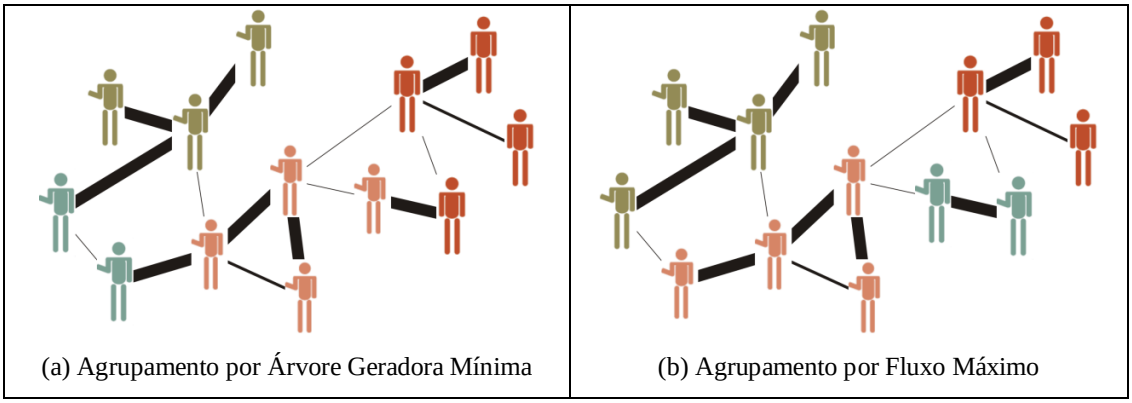
Figura 33 - Exemplo do algoritmo de agrupamento por árvore geradora mínima.

Existem alguns problemas nos quais as relações entre os objetos são mais importantes que a similaridade entre os seus atributos de perfil, como é o caso de alguns problemas de análise em Redes Sociais. Isto é devido ao fato de que dois elementos podem ter atributos diferentes, mas estarem fortemente relacionados através de algum tipo de relacionamento. Por exemplo, considere um conjunto de dados no qual os objetos são as pessoas e os atributos de perfil são cor e salário. Nesse caso, uma pessoa rica e negra nunca estará no mesmo grupo de uma pessoa pobre e branca, mesmo se essas duas pessoas tiverem uma forte relação de amizade.

O exemplo descrito acima pode ser ilustrado conforme apresentado na Figura 34, onde as linhas mais grossas representam relacionamentos fortes e as linhas mais finas representam os relacionamentos mais fracos.

Na Figura 34(a) os grupos são identificados através do algoritmo de agrupamento por Árvore Geradora Mínima, utilizando as informações de perfil e dos relacionamentos dos pesquisadores. Observe que pode ocorrer de pessoas com relacionamentos fracos pertencerem ao mesmo grupo, pois elas possuem perfis muito semelhantes.

Por outro lado, na Figura 34(b) os grupos foram identificados pelo algoritmo de agrupamento por Fluxo Máximo, utilizando apenas informações sobre os relacionamentos. Nesse caso, apenas pesquisadores com bom fluxo de informação entre si pertencem ao mesmo grupo.



(a) Agrupamento por Árvore Geradora Mínima

(b) Agrupamento por Fluxo Máximo

Figura 34 - Exemplo gráfico do resultado de agrupamento por AGM e por Fluxo Máximo.

Os algoritmos de agrupamento podem considerar vários tipos de relações no contexto das redes sociais, como a amizade, as relações de parentesco e co-autoria. Para resolver esse tipo de problema, é importante adotar uma métrica de similaridade que considere as relações entre os objetos, para que os objetos fortemente relacionados sejam mantidos no mesmo grupo, mesmo que seus perfis não sejam muito semelhantes.

Em estudos anteriores (Ströele, Oliveira, Zimbrao and Souza, 2009; Ströele, Silva, Oliveira, Souza and Zimbrao, 2009; Ströele, Silva, Souza, Mello, Souza, Zimbrao and Oliveira, 2011) foi utilizado o algoritmo baseado na árvore geradora mínima. No entanto, como esse algoritmo utiliza a árvore geradora mínima do grafo que representa a Rede Social, uma grande parte dos relacionamentos é excluída. Assim, apenas as informações das relações mais fortes são utilizadas na fase de agrupamento e, conseqüentemente, muitos relacionamentos que poderiam influenciar na formação dos grupos não são considerados.

Existem vários problemas nos quais os atributos de perfil possuem a influência máxima no critério de agrupamentos. Entretanto, existem problemas nos quais esses atributos não possuem grande influência na formação dos grupos, como é o caso das redes sociais científicas. Na análise das redes sociais o fator de maior interesse é compreender como os relacionamentos se formaram e como as pessoas interagem através desses relacionamentos. Assim, os atributos dos perfis de cada pessoa da rede social devem ser analisados, mas deve ser feito em uma segunda etapa.

Além disso, foi observado que no algoritmo de agrupamento por árvore geradora mínima os relacionamentos não eram o critério com maior influência na formação dos

grupos, já que os atributos de perfil têm uma grande influência nessa formação. Assim, os pesquisadores que tinham relacionamentos fortes e perfis diferentes foram atribuídos a grupos distintos. Esse foi o fator mais importante na decisão de se optar pelo desenvolvimento do algoritmo de agrupamento por fluxo máximo para o estudo das redes sociais com relacionamentos científicos, pois nessas redes o relacionamento entre os pesquisadores é o critério mais importante para a construção dos grupos.

5.4 Análise das Funções de Penalização

O critério adotado para avaliar a melhor função de penalização com base no ano foi a análise do fluxo interno dos grupos formados pelo algoritmo de mineração de dados. O algoritmo de agrupamento por fluxo máximo foi aplicado em três conjuntos de dados formados por cada uma das três funções de penalização definidas no capítulo 4. O objetivo dessa análise é identificar qual função proporciona o maior fluxo interno nos grupos.

Tabela 4 - Variação dos pesos por função de penalização

Ano	Sigmóide	Potência	Exponencial
2011	1,00752738	1,00000000	1,71828183
2010	1,00550373	0,70000000	0,64872127
2009	1,00183743	0,49000000	0,39561243
2008	0,99522597	0,34300000	0,28402542
2007	0,98340301	0,24010000	0,22140276
2006	0,96257413	0,16807000	0,18136041
2005	0,92682730	0,11764900	0,15356499
2004	0,86814894	0,08235430	0,13314845
2003	0,77852478	0,05764801	0,11751907
2002	0,65565631	0,04035361	0,10517092
2001	0,51000000	0,02824752	0,09516944

Como as funções de penalização reduzem o peso do relacionamento de formas diferentes, o cálculo do fluxo interno para essa análise foi feito desconsiderando o peso do ano. A Tabela 4 apresenta a variação do peso no ano para cada função de penalização. É fácil ver, com o auxílio dessa tabela, que cada função de penalização aplica um fator de redução diferente no peso dos relacionamentos. Assim, caso essas funções fossem

utilizadas no cálculo do fluxo interno nos grupos, a análise de qual função produz o melhor fluxo interno não seria real.

A análise do fluxo interno seguiu os seguintes passos: modelagem da rede social científica multi-relacional para uma função de penalização específica; execução do algoritmo de agrupamento por fluxo máximo; cálculo do fluxo interno para o conjunto de grupos formados considerando o peso do ano igual a 1; e armazenamento do fluxo interno encontrado para a função de penalização utilizada no processo de modelagem.

Esse processo foi executado individualmente para cada uma das três funções e foi construído um gráfico para analisar os fluxos internos dos grupos. Esse gráfico está ilustrado na Figura 35 e representa o fluxo interno de cada função de penalização dado um número de grupos. Analisando o gráfico é fácil ver que a função exponencial apresenta um fluxo interno melhor que as outras duas funções.

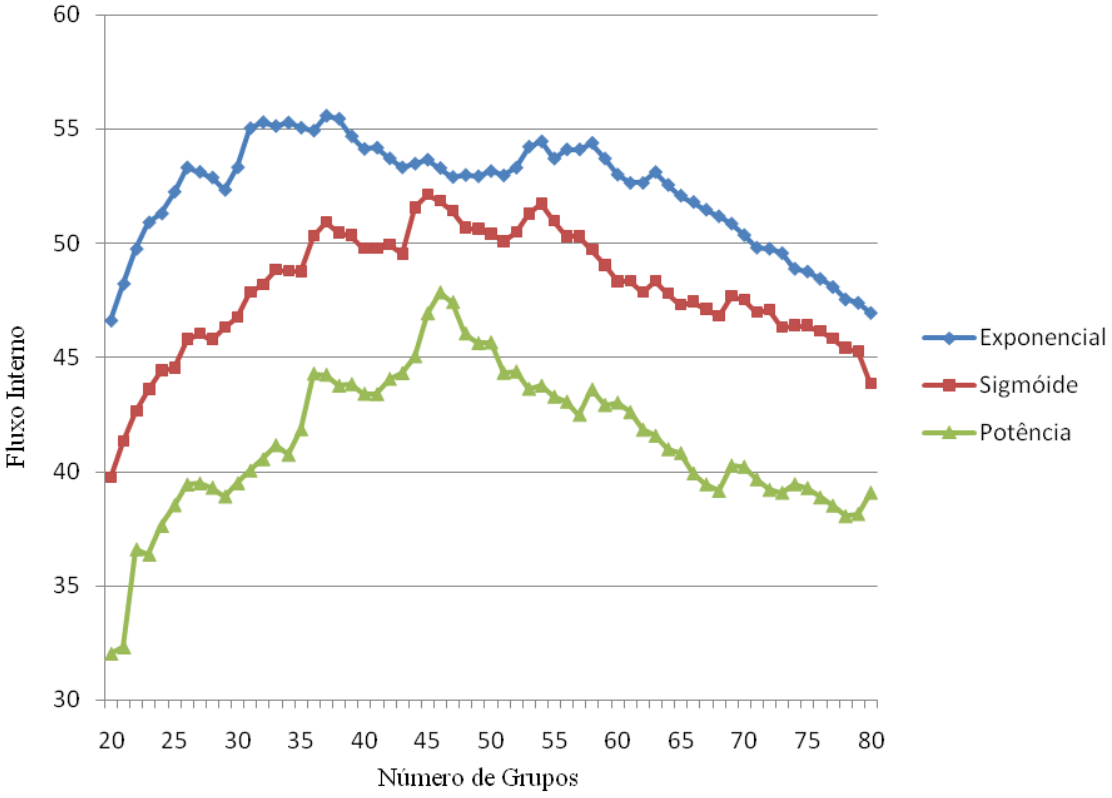


Figura 35 - Variação do fluxo interno para cada função de penalização

O fluxo interno máximo atingido por cada função de penalização está representado no gráfico de barras na Figura 36. Analisando esse gráfico pode ser observado que a diferença entre o fluxo interno máximo obtido pela função exponencial e pela função sigmóide é menor que a diferença do fluxo máximo da função sigmóide e da função potência.

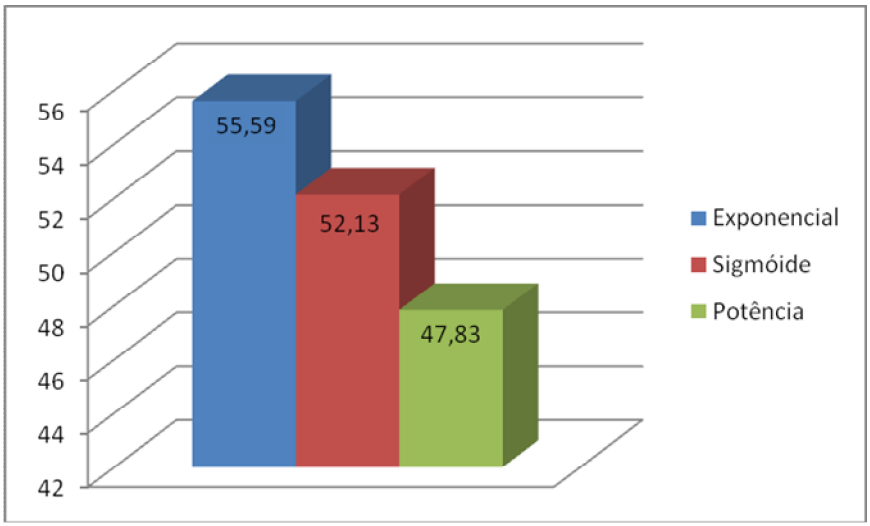


Figura 36 - Fluxo interno máximo obtido para cada função de penalização

Com essas análises fica claro que a função exponencial produz resultados melhores que as outras duas funções. Assim, o estudo de caso deste trabalho foi realizado modelando a rede social científica com a função de penalização exponencial.

5.5 Definição do Número de Grupos

Uma das maiores dificuldades dos problemas de agrupamento não hierárquicos é a definição do número de grupos ideal para o problema que se deseja solucionar, ou seja, descobrir em quantas classes os dados estão distribuídos da melhor forma. Como o número de grupos do problema solucionado neste trabalho é desconhecido foi necessário adotar uma medida para definir o número ideal de agrupamentos.

A análise de agrupamento tem por objetivo identificar grupos homogêneos de forma que a soma das diferenças dentro dos grupos (intragrupos) seja minimizada e a

soma das diferenças entre grupos (intergrupos) seja maximizada (Aldenderfer and Blashfield, 1984). A validação dos grupos é uma avaliação de qual conjunto de grupos tem a melhor estrutura de agrupamento.

Existem várias técnicas que podem ser utilizadas para auxiliar na definição desse número, dentre elas: índice PBM, distância intragrupos, distância intergrupos, dentre outros (Bezdek and Pal, 1998; Pakhira, Bandyopadhyay and Maulik, 2004; Xie and Beni, 1991).

A soma das diferenças intergrupos é dada pela soma das distâncias dos medóides de cada grupo. Essa medida parte do princípio que quanto mais distantes estão os medóides de cada grupo, melhor é o agrupamento. No entanto, essa medida pode não dar bons resultados, pois pode ocorrer de dois medóides estarem distantes e os elementos dos grupos não estarem bem separados, conforme representado hipoteticamente na Figura 37, na qual os medóides são representados pelos dois objetos com cores mais suaves e contornados por linhas grossas.

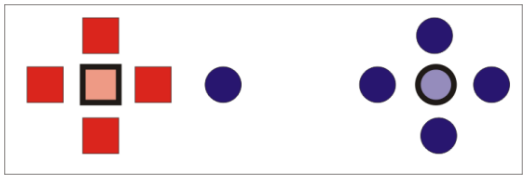


Figura 37 - Distância intergrupos e intragrupos.

A soma das diferenças intragrupos é uma boa medida para avaliar a homogeneidade dos grupos formados. Essa medida avalia a posição dos objetos no espaço de variáveis dentro dos seus respectivos grupos. Essa soma das distâncias internas irá indicar se os objetos de um mesmo grupo estão próximos ao medóide. Com esse tipo de avaliação é mais fácil identificar possíveis erros como o demonstrado na Figura 37.

A soma das diferenças intragrupos irá indicar a distância de todos os elementos do grupo ao medóide. Essa soma é definida por

$$IntraGrupo = \sum_{i=0}^{k} \sum_{j=0}^{m} \left\| x_j - \overline{x}_i \right\|. \qquad (18)$$

Onde *k* é o número de grupos, *m* é o número de elementos do grupo *i*, x_j é um elemento do grupo *i* e $\overline{x}_i$ é o medóide do grupo *i*.

O valor de *k* que produzir a menor soma das distâncias intragrupos, representada na equação (18), indica o número de grupos ideal para o problema que estiver sendo solucionado.

Outra medida utilizada para avaliar o número ideal de grupos é o índice PBM (Pakhira, Bandyopadhyay and Maulik, 2004). Esse índice é definido como segue:

$$PBM(k) = \left(\frac{1}{k} \times \frac{E_1}{E_k} \times D_k \right)^2 \tag{19}$$

onde *k* é o número de grupos e

$$E_k = \sum_{i=1}^{k} E_i \tag{20}$$

$$E_i = \sum_{t=1}^{n} u_{ti} d(x_t, \bar{x}_i), \text{ tal que} \tag{21}$$

$$D_k = \max_{i,j=1}^{k} d(\bar{x}_i, \bar{x}_j) \tag{22}$$

Tem-se que ***n*** é o número de pontos no conjunto de dados, $U(X) = [u_{ti}] n \times k$ é a matriz que indica o grupo de cada elemento e $\bar{x}_i$ é o medóide do *i*-ésimo grupo. O objetivo é maximizar o índice PBM para obter o melhor número de grupos, ou seja, o valor máximo desse índice indica o melhor particionamento (Pakhira, Bandyopadhyay and Maulik, 2004).

Analisando o índice PBM detalhadamente, pode ser observado que ele é composto por três fatores: $1/k$, E_1/E_K e D_K. O primeiro fator irá reduzir o valor do índice à medida que o número de grupos aumenta.

O denominador do segundo fator é a soma dos desvios da posição de cada objeto no espaço de variáveis ao medóide do seu respectivo grupo. E o numerador E_1 é constante, sendo a soma dos desvios para todos os objetos alocados em uma única partição. Assim, E_1/E_K é diretamente proporcional à homogeneidade dos grupos formados. Conseqüentemente, quanto menor E_K, maior a homogeneidade dos grupos e maior é o valor do índice.

Por último, o fator D_K é a distância máxima entre o medóide de dois dos k grupos formados. Quanto maior a distância entre os grupos, maior é o índice de qualidade.

A Figura 38 ilustra a variação do índice PBM pelo número de grupos para o problema proposto. Como mencionado anteriormente, o valor mais elevado para esse índice indica o número ideal de grupos. Conseqüentemente, com base na Figura 38, o valor ideal para o número de grupos é $k = 31$, já que esse é o valor que proporciona um particionamento ideal.

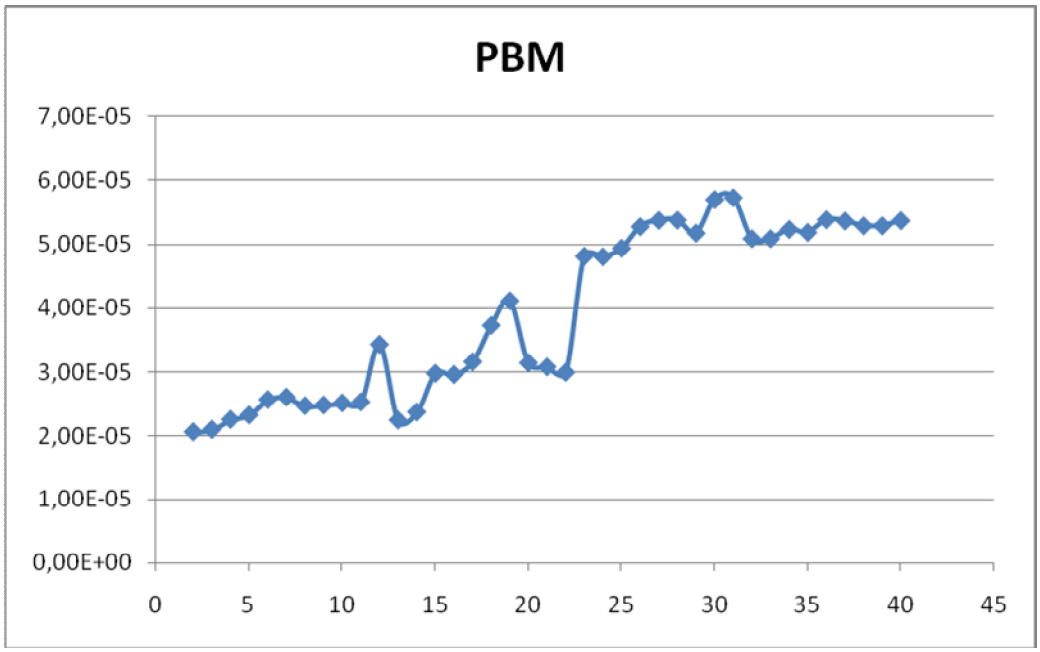


Figura 38 - Variação do Índice PBM.

Nos experimentos realizados neste trabalho foi observado que a estrutura de grupos sugerida pelo índice PBM nem sempre possuía o melhor fluxo interno. Por isso foi adotada uma nova metodologia para definir o número de grupos para o problema de agrupamento em redes sociais científicas multi-relacionais, que é a maximização do fluxo interno dos grupos.

A distância intragrupo é dada pela soma das distâncias de cada elemento do grupo ao medóide, conforme a equação (18). Nos problemas de análise de redes sociais a distância de cada elemento ao medóide deve ser calculada com base nos pesos dos

relacionamentos entre os dois elementos sociais. A proposta deste trabalho é utilizar o fluxo máximo entre os pesquisadores para definir uma medida de avaliação dos agrupamentos. Essa estratégia é semelhante à análise da distância intragrupo, porém utiliza a informação do fluxo de informações dentro de cada grupo.

Dessa maneira, a medida de avaliação dos grupos será feita da seguinte forma

$$FluxoIntraGrupo = \sum_{i=0}^{k} \sum_{j=0}^{m} MaxFlow(x_j, \overline{x}_i). \quad (23)$$

Onde $MaxFlow(x_j, \overline{x}_i)$ é o fluxo máximo de informação entre os pesquisadores x_j e o medóide $\overline{x}_i$. O algoritmo para o cálculo do fluxo máximo entre dois pesquisadores foi apresentado neste capítulo na seção anterior.

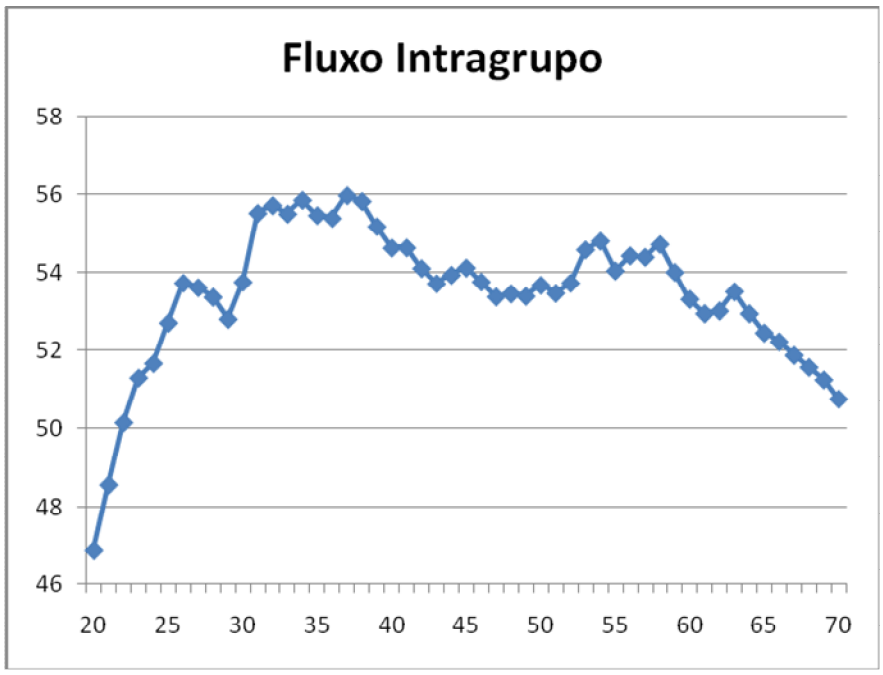


Figura 39 - Variação do Fluxo Intragrupo.

A Figura 39 ilustra a variação da medida proposta na equação (23) pelo número de grupos para o estudo de caso deste trabalho. Pode ser observado que os melhores fluxos internos estão para $31 \leq k \leq 38$, ou seja, no estudo de caso deste trabalho o número de grupos ideal está no intervalo de 31 a 38 grupos. De maneira mais precisa, o número

de grupos ideal para este problema é $k = 37$, já que o maior fluxo interno foi obtido para esse número de grupos.

Analisando os resultados produzidos pelo índice PBM e pelo Fluxo Intragrupo pode ser observado que embora $k = 31$ tenha um bom fluxo interno, esse número de grupos não produz o máximo fluxo conforme apresentado na Figura 39.

No estudo de caso desenvolvido neste trabalho a maximização do fluxo interno será adotada como a medida para definir o número de grupos ideal. Como um dos objetivos deste trabalho é encontrar comunidades de pesquisa que possuam a melhor comunicação interna é natural que esse critério seja utilizado para definir o número de grupos a ser adotado. Assim, será utilizado $k = 37$ na análise dos agrupamentos do estudo de caso que será apresentado no próximo capítulo.

Capítulo 6 – Estudo de Caso: Comunidades Científicas

Neste capítulo será apresentada a primeira etapa do estudo de caso deste trabalho, que é a descoberta das comunidades científicas. Os resultados foram gerados pelo algoritmo de mineração de dados por fluxo máximo.

Além disso, serão feitas várias análises sobre os resultados obtidos, partindo de uma visualização mais global da rede social científica até uma visão mais detalhada da mesma, na qual cada pesquisador pode ser analisado individualmente.

6.1 Introdução

A primeira etapa do estudo de caso deste trabalho visa identificar as comunidades de pesquisa inter e intra-universidades brasileiras através de uma rede social científica multi-relacional ponderada formada por pesquisadores e quatro tipos de relacionamentos diferentes que existem nas instituições científicas. O conjunto de dados e o processo de modelagem dessa rede social foram descritos detalhadamente no Capítulo 4.

A segunda etapa do estudo de caso tem como objetivo sugerir novos relacionamentos para os pesquisadores. Os estudos sobre previsão de relacionamentos e as análises sobre a sugestão de relacionamentos estão nos capítulos 7 e 8, respectivamente.

Com base nas pesquisas anteriores, essa primeira etapa do estudo de caso utilizou o algoritmo de agrupamento por fluxo máximo. Conforme descrito no capítulo anterior, esse algoritmo avalia apenas as informações contidas nos relacionamentos durante o processo de agrupamento na Rede Social Científica.

Todas as análises da rede social científica multi-relacional foram feitas através da ferramenta de visualização de redes sociais desenvolvida também neste trabalho. Essa ferramenta permite a análise da rede sobre diferentes níveis de visualização, permitindo um detalhamento maior ou menor da rede segundo as análises que o usuário deseja realizar. Além disso, é possível fazer filtros por pesquisadores específicos e fazer a análise temporal da rede social. Os detalhes sobre os níveis de visualização e todas as

funcionalidades da ferramenta de visualização desenvolvida neste trabalho estão no Capítulo 9.

O desenvolvimento do módulo de agrupamento, de sugestão de relacionamentos e de visualização da rede social científica multi-relacional tem como intuito facilitar a análise das redes sociais. Analisando a rede social científica brasileira pretende-se aprender como ocorre a troca de informações entre os pesquisadores, entre as comunidades de pesquisas e entre as instituições de ensino.

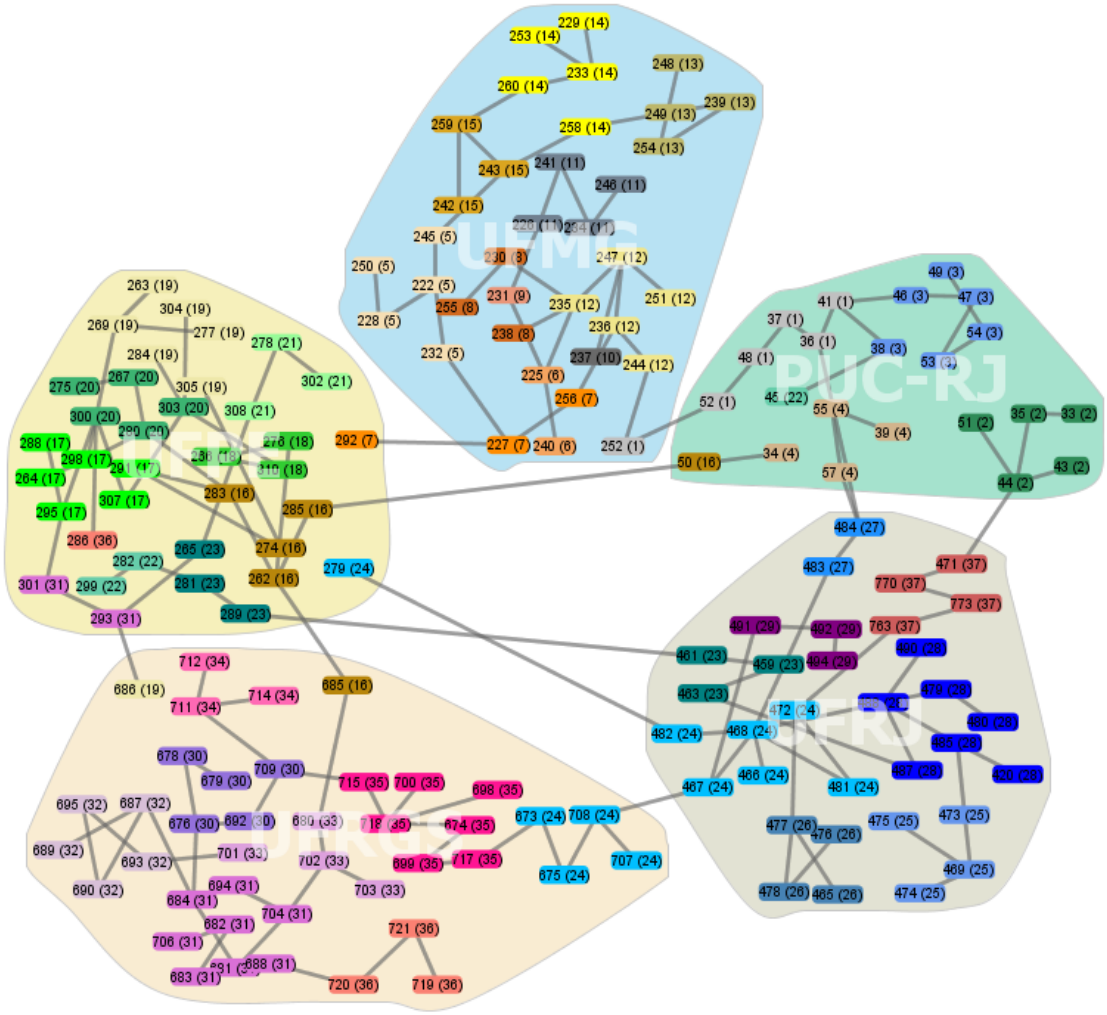


Figura 40 - Relacionamentos Internos e Externos (Árvore Geradora Mínima)

A Figura 40 mostra os resultados obtidos pela técnica de mineração de dados. As regiões maiores mostram as instituições brasileiras e os retângulos menores mostram os

grupos dentro de uma organização. Cada retângulo possui dois números, o primeiro identifica de maneira única o pesquisador e o segundo número, que está entre parênteses, identifica um grupo gerado pelo método de detecção de grupo.

Para facilitar a visualização da rede social, na Figura 40 as arestas representam apenas os relacionamentos mais fortes, ou seja, nessa figura é exibida a árvore geradora mínima do grafo social. Além disso, essas arestas têm informações sobre os quatro tipos de relacionamentos, como mostrado no Capítulo 4.

6.2 Análise dos Grupos

A Análise dos Grupos permite que seja avaliado como ocorre a comunicação entre os pesquisadores. Além disso, pode ser estudada como ocorre a troca de conhecimento entre os grupos e entre as instituições de ensino. Nessa etapa do trabalho serão analisadas as comunidades científicas encontradas pelo algoritmo de agrupamento por fluxo máximo.

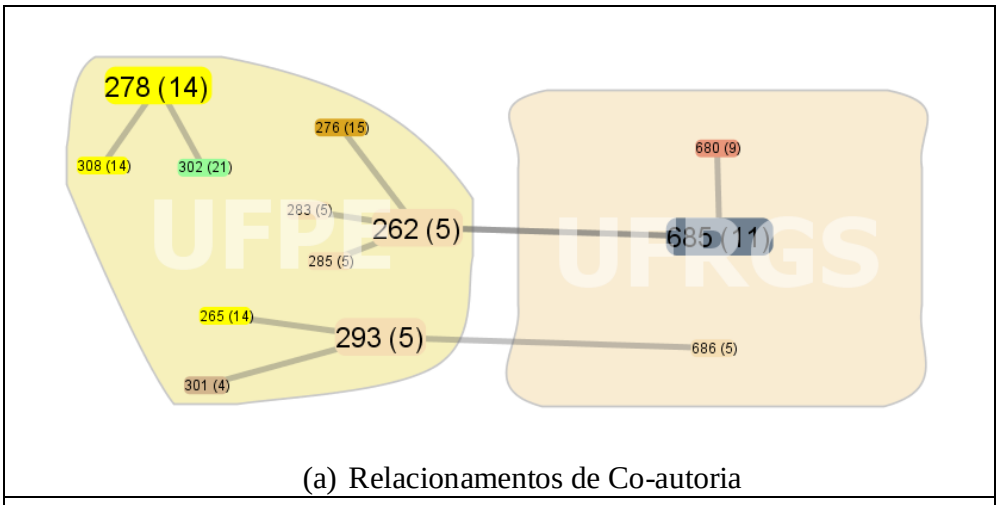
Nos primeiros estudos foram considerados apenas os relacionamentos de co-autoria (Ströele, Silva, Oliveira, Souza and Zimbrao, 2009; Ströele, Silva, Souza, Mello, Souza, Zimbrao and Oliveira, 2011). Comparando os resultados obtidos nos estudos anteriores com os obtidos pela análise da rede social multi-relacional, pode ser observado que, com a adição dos novos tipos de relacionamentos na rede social científica, a estrutura dos grupos mudou.

Em alguns casos, pesquisadores que estavam em grupos diferentes se mudaram para o mesmo grupo. Isso ocorreu porque esses pesquisadores possuem múltiplos tipos de relacionamentos uns com os outros. Conseqüentemente, as conexões desses pesquisadores são mais forte que as conexões dos pesquisadores que possuem apenas relacionamentos de co-autoria.

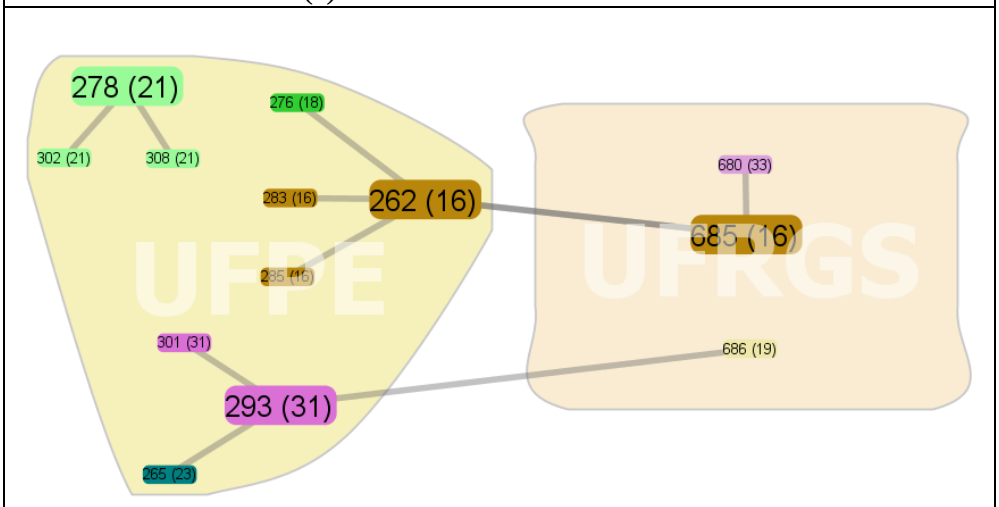
Na Figura 41 está ilustrada uma pequena parte da rede social científica utilizada no estudo de caso deste trabalho. Essa parte da rede social contempla o cenário descrito no parágrafo anterior. Na Figura 41a, estão representados os grupos gerados usando apenas os relacionamentos de co-autoria na etapa de modelagem da rede social. Por outro

lado, na Figura 41b, estão os grupos formados com a inclusão dos novos tipos de relacionamentos descritos no Capítulo 4.

Analisando a rede social multi-relacional, pode ser observado que o pesquisador *302* mudou-se para o grupo do pesquisador *278*. Por outro lado, o pesquisador *685* migrou para o grupo do pesquisador *262*, formando assim um grupo interinstitucional. Outras mudanças podem ser observadas analisando apenas essa pequena parte da rede social. Assim, considerando a rede social como um todo, pode-se concluir que através da construção de uma rede multi-relacional, chega-se muito mais perto da realidade das instituições brasileiras.



(a) Relacionamentos de Co-autoria



(b) Múltiplos Relacionamentos

Figura 41 - Análise da Estrutura dos Grupos.

Como mencionado no capítulo anterior, a análise dessa rede social científica foi realizada em trabalhos anteriores utilizando o algoritmo de agrupamento por árvore geradora mínima (Ströele, Oliveira, Zimbrao and Souza, 2009; Ströele, Silva, Oliveira, Souza and Zimbrao, 2009; Ströele, Silva, Souza, Mello, Souza, Zimbrao and Oliveira, 2011). Analisando os resultados obtidos com o novo algoritmo de agrupamento por fluxo máximo foi observado que o número de grupos unitários reduziu, ou seja, o número de grupos formados por apenas um pesquisador reduziu consideravelmente. Utilizando o algoritmo de agrupamento por árvore geradora mínima foram encontrados 10 grupos unitários, enquanto que utilizando o algoritmo de agrupamento por fluxo máximo foram encontrados apenas 2.

A redução do número de grupos unitários também foi um fator que contribuiu para a escolha do algoritmo de agrupamento por fluxo máximo. O objetivo desse módulo é identificar grupos de pesquisadores que trabalham em conjunto, portanto, quanto menor o número de grupos unitários melhor será a informação de quais pesquisadores compartilham de interesses comuns.

Para analisar a distribuição dos grupos entre as universidades a rede social científica foi analisada em um nível de visualização mais abstrato no qual é possível visualizar apenas os grupos obtidos pelo método de mineração de dados. Essa facilidade de visualização em níveis é uma das vantagens da ferramenta desenvolvida neste trabalho.

A Figura 42 ilustra o nível de visualização dos grupos. Caso o usuário deseje visualizar as medidas de cada agrupamento basta ele acessar o componente de informações da ferramenta de visualização que contém a síntese dos grupos encontrados. O acesso e o conteúdo desse componente estão descritos em detalhes no capítulo de visualização.

Na Figura 42 cada retângulo representa um grupo dentro da universidade. Cada grupo é identificado pelo nome da universidade ao qual ele pertence e pelo número que o representa. Por exemplo, o grupo “PUC-RJ;1” é o grupo de número 1 na PUC-RJ, já o grupo “UFMG;1” é o mesmo grupo de número 1, porém na UFMG. Os retângulos que representam o mesmo grupo possuem cores iguais para facilitar a identificação dos grupos inter institucionais.

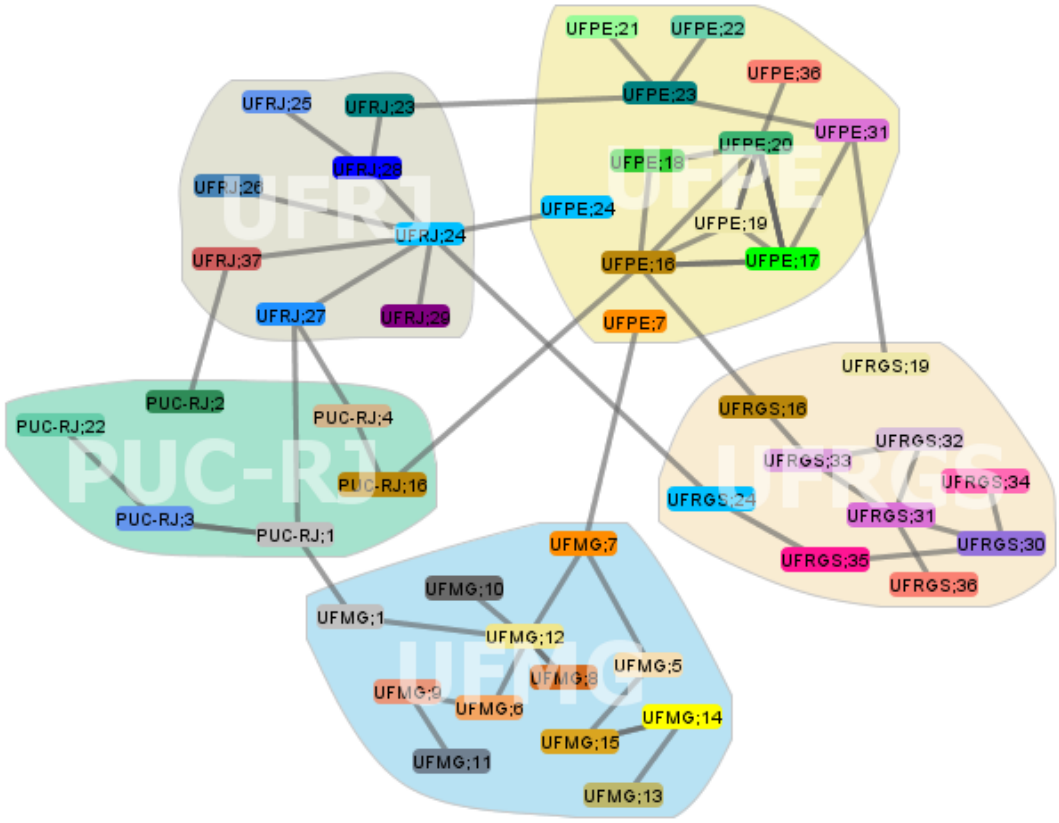


Figura 42 - Visualização abstrata para análise da distribuição dos grupos

A Tabela 5, construída com o auxílio da Figura 42, mostra de maneira consolidada a distribuição dos grupos intra e inter instituições. A diagonal principal representa o número de grupos que existem apenas na universidade específica. Já as outras células indicam o número de grupos que são comuns às duas universidades.

Embora a maioria dos grupos seja formada por pesquisadores de apenas uma universidade, alguns deles possuem pesquisadores de universidades diferentes. Nestes casos, pode-se dizer que há uma grande troca de informações entre as instituições através desses grupos.

Analisando a Tabela 5 pode ser visto que o par (UFRJ, UFPE) representa as únicas universidades que possuem mais de um grupo em comum. Todos os outros pares possuem no máximo um grupo em comum.

Ainda com base na Tabela 5 pode-se dizer que não há grandes interesses em se desenvolver pesquisas em conjunto entre os pares de instituições (UFRJ, PUC-RJ) e

(UFRJ, UFMG), já que esses pares não possuem nenhum grupo em comum. Assim, o fluxo de informações entre essas instituições é menor que entre as que possuem grupos em comum.

Tabela 5 - Distribuição dos Grupos

	UFRJ	PUC-RJ	UFMG	UFRGS	UFPE
UFRJ	6	0	0	1	2
PUC-RJ	0	4	1	0	1
UFMG	0	1	10	0	1
UFRGS	1	0	0	8	1
UFPE	2	1	1	1	8
Grupos Comuns	2	2	2	2	4

A linha de Grupos Comuns da Tabela 5 indica o número de grupos que cada universidade possui que é formado por pesquisadores de universidades diferentes. Analisando essa linha da tabela observa-se que a UFPE é a universidade que possui o maior número de grupos inter institucionais.

Com base na Figura 40 e na Figura 42, pode ser observado que existem pesquisadores que são responsáveis por estabelecer uma forte ligação com pesquisadores de instituições externas, sendo por isso que existem grupos interinstitucionais. Com a ajuda desses pesquisadores, as informações podem ser propagadas através da rede social com maior facilidade, ou seja, a informação de uma instituição é transferida para outra mais facilmente através deles.

É importante dizer que existem muitos outros relacionamentos entre as instituições do que os mostrados na Figura 40, na qual está sendo exibida apenas a árvore geradora mínima. Portanto, embora esses pesquisadores sejam muito importantes no processo de troca de informações, o fluxo de conhecimento interinstitucional não depende apenas deles.

Além da análise dos grupos interinstitucionais também foram analisadas as áreas de interesse de cada pesquisador dos grupos. A distribuição das áreas pelos grupos está na Tabela 6. É esperado que os pesquisadores de um mesmo grupo desenvolvam seus trabalhos na mesma área ou em áreas afins.

Conforme dito no capítulo 4, na seção de análise dos atributos de perfil, nem todas as áreas dos pesquisadores foram analisadas. Além disso, os grupos unitários e os grupos que possuem pesquisadores sem experiências nessas áreas também não foram analisados.

A Tabela 6 mostra que os grupos realmente possuem áreas de atuação semelhantes. Entretanto, existem alguns grupos que possuem pesquisadores de áreas diferentes. Esses grupos são chamados de *grupos interdisciplinares* e são muito importantes para o desenvolvimento dos estudos científicos, pois eles indicam que existem pesquisadores "misturando" soluções de diferentes áreas. Através desses estudos podem surgir idéias bastante inovadoras.

Os grupos 1 e 31 são exemplos de grupos interdisciplinares, pois esses grupos possuem pesquisadores da área de Banco de Dados e também da área de Inteligência Artificial. Pesquisadores que trabalham em conjunto nessas duas áreas têm sido bastante comum, já que é possível aplicar as soluções encontradas na área de Inteligência Artificial para solucionar os problemas da área de Banco de Dados e vice-versa.

Tabela 6 - Distribuição das áreas de atuação por grupo

	1	2	3	4	5	6	7	13	14	15	16	17	18	19	20	22	23	24	26	28	29	30	31	33	35	36	37
Análise e Complexidade de Algoritmos		x		x													x										
Arquitetura de Sistemas de Computação									x	x	x							x	x						x	x	
Banco de Dados	x				x								x					x					x	x			x
Engenharia de Software		x	x						x		x	x		x	x				x				x	x			
Hardware																		x							x		
Inteligência Artificial	x													x									x				
Linguagem de Programação			x					x				x			x												
Mineração de Dados																x											
Modelos Analíticos e de Simulação																		x			x		x				
Otimização																				x							
Processamento Gráfico							x															x					
Redes de Computadores						x																					
Sistema de Informação		x			x						x	x	x			x			x				x	x			x
Teleinformática			x																		x					x	

6.3 *Análise dos Relacionamentos*

Os resultados obtidos nesse estudo de caso permitiram que a rede social científica fosse analisada global e localmente. Através de uma perspectiva global, é possível ver todos os relacionamentos entre as instituições de ensino, grupos e pesquisadores, analisando o fluxo de conhecimento entre cada um desses níveis.

Na análise local é possível fazer buscas por pesquisadores específicos e analisar como eles colaboram uns com os outros. Como dito anteriormente, os níveis de visualização e as funcionalidades da ferramenta de visualização desenvolvida para a análise da rede social científica multi-relacional estão descritos em detalhes no Capítulo 9.

Inicialmente foi estudada a força com a qual as universidades estabelecem os seus relacionamentos. Os relacionamentos foram classificados em internos ou externos. Os relacionamentos entre pesquisadores que fazem parte da mesma instituição de ensino são chamados de *relacionamentos internos*, já os relacionamentos formados por pesquisadores de instituições diferentes são chamados de *relacionamentos externos*.

Os relacionamentos entre cada par de universidades foram examinados através de uma matriz simétrica, mostrada na Tabela 7, na qual o par *ij* representa o número total de relacionamentos entre as universidades *'i'* e *'j'*.

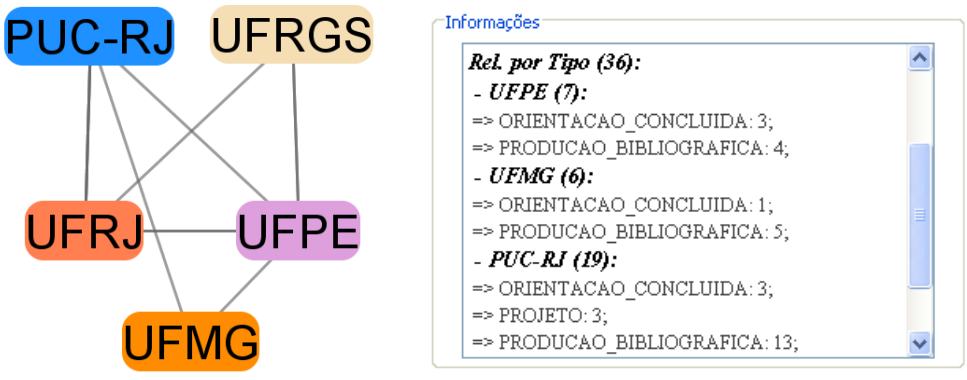


Figura 43 - Análise dos relacionamentos interinstitucionais

Para construir essa tabela, o conjunto de dados foi cuidadosamente analisado para identificar todos os relacionamentos externos de cada universidade da rede social

científica e para totalizar o número de pesquisadores de cada uma delas. As universidades UFRJ, PUC-RJ, UMFG, UFRGS e UFPE possuem, respectivamente, 33, 22, 36, 41 e 37 pesquisadores. Todas as informações dessa tabela estão disponíveis na ferramenta de visualização, conforme ilustra a Figura 43. O painel de informações apresenta os totais de relacionamentos de cada universidade, no caso desta figura as informações são da UFRJ.

Tabela 7 - Totais de relacionamentos inter institucionais

	UFRJ	PUC-RJ	UFMG	UFRGS	UFPE
UFRJ	–	19	6	4	7
PUC-RJ	19	–	24	3	23
UFMG	6	24	–	11	8
UFRGS	4	3	11	–	10
UFPE	7	23	8	10	–
TOTAL	36	69	49	28	48
TOTAL (RP)	1.09	3.14	1.36	0.68	1.30

O total de relacionamentos externos foi analisado relativamente ao número de pesquisadores (Relacionamentos por Pesquisador – RP), pois o número de pesquisadores de cada instituição é diferente. É mais provável que uma instituição com muitos pesquisadores tenha mais relacionamentos externos do que uma instituição com poucos pesquisadores.

Analisando a última linha da Tabela 7 pode ser observado que a PUC-RJ é a instituição com o maior número de relacionamentos externos por pesquisador, ou seja, dentre as instituições analisadas nesse estudo de caso, a PUC-RJ é a que realiza o maior número de trabalhos interinstitucionais. Por outro lado, a UFRGS é a universidade com a menor relação de pesquisadores com ligações externas.

A Tabela 8 também é uma matriz simétrica na qual o par *ij* representa o total de relacionamentos fortes entre as instituições *'i'* e *'j'*. Os relacionamentos fortes são aqueles que formam a árvore geradora mínima da rede social científica. Cada relacionamento forte indica uma forte cooperação entre os pesquisadores. A árvore geradora mínima da rede social deste trabalho pode ser visualizada na Figura 40.

Nessa tabela foi analisado o total de relacionamentos externos fortes por pesquisador (RFP). É possível ver que a PUC-RJ e a UFRJ são as instituições mais

fortemente relacionadas a outras instituições. Por outro lado, a UFMG e a UFRGS são as menos fortemente ligadas a outras universidades.

Tabela 8 - Totais de relacionamentos fortes

	UFRJ	PUC-RJ	UFMG	UFRGS	UFPE
UFRJ	–	3	0	1	2
PUC-RJ	3	–	1	0	1
UFMG	0	1	–	0	1
UFRGS	1	0	0	–	2
UFPE	2	1	1	2	–
TOTAL	6	5	2	3	6
TOTAL (RFP)	0.18	0.22	0.05	0.07	0.16

Comparando o resultado das duas tabelas apresentadas anteriormente, pode ser observado que existem pesquisadores que possuem um perfil de trabalhar mais estreitamente com pesquisadores de outras instituições. Os relacionamentos estão concentrados em poucos pesquisadores que acabam se relacionando com freqüência com a outra universidade, criando um vínculo forte com a mesma. Assim, os relacionamentos fortes entre as instituições dependem de um pequeno grupo de pesquisadores, como é o caso da PUC-RJ e da UFRJ.

No entanto, existem universidades onde os relacionamentos externos são formados por um grande número de pesquisadores. Vários pesquisadores possuem relacionamentos externos, mas nenhum deles em grandes quantidades. Assim, esse tipo de instituição possui muitos pesquisadores com relacionamentos externos fracos. Esse cenário ocorre com a UFMG, que possui um bom número de relacionamentos interinstitucionais (Tabela 7), mas não possui muitos relacionamentos externos fortes (Tabela 8).

Como já era esperado, os relacionamentos intra-institucionais são geralmente mais fortes que os interinstitucionais. Assim, conclui-se que pesquisadores que estão locados na mesma instituição de ensino possuem uma tendência maior de trabalharem juntos do que com pesquisadores de outras instituições.

Os relacionamentos de alguns pesquisadores específicos também foram analisados. A idéia foi identificar alguns centralizadores de conhecimento. Esse estudo

foi feito utilizando três pesquisadores da UFRJ e um pesquisador da PUC-RJ. Os centralizadores de conhecimento são considerados *conectores centrais*: pessoas que, em virtude de suas relações com as pessoas em diferentes organizações ou grupos, servem como chaves de fronteira (passando informações e contexto de uma instituição ou grupo para outro). Esses pesquisadores são facilmente identificados com o auxílio da ferramenta de visualização, mas para selecioná-los foi feita uma lista dos pesquisadores com o maior número de relacionamentos internos e externos.

A partir dessa lista foram selecionados quatro pesquisadores com perfis de relacionamento diferentes e, através do filtro da ferramenta de visualização, foi construída uma rede social científica exibindo apenas os relacionamentos desses pesquisadores, a saber: 773, 485, 468 e 44. Por uma questão de privacidade todas as imagens exibidas neste trabalho não apresentam nem os nomes e nem as fotos dos pesquisadores. Entretanto, a ferramenta permite que essas informações sejam visualizadas.

A Figura 44 mostra a visualização local da rede social produzida pela ferramenta. Os pesquisadores selecionados ficam em destaque para facilitar a análise por parte do usuário. Cada nó da rede social apresenta o número que identifica unicamente um pesquisador. Caso seja solicitado, a ferramenta irá exibir em cada nó a foto do pesquisador, caso a mesma esteja disponível no currículo *lattes* (LATTES, 2008), o nome do pesquisador e o número do grupo ao qual ele pertence.

Na Figura 44 as cores das arestas indicam tipos de relacionamentos diferentes. Pode ser observado que os pesquisadores 468 e 773 possuem um perfil de relacionamento interno, ou seja, eles possuem mais relacionamentos com pesquisadores da mesma instituição que a deles do que com pesquisadores de instituições diferentes. Por ter um perfil de relacionamento interno, esses pesquisadores são definidos como *Conectores Centrais Internos*.

Por outro lado, os pesquisadores externos, que são uma minoria na rede social científica, são aqueles com mais relacionamentos externos do que relacionamentos internos. Também é possível ver na Figura 44 que o pesquisador 485 possui mais conexões externas do que internas, sendo, portanto, considerado um *Conector Central Externo*.

Finalmente, há um grupo de pesquisadores que possuem relacionamentos internos e externos na mesma proporção. O pesquisador 44 tem o mesmo nível de colaboração interna e externa, podendo ser definido como um *Conector Central Interno/Externo*.

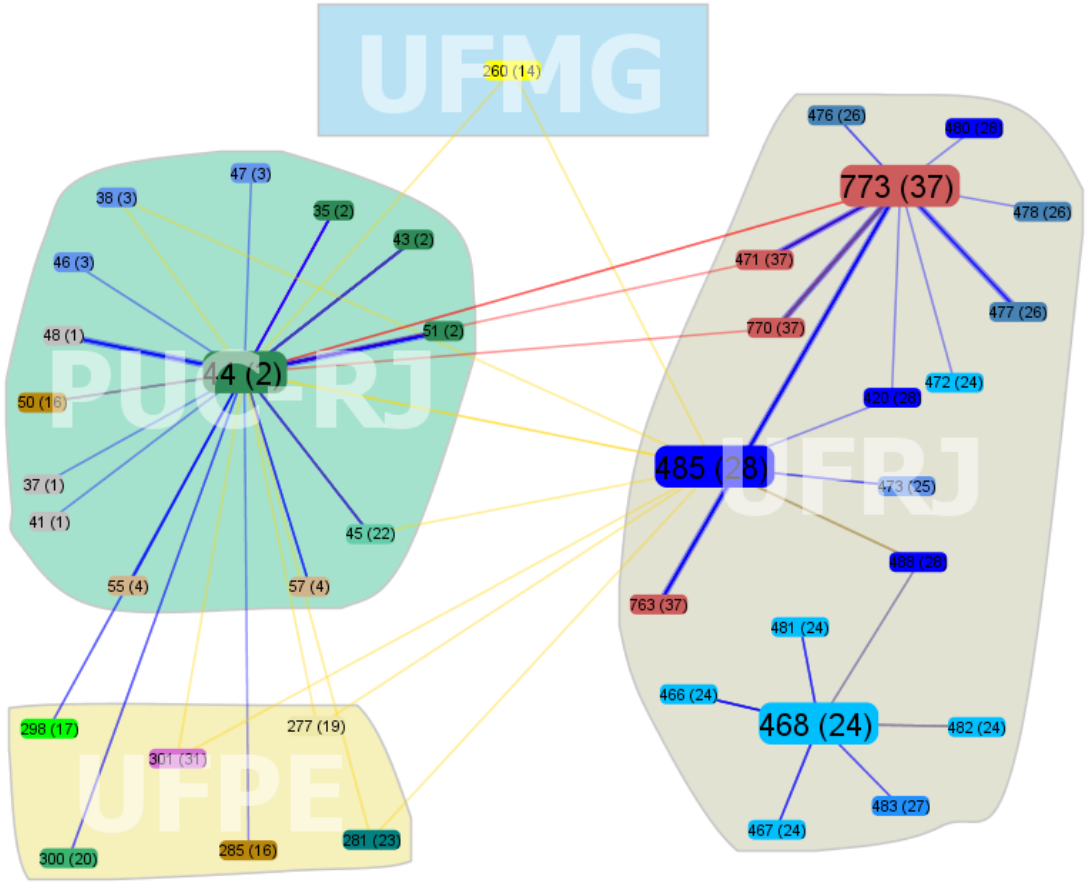


Figura 44 - Visualização local da Rede Social Científica Multi-relacional.

Com a análise local da rede social pode ser observado que surgiram novos fluxos de conhecimento na rede social multi-relacional quando comparado com a rede social homogênea. Concluiu-se que, por vezes, os pesquisadores não possuem o tipo de relacionamento representado pela rede social homogênea, mas possui outros tipos de relacionamentos que são representados na rede social multi-relacional. Essa é uma grande vantagem da análise de redes sociais multi-relacionais.

A fim de analisar as principais diferenças entre redes multi-relacionais e redes homogêneas, foi feita uma rede social apenas com os relacionamentos de co-autoria. Essa

rede social está ilustrada na Figura 45 e possui os mesmos pesquisadores utilizados na análise dos conectores centrais. A ferramenta de visualização possui um componente onde é possível selecionar os tipos de relacionamentos da rede social que devem ser exibidos. Assim, para criar a figura, bastou selecionar para exibição o relacionamento de co-autoria e desmarcar os outros tipos.

Observando a Figura 45, pode ser visto que existem alguns pesquisadores que não se relacionam por relacionamentos de co-autoria, como é o caso de todos os relacionamentos externos dos pesquisadores 773, 485. No entanto, esses pesquisadores têm outros tipos de relacionamentos, como mostrado na Figura 44. A rede social multi-relacional tem muitos fluxos de conhecimento alternativos que não estão representados na rede social homogênea. Conseqüentemente, quando esses dois tipos de redes são comparados, observa-se que a rede social multi-relacional tem um fluxo de conhecimento melhor do que as redes sociais homogêneas.

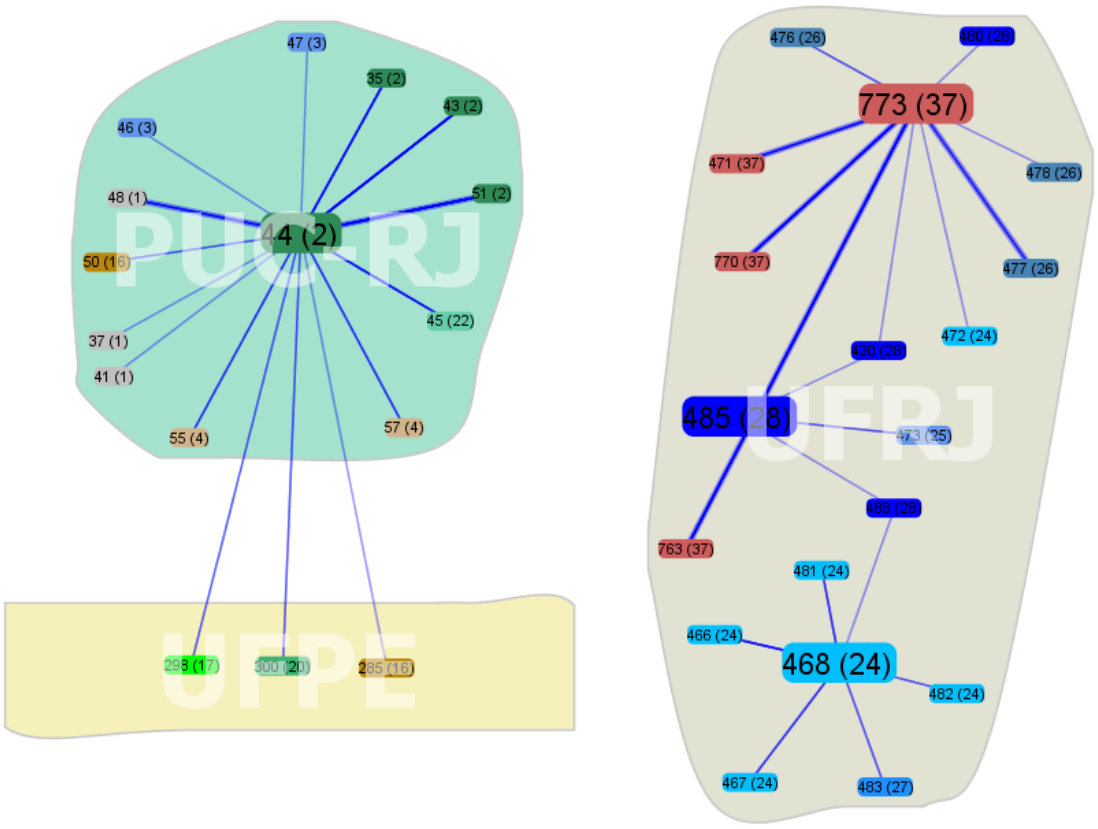


Figura 45 - Visualização local da Rede Social Científica de Co-autoria.

6.4 Validação dos Resultados

Todos os dados utilizados nos experimentos descritos anteriormente são reais. Conforme apresentado nas seções anteriores, os dados representam informações acadêmicas fornecidas pelos próprios pesquisadores. Assim, tem-se a garantia da veracidade dos dados utilizados na construção da rede social científica.

Além da consistência dos dados, também foram validadas as análises feitas no desenvolvimento do módulo de agrupamento. A análise dos relacionamentos e os grupos formados pelo método proposto foram validados utilizando uma avaliação qualitativa. Foram entrevistados – com o auxílio de um questionário – os pesquisadores de uma das universidades, as respostas foram analisadas, e, em seguida, comparadas com os resultados de nossa abordagem.

Nesse formulário foram adicionadas perguntas que permitissem identificar as características da rede social estudada neste trabalho. As principais informações que precisam ser levantadas são: se as áreas de pesquisas são interligadas; como ocorrem os relacionamentos entre os pesquisadores; como ocorrem os relacionamentos entre as instituições de ensino; etc.

Além disso, os pesquisadores também foram questionados sobre suas áreas de interesse; se ele/ela trabalha com pesquisadores de outras áreas; se ele/ela geralmente se relaciona com pesquisadores de outras instituições; e outros tipos de perguntas, a fim de mapear o comportamento científico do pesquisador.

Com o intuito de validar as análises feitas sobre o perfil relacional dos pesquisadores, foram feitas perguntas para verificar com que freqüência os pesquisadores interagem com pesquisadores da mesma instituição e com pesquisadores de instituições diferentes. Assim, é possível verificar se eles possuem um perfil de relacionamento interno, externo ou interno/externo. Através dessas informações foi também foi possível verificar a força dos relacionamentos através dessas informações.

Durante essa avaliação qualitativa, os pesquisadores indicaram as áreas de pesquisas com as quais eles estão mais envolvidos. Assim, foram validadas as análises feitas sobre os grupos interdisciplinares, e, em alguns casos, foi verificado que o relacionamento entre pesquisadores de áreas diferentes é realmente forte. Com isso, esses

pesquisadores fazem parte do mesmo grupo de pesquisa, validando os grupos interdisciplinares identificados pelo método de mineração de dados.

Através dos questionários respondidos, foi possível ver que as áreas de Banco de Dados, Engenharia de Software e Sistema de Informação estão fortemente relacionadas. O mesmo caso ocorre, por exemplo, com a área de Inteligência Artificial e Sistema de Informação.

Ao todo, 34 questionários foram respondidos, o que representa cerca de 20% dos pesquisadores do conjunto de dados deste trabalho. A maioria dos pesquisadores que responderam ao questionário faz parte da COPPE/UFRJ. Entretanto, foram obtidas respostas de pesquisadores de todas as instituições de ensino analisadas neste trabalho. O modelo do formulário utilizado está no anexo A.

Todos os pesquisadores que responderam ao questionário afirmam estar mais diretamente relacionado a pesquisadores da mesma instituição do que com pesquisadores de instituições externas. Cada pesquisador indicou os nomes dos colegas com os quais ele se considera mais estreitamente relacionado profissionalmente, o que permitiu avaliar os relacionamentos fortes e fracos. Os resultados obtidos com o questionário mostraram que tanto os métodos propostos quanto as análises da rede social científica estão muito próximos da realidade.

Como a maioria dos questionários foi respondida por pesquisadores da COPPE/UFRJ, foi possível validar os relacionamentos externos dessa instituição de ensino. Os resultados obtidos com os questionários mostraram que tanto o método proposto quanto as análises dos relacionamentos estão corretas.

Capítulo 7 – Previsão/Sugestão de Relacionamentos

O estudo de previsão de relacionamentos em redes sociais, do inglês *link prediction*, sempre teve grande interesse por parte dos pesquisadores. Saber com antecedência as ações futuras de uma rede social pode auxiliar, por exemplo, a tomada de decisão dos agentes dessa rede social.

No âmbito das redes sociais científicas a sugestão de relacionamentos pode acelerar o processo de comunicação e, conseqüentemente, melhorar a produção científica da instituição de ensino.

Neste capítulo serão apresentados os principais conceitos para o desenvolvimento do módulo de sugestão de relacionamentos. Serão apresentadas algumas técnicas já existentes, as principais métricas utilizadas no estudo desse tipo de problema, bem como uma nova métrica que está sendo proposta neste trabalho.

7.1 Previsão de Relacionamentos

Como pode ser observado nos estudos apresentados anteriormente, a Análise das Redes Sociais é uma área que visa compreender um determinado contexto social representando a rede social por grafos matemáticos. Entretanto, até o presente momento, a grande maioria dos estudos foi realizada em uma Rede Social Estática, ou seja, não são incluídos e nem removidos nós ou relacionamentos no grafo social. Assim, no contexto das redes sociais científicas, um relacionamento existe entre dois pesquisadores se eles tiveram alguma relação científica no passado, de forma que, uma vez estabelecido o vínculo, esse relacionamento nunca mais será removido.

Embora muitos estudos tenham sido desenvolvidos para análise de redes sociais estáticas, não é difícil ver que no mundo real, a maioria das redes sociais é dinâmica. No caso de redes sociais científicas é fácil imaginar como relacionamentos mais antigos podem não existir, pelo fato de um dos pesquisadores estar envolvido em outra área de pesquisa, ou talvez por ele ter saído da universidade e esses dois pesquisadores terem perdido contato. A entrada ou saída de uma pessoa em uma rede social resulta na criação

ou remoção de muitos relacionamentos, formando um ciclo de entradas e saídas, mantendo assim a rede social viva e em constante alteração.

Identificar as alterações nas redes sociais e analisar o comportamento futuro da rede através de informações sobre o seu passado é chamado de Previsão de Relacionamentos (Link Prediction). O problema de Previsão de Relacionamentos foi definido por Liben-Nowell & Kleinberg (2007) da seguinte maneira:

> *Dado uma foto de uma rede social em um instante* ***t****, busca-se prever com precisão as arestas que serão adicionadas à rede durante o intervalo de tempo* ***t*** *até um tempo futuro* ***t'***. (Potgieter, April, Cooke and Osunmakinde, 2009)

Embora o nome e a definição formal do problema de previsão de relacionamentos só mencionem os relacionamentos das redes sociais, essa área abrange a previsão de qualquer alteração na rede, ou seja, prever a entra de uma nova pessoa na rede social também é um problema de previsão.

A previsão de relacionamentos é abordada de duas formas diferentes, porém complementares. A primeira examina as estruturas sociais em desequilíbrio, a fim de equilibrá-lo. Por exemplo, duas pessoas que têm muitos amigos em comum deveriam conhecer um ao outro.

A segunda abordagem analisa o conteúdo da comunicação entre os indivíduos, à procura de indivíduos que compartilham interesses comuns. Por exemplo, duas pessoas que debateram sobre pesca e álgebra abstrata em e-mails devem finalmente se encontrar, pois são assuntos muito específicos.

Neste trabalho será adotado a primeira abordagem para prever/sugerir novos relacionamentos. Para esse fim, serão apresentadas ainda neste capítulo as principais métricas utilizadas na literatura, bem como a métrica proposta neste trabalho.

7.2 Trabalhos relacionados

As técnicas de previsão de relacionamentos podem usar as métricas da instância de um grafo social para determinar onde é provável surgirem novas ligações. Por exemplo, é mais provável que uma nova ligação seja incidente em um nó com um grau elevado do que um nó com um grau pequeno. No entanto, embora a previsão de novas ligações seja usada em muitas aplicações, tem havido pouca pesquisa nessa área. Esta seção lista os trabalhos que discutem técnicas de previsão de relacionamentos, e na próxima seção será discutido o uso de alguns aplicativos.

Em 2003 Popescul e Ungar citam sistemas de previsão feitos usando aprendizagem estatística (Popescul and Ungar, 2003). O sistema deles aprende os padrões da previsão de relacionamentos a partir de consultas a um banco de dados relacional, incluindo associações, seleções e agregações.

Taskar, Abbeel e Koller utilizaram modelos de Markov para aprender os padrões dos cliques e transitividade em páginas Web (Huang, Taskar, Abbeel, Wong and Koller, 2003). Ambos os sistemas de previsão incluem os atributos dos nós (por exemplo, texto da página web), além dos aspectos relacionais desses nós. Isso os torna mais poderosos do que os sistemas de previsão utilizando apenas métricas topológicas, porém possuem um domínio bastante específico para o problema que está sendo solucionado.

Em 2004 Popescul e Ungar aperfeiçoaram a previsão de relação entre autores e documentos em redes bipartidas usando técnicas de agrupamento (Popescul and Ungar, 2004). Eles agruparam os documentos por assunto e os autores por comunidades científicas, a fim de gerar novas entidades que foram utilizadas na regressão lógica das características e relações. O sistema foi testado em dados composto por um número igual de casos positivos e negativos. Eles conseguiram aumentar a precisão em relação aos modelos que não utilizam técnicas de agrupamento por cerca de quatro por cento em média.

Zhou e Schölkopf abordaram, de uma maneira diferente, três problemas relacionados a grafos: classificação, avaliação (*ranking*) e previsão de ligações (Zhou and Scholkopf, 2004). Eles definiram cálculos discretos para grafos e em seguida adaptaram a

regularização clássica dos casos contínuos para os dados do grafo. Embora seja matematicamente interessante esse estudo não inclui testes empíricos.

Huang, Liben-Nowell e Kleinberg testaram o poder de previsão apenas das métricas de proximidade, incluindo vizinhos comuns, a métrica de Katz e variantes do PageRank (Huang, Liben-Nowell and Kleinberg, 2003). Eles perceberam que algumas dessas métricas tinham uma precisão de previsão de até 16%, enquanto que a previsão aleatória tem precisão inferior a um por cento. Liben apresenta em um novo trabalho um amplo estudo sobre previsão de relacionamentos em redes sociais (Liben-Nowell and Kleinberg, 2007). Neste trabalho ele revê as análises feitas no artigo de 2003 (Huang, Liben-Nowell and Kleinberg, 2003). Sua hipótese era de que a previsão de ligações poderia ser feita apenas a partir da topologia. Isso foi considerado verdade já que seu sistema superou as previsões aleatórias por um fator de 50 vezes.

No entanto, as redes de colaboração, extraídas de *www.arxiv.org*, eram bem menores que as redes usadas em seus experimentos iniciais (entre 486 e 1790 nós). Assim, ele teria sido capaz de armazenar todas as suas redes em uma matriz de adjacência, fazendo os cálculos de caminho mais curto e métricas similares com custo computacional baixo. Por outro lado, as redes sociais de tamanho médio, compostas de cerca de cinqüenta mil nós, não podem ser armazenadas em matrizes de adjacência por causa das limitações de acesso a memória. Isso ocorre porque uma matriz de adjacência tem que ter uma entrada para cada um dos pares de nós, fazendo o total de memória necessária ser proporcional ao quadrado do número de nós na rede.

Assim, seus métodos de previsão baseados em caminho mais curto é um tanto impraticável para análise em grande escala. Felizmente, ele descobriu que os métodos de vizinhos comuns apresentavam resultados bastante satisfatórios. Entretanto, por definição, as abordagens que usam a métrica de vizinhos comuns não podem prever as ligações entre nós a uma distância superior a três, uma vez que a distância entre nós com vizinhos comuns será no máximo dois.

Apesar dos obstáculos computacionais envolvidos com as grandes redes sociais, elas podem, de fato, ser mais adequadas no estudo de previsão de relacionamentos. Liben-Nowell diz que quanto mais diversificada é uma rede social, mais fácil para que os nós sejam separados em grupos de interesse comum (por exemplo, grupos de interesses

de pesquisa comuns em redes de publicação). Com grafos menores, as pessoas tendem a formar ligações mais ao acaso, já que todos os nós tendem a ser mais semelhante.

7.3 Aplicações

Huang, Li e Chen investigaram a utilização de previsão de relacionamentos para melhorar a filtragem colaborativa em sistemas de recomendação (Huang, Li and Chen, 2005). Eles descobriram que a métrica de Katz foi a mais útil, seguida pela métrica de ligação principal, vizinhos comuns e, finalmente, a métrica Adamic\Adar. As métricas baseadas no caminho e nos vizinhos comuns superaram as métricas mais simples. Eles descobriram que a medida da distância entre nós não foi uma informação útil, provavelmente porque a maioria dos pares de nós do seu conjunto de dados pode ser ligada por um caminho curto.

Farrell, Campbell e Myagmar usam previsão de relacionamentos para projetar um sistema que recomenda novas ligações acadêmicas para pesquisadores de uma conferência sobre ciência da computação (Farrel, Campbell and Mayagmar, 2005). Os pesquisadores mais experientes que usaram o sistema encontraram pouca utilidade para o mesmo, mas os pesquisadores mais novos acharam que a ferramenta era útil ao recomendar colegas em potencial e palestras de seus interesses.

Os autores desse trabalho são defensores das relações humanas orientadas computacionalmente e acreditam que os sistemas de rede social serão úteis para ajudar os humanos a lidar com o grande número de contatos profissionais que necessitam manter. O sistema mantém informações sobre os seus contatos e seus relacionamentos através da análise dos dados em grafos bipartidos, tais como: trabalhos publicados e pesquisadores, ou reuniões e empresários.

Em sua tese de doutorado Zhu (2003) utilizou previsão de relacionamentos para determinar qual página Web um usuário provavelmente irá visitar no seu próximo acesso, a fim de melhorar a navegação e a eficiência dos sites. Isso foi feito armazenando listas de várias páginas Web visitadas como uma cadeia de Markov. Normalmente, quando são previstos novos relacionamentos é assumido que esses relacionamentos são independentes. Em outras palavras, uma pessoa faz novos relacionamentos um a um, e

não uma seqüência de novas conexões. Isso é claramente diferente de como Zhu prevê visitas a páginas web. Essa técnica seria útil apenas se fosse possível supor que com a formação de um novo relacionamento com a pessoa A será criado um novo relacionamento com a pessoa B, porém não é dessa maneira que os relacionamentos surgem nas redes sociais.

7.3.1 Complemento das Ligações (Link Completion)

Link Completion é também um problema de análise de relacionamentos bastante semelhante ao problema de previsão de relacionamentos (Goldenberg, Kubica, Komarek, Moore and Schneider, 2003). A diferença dessa técnica está no fato de seu objetivo ser determinar o nó que está faltando em um par de nós que definem uma ligação do grafo. Em outras palavras, caso se tenha um conjunto de dados com alguns relacionamentos parciais, nos quais não se conhecem os dois elementos responsáveis pelos relacionamentos, é preciso determinar qual é o nó faltante desses relacionamentos.

Assim, *link completion* é considerado um problema mais difícil e mais genérico que o problema de previsão de relacionamentos, já que ao invés de tentar determinar qual o par de nós têm maior probabilidade de estabelecerem um novo relacionamento, tenta-se determinar para qual nó outro nó está se relacionando. Um exemplo desse problema é quando um usuário compra cinco livros on-line e o nome de um livro é corrompido na transferência. Um algoritmo de *link completion* poderia prever o nome do livro extraviado com base no nome do usuário e os outros livros que ele comprou.

7.3.2 Descoberta de Ligações Anômalas

Em 2005 Rattigan e Jensen (Rattigan and Jensen, 2005) apresentaram o problema da descoberta de ligações anômalas em resposta às enormes dificuldades dos problemas de previsão de relacionamentos. Eles argumentaram que a previsão de relacionamentos tem dificuldades insuperáveis porque o número de pares de nós que precisam ser avaliados aumenta de forma quadrática em proporção ao número de nós em uma rede.

Além disso, segundo eles, as redes sociais são muito esparsas, levando a pouquíssimos casos positivos. Isso torna quase impossível para os sistemas de previsão aprenderem as diferenças significativas entre as métricas, em casos positivos e negativos. Há tão poucos casos positivos que eles acabam sendo “engolidos” pelos casos negativos que têm métricas semelhantes.

Rattigan e Jensen recomendam que o foco da pesquisa seja em descobrir as ligações anômalas. Isso significa detectar quais são as ligações em uma rede social que surgiram em um determinado intervalo de tempo e que são surpreendentes, ou seja, que não era esperado de acorrer; tais como, dois nós que passam a se relacionar, mas que possuíam muito poucos vizinhos em comum ou que estavam muito distantes na etapa de tempo anterior à que ocorreu a ligação.

Uma aplicação importante dessa técnica é a descoberta de ligações anômalas entre dois indivíduos indicando uma colaboração criminosa. A descoberta de ligação anômala é complementar a previsão de relacionamentos no sentido de que ambos usam as mesmas métricas para avaliar quais ligações são surpreendentes e quais são as ligações esperadas. Assim, as pesquisas que envolvem qualquer um dos temas irão beneficiar o outro.

7.3.3 Detecção de Ligações

Detecção de ligações é semelhante à previsão de ligações, porém envolve determinar onde existe um vínculo já estabelecido que não esteja sendo representado no sociograma, talvez devido à falta de dados. Nessa técnica não se deseja prever o que irá acontecer no futuro, o objetivo é saber quais são as relações que de certa forma já existem na rede social, mas que não estão explícitas, ou seja, estão ocultas.

Um exemplo é prever a influência, considerado um tipo de ligação escondida na análise de redes sociais, que um investigador exerce sobre outro, baseado em uma rede social de publicações. Os criminosos gostariam que todas as suas ligações para outros criminosos não fossem conhecidas e talvez intencionalmente eles possam esconder suas ligações o máximo possível. Da mesma forma, as empresas gostariam de saber todas as ligações entre as pessoas mesmo quando baseadas em informações incompletas para que suas estratégias de *marketing* sejam o mais eficiente e abrangente possíveis.

Liben-Nowell (Liben-Nowell and Kleinberg, 2007) e Taskar (Taskar, Wong, Abbeel and Koller, 2004) mencionam esse problema como uma aplicação de previsão de ligações na introdução de seus trabalhos. Alguns pesquisadores utilizam as técnicas de detecção de ligações como um aplicativo de testes para seus trabalhos (Komarek, 2004; Popescul and Ungar, 2003). A técnica de detecção de ligações em redes sociais usando métricas topológicas não deve ser confundida com outros tipos de detecção de ligação, por exemplo, a detecção de ligações históricas (Chen, Farahat and Brants, 2004), onde artigos de notícias são comparados para determinar se eles são sobre o mesmo evento.

7.4 Métricas

A previsão de relacionamentos é baseada no cálculo de várias métricas, de forma que, através de seus resultados, é possível prever/sugerir novos relacionamentos. A métrica mais comum e mais utilizada para sugerir um relacionamento entre duas pessoas é verificar o número de vizinhos em comum dessas pessoas, quanto maior o número de vizinhos comuns entre eles maior a probabilidade dessas duas pessoas desenvolverem um novo relacionamento.

A Figura 46 ilustra o caso descrito anteriormente. A Figura 46(a) ilustra uma rede social com relacionamento de amizade, na qual as pessoas ‘A’ e ‘B’ não estão relacionadas. Entretanto, essas pessoas possuem vários amigos comuns, sendo natural a sugestão de um novo relacionamento entre elas, como representado na Figura 46(b).

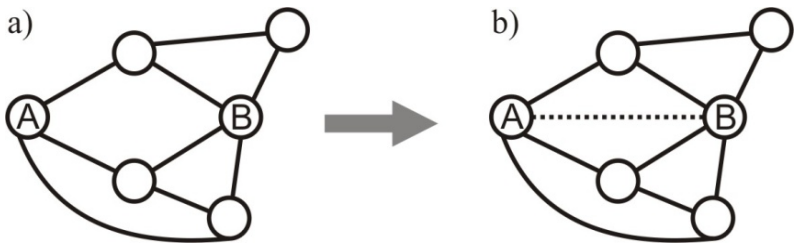


Figura 46 - Exemplo de Previsão de Relacionamentos

As métricas mais populares para o estudo de previsão de relacionamentos são: grau do nó, vizinhos comuns e a métrica de Katz. O *grau do nó* indica o número de

conexões que o nó possui. No caso das redes sociais científicas essa métrica irá indicar quantos relacionamentos o pesquisador possui.

A métrica de *vizinhos comuns* indica o número de nós comuns conectados a ambos os nós, ou seja, o número de pesquisadores que são conhecidos pelos dois pesquisadores.

Finalmente, a *métrica de Katz* contabiliza um valor que indica como é o acesso entre os dois nós, ou seja, baseando-se nos caminhos do grafo social essa métrica irá mensurar como a informação trafega de um pesquisador para outro (Huang, Liben-Nowell and Kleinberg, 2003; Potgieter, April, Cooke and Osunmakinde, 2009).

As métricas são utilizadas para quantificar o comportamento local dos agentes sociais. O surgimento de novos relacionamentos está vinculado a esse comportamento dos agentes na rede social.

Neste trabalho foi utilizada uma nova métrica composta por estas três métricas: grau do nó, os vizinhos comuns e métrica de Katz. A principal diferença entre as métricas grau do nó, vizinhos comuns e a métrica de Katz é que as duas primeiras são baseadas na análise da vizinhança do nó, enquanto que a métrica de Katz é baseada nos caminhos dos nós. Métricas baseadas no caminho avaliam todos os relacionamentos entre uma dupla de pesquisadores, enquanto que as métricas baseadas na vizinhança avaliam apenas as conexões diretas de cada pesquisador.

A definição dessas e de outras métricas que também são utilizadas na resolução de problemas de previsão de relacionamentos está na Tabela 9 (Potgieter, April, Cooke and Osunmakinde, 2009). Para um melhor entendimento da tabela considere a as seguintes definições: # é o número de elementos de um conjunto. U_n é o conjunto de relações bidirecionais da rede social no tempo *n*. Uma relação bidirecional entre os nós v_i e v_j é definida por $u_{i,j}$ e $u_{j,i}$. $\Gamma(v_i)$ é o conjunto dos vizinhos do nó v_i, isto é, o conjunto $\{v_j : u_{i,j} \in U\}$. Além disso, $P(v_i, v_j)$ é o conjunto de todos os caminhos mais curtos do vértice v_i ao vértice v_j. $P(v_i, v_j, v_x)$ é o conjunto de todos os caminhos mais curtos entre os nós v_i e v_j que passam pelo nó v_x. Finalmente, $\mathbf{caminhos}_{v_i,v_j}^{<l>}$ é o conjunto de todos os caminhos mais curtos de tamanho *l* entre v_i e v_j.

Tabela 9 - Conjunto de Métricas adotadas pela literatura.

Métrica	*Definição Matemática*	*Descrição*
Grau do nó v_i	$\#\{u_{j,i} : u_{i,j} \in U\}$ ou $\#\Gamma(v_i)$	Número de ligações do nó v_i a qualquer nó no momento *n*.
Caminhos Comuns	$\sum_{v_i \in V} \sum_{v_k \in V} \frac{\#P(v_j, v_k, v_i)}{\#P(v_j, v_k)}, \quad v_k \neq v_j$	Total de caminhos mínimos entre v_j e v_k que passam por v_i, dividido pelo total de caminhos mínimos entre esses dois elementos.
Vizinhos Comuns	$\#\{v_k : u_{i,k} \in U_n, u_{k,j} \in U_n\}$ ou $\#\{\Gamma(v_i) \cap \Gamma(v_j)\}$	Número de nós ligados a ambos os nós v_i e v_j, ou seja, número de vizinhos comuns que esses elementos possuem.
Similaridade Adamic/Adar	Caso Geral: $\sum_{\substack{z: \text{característica} \\ \text{comum}}} \frac{1}{\log(\text{freqüência}(z))}$ Caso dos Vizinhos Comuns: $\sum_{v_z \in \{\Gamma(v_i) \cap \Gamma(v_j)\}} \frac{1}{\log(\#\Gamma(v_z))}$	Número de características compartilhadas pelos nós, dividido pelo *log* da freqüência das características. Essa métrica avalia as características raras mais fortemente.
Ligação Principal	$\#\{v_k : u_{i,k} \in U_n\}.\#\{v_k : u_{j,k} \in U_n\}$ ou $\#\Gamma(v_i).\#\Gamma(v_j)$	Produto do número de arestas incidentes nos dois nós.
Métrica de Katz	$\sum_{l=1}^{\infty} \beta^l.\#(\mathbf{caminhos}_{v_i,v_j}^{<l>})$ onde $\mathbf{caminhos}_{v_i,v_j}^{<l>}$ é o conjunto de caminhos de tamanho *l* do nó v_i ao nó v_j.	Soma ponderada de todos os caminhos de tamanhos variados existentes entre os nós.

Analisando a definição das métricas, pode ser observado que o valor da métrica de Katz para pesquisadores que possuem muitos relacionamentos será maior do que o valor dessa mesma métrica para os pesquisadores que possuem poucos relacionamentos. Isso ocorre porque a quantidade de caminhos para um pesquisador está diretamente relacionada aos seus relacionamentos, isto é, ao seu grau. O objetivo da métrica proposta neste trabalho é a obtenção de uma pontuação que reflita as chances reais de dois pesquisadores se relacionarem no futuro. A definição da métrica composta está descrita em detalhes na próxima seção.

7.5 Análise das Redes Dinâmicas

Os problemas de previsão de relacionamentos com embasamento temporal são diferentes dos problemas de previsão de relacionamentos tradicionais, nos quais não há nenhum aspecto temporal (Clauset, Moore and Newman, 2008).

Muitas vezes a análise do passado pode auxiliar na descoberta de um comportamento futuro. Por exemplo, suponha que dois pesquisadores tenham se relacionado em 1980, mas há 20 anos esses pesquisadores não estabelecem novas conexões. Analisando a rede social dessa instituição ao longo do tempo ficará claro que, provavelmente, esses pesquisadores mudaram os rumos de suas pesquisas e deixaram de trabalhar em conjunto. Essa conclusão se torna óbvia quando é feita a análise temporal da rede social, porém pode ser bastante complexa utilizando apenas as informações de um momento dessa mesma rede.

Muitos estudos em redes sociais descobrem padrões em grafos estáticos, identificando propriedades em um 'retrato' da rede social. Entretanto, dada a falta de informação sobre a evolução das redes ao longo dos períodos, tem sido difícil analisar as tendências ao longo do tempo (Leskovec, Kleinberg and Faloutsos, 2005).

Atualmente, existem poucos estudos que utilizam informações sobre o tempo na Previsão de Relacionamentos (Acar, Dunlavy and Kolda, 2009; Potgieter, April, Cooke and Osunmakinde, 2009). Nas redes sociais científicas é possível fazer uma análise das tendências dos relacionamentos com base no passado da rede social, já que nesse tipo de rede os relacionamentos possuem informações temporais.

Os estudos nessa seção visam analisar o comportamento das redes sociais científicas no decorrer do tempo. O objetivo é analisar as tendências nessas redes e como a análise do tempo pode influenciar na previsão e sugestão dos novos relacionamentos.

7.5.1 Trabalhos Relacionados

A análise das mudanças nas redes sociais ao longo do tempo é chamada de análise de redes dinâmicas, que neste trabalho também é referenciado como análise temporal das redes sociais. Hoje o estudo dessas redes é uma vertente das agências de inteligência,

tendo em conta o aumento das atividades terroristas e outros grupos criminosos (Coffman, Greenblatt and Marcus, 2004). Tais grupos têm sido rotulados por *redes escuras*, e sua estrutura e comportamento normal variam muito quando comparadas com as redes sociais tradicionais. Nessas redes, conforme descrito no capítulo 2, os relacionamentos não são explícitos, já que o objetivo é esconder as pessoas ou alguma informação da rede. Por exemplo, para maior eficiência do comércio há um sigilo na troca de informações, produzindo uma estrutura com padrões de comunicações incomuns (Fellman and Wright, 2004).

Carley é uma das pesquisadoras mais produtivas na modelagem de redes escuras usando as técnicas de redes dinâmicas. Ela criou um programa de rede dinâmica, *DyNet*, onde múltiplos agentes modelam o comportamento social dos seres humanos, com acesso aos recursos e organizações (Carley, 2003). Esse programa é usado para entender a evolução da rede, e a melhor maneira de desestabilizar as redes terroristas. Essas técnicas são bastante poderosas e possuem um domínio muito específico e complexo, o que torna difícil a expansão desse modelo para outros problemas.

Houve também alguns estudos puramente teóricos feitos para entender as alterações sofridas pela estrutura das redes sociais ao longo do tempo. Os trabalhos de Holme concentram-se nesse assunto, incluindo estudos sobre mudanças das métricas de uma rede de namoro da Internet sueca, rede chamada Pussokram (Holme, 2003; Holme, Edling and Liljeros, 2004). Os trabalhos de Holme investigaram as tendências das métricas agregadas aos grafos, como, por exemplo, o comprimento do percurso médio e grau médio de um nó.

7.5.2 Análise Temporal

Leskovec, Kleinberg e Faloutsos (2005) afirmam que pouco tem sido feito na análise de tendências do grafo em longo prazo:

> *Muitos estudos têm descoberto padrões em grafos estáticos, identificando as propriedades em um único instante de uma grande rede social, ou em um número muito pequeno de instâncias. Entretanto, dada a falta de informação sobre a evolução da rede durante longos períodos, tem sido*

difícil para converter essas propriedades em informações que representem as tendências ao longo do tempo.

Seus estudos constataram que ao longo do tempo os grafos aumentam em densidade e a distância média entre os nós diminui. Essa constatação foi contrária às crenças existentes de que, em média, o grau do nó permanece constante e a distância média entre os nós aumenta lentamente. Eles alegaram que os modelos de geração de grafos não são realistas e fizeram a proposta de um novo modelo de geração.

Desikan e Srivastava estudaram as mudanças sofridas ao longo do tempo pelas métricas de um grafo que representa um conjunto de páginas Web (Desikan and Srivastava, 2004). Eles descobriram que as métricas temporais, tais como a sua utilização na descoberta da popularidade das páginas Web, podem ser usadas na reorganização das filas de classificação. Por exemplo, as páginas que se tornaram recentemente mais populares devem aparecer antes na fila de classificação do que as páginas que estão mais obsoletas.

As métricas temporais são um complemento útil para as métricas tradicionais estáticas no estudo de algumas redes sociais. Basicamente, as métricas temporais se derivaram das métricas estáticas. De forma que a formulação matemática da métrica continua a mesma, mas o conteúdo das ligações na rede social carrega informações temporais que acabam sendo refletidas na métrica.

A grande maioria das técnicas de previsão dos relacionamentos é limitada pelo fato delas preverem a evolução de uma entidade complexa analisando um único momento do passado. Fazendo uma analogia, considere que uma pessoa fez o lançamento de uma bola. Caso se deseje prever qual a trajetória dessa bola serão necessários no mínimo dois instantes diferentes da trajetória para que se possa ter uma boa previsão. Caso seja informado apenas um instante não é possível saber nem se a bola está movendo da direita para a esquerda ou da esquerda para a direita.

Em outras palavras, é necessário ter o conhecimento da evolução da rede, examinando mais intervalos de tempo do que somente a última posição da rede social. A posição de uma rede social é calculada pelas métricas tradicionais de análise, mas a evolução da rede só pode ser determinada através de métricas temporais usando o histórico de mudanças da rede.

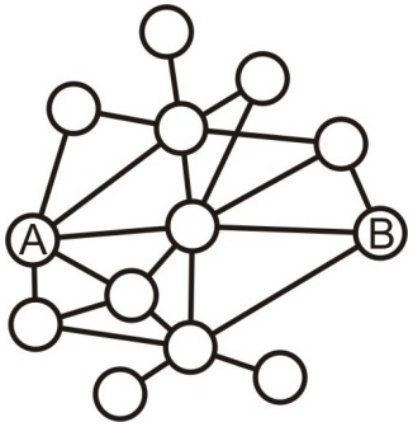


Figura 47 - Previsão de relacionamentos desconsiderando a evolução da Rede Social

Para ilustrar a fragilidade das técnicas de previsão de relacionamentos tradicionais, considere a tentativa de prever novas ligações em um grafo avaliando apenas a última evolução da rede. Suponha que se deseja prever se surgirá algum relacionamento entre os elementos A e B representados na Figura 47. Pode ser observado que esses nós não possuem os graus mais altos, nem os caminhos mais curtos e nem o maior número de vizinhos em comum. Assim, pode até ser sugerido que uma ligação irá se formar entre eles, mas o valor da métrica será baixo o que é um sinal de incerteza com relação a essa previsão.

Por outro lado, considere o mesmo grafo apresentado como parte da seqüência da evolução temporal da rede social, como ilustrado na Figura 48. Esses grafos proporcionam muito mais informação do que a etapa final sozinha. Analisando essa seqüência pode ser dito com bastante certeza que os nós A e B possuem grandes chances de formar um novo relacionamento. Esses nós têm sido muito mais ativos na rede do que os outros nós. O grau desses nós cresceu muito acima da média do crescimento dos outros nós.

Caso fosse possível analisar quantas mensagens foram trocadas entre os membros dessa rede social, provavelmente se chegaria à conclusão de que esses nós também foram muito mais ativos que os outros. Como esses elementos estão mais ativos, eles têm mais chance de estabelecerem novos contatos e criarem novas ligações na rede, dando uma maior confiança à sugestão de um relacionamento entre eles.

A proposta deste trabalho no módulo de previsão de relacionamentos é prever novos relacionamentos baseando-se na evolução da rede social e não apenas através de uma análise estática da rede. Com a análise temporal da rede será possível encontrar padrões evolutivos que facilitem a previsão de novos relacionamentos, reduzindo a taxa de erros.

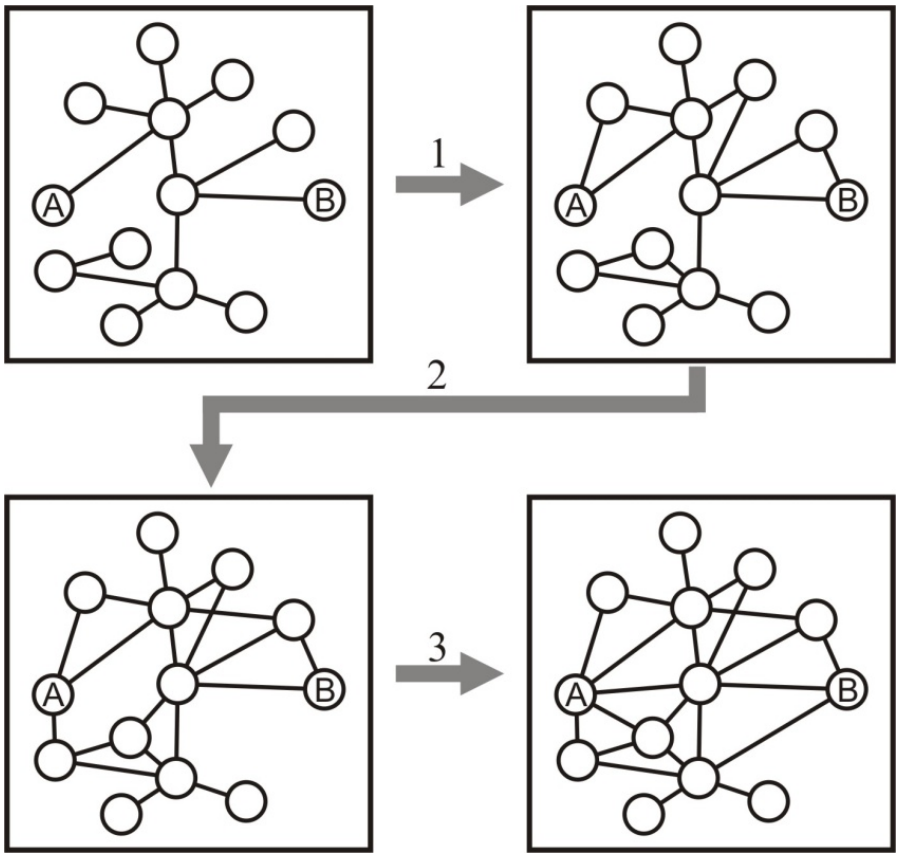


Figura 48 - Previsão de relacionamento avaliando a evolução da Rede Social

Como dito anteriormente, a informação temporal é adicionada na rede social através dos seus relacionamentos. Neste trabalho foram utilizadas funções de penalização com base no ano do relacionamento para incorporar a informação da evolução da rede nos relacionamentos. Alguns estudos adicionam a análise do tempo de maneira linear, ou seja, apenas acumulam todas as informações do passado sem diferenciar as informações mais antigas das mais recentes (Liben-Nowell and Kleinberg, 2007). Outros utilizam funções para diferenciar as informações mais antigas das mais recentes (Acar, Dunlavy

and Kolda, 2009), como é o caso deste trabalho. As funções de penalização utilizadas neste trabalho estão definidas no capítulo 4.

Com o uso das funções de penalização com base no ano do relacionamento, os pesquisadores que não reforçarem seus vínculos com o passar dos anos terão seus relacionamentos removidos da rede social. Assim, os relacionamentos terão mais ou menos peso de acordo com as evoluções da rede social científica. Conseqüentemente, as métricas também terão seus valores influenciados por essas funções. Pode ocorrer, por exemplo, do grau de um nó reduzir com o passar dos anos quando um dos seus relacionamentos for removido da rede.

A métrica composta visa unir as informações temporais e a atividade dos nós na rede social. Os conceitos e a definição da métrica proposta neste trabalho estão na próxima seção.

7.6 Definição da Métrica Composta

Como dito anteriormente, as métricas mais utilizadas são a métrica de Katz, vizinhos comuns e grau do nó. Com o intuito de analisar o comportamento evolutivo da rede social científica, todos os relacionamentos novos de cada ano foram analisados juntamente com os valores das três métricas calculadas para eles.

Por ser considerada pela literatura a melhor métrica para previsão de relacionamentos a métrica de Katz foi utilizada como base nessa etapa de análise do comportamento evolutivo (Acar, Dunlavy and Kolda, 2009; Huang, Li and Chen, 2005; Huang and Lin, 2009; Liben-Nowell and Kleinberg, 2007). O cálculo da métrica de Katz irá produzir um valor e quanto maior for esse valor, maior será a probabilidade de dois pesquisadores se relacionarem no futuro.

O valor da métrica de Katz está diretamente relacionado ao peso dos relacionamentos da rede social. Se os caminhos que ligam dois pesquisadores são formados por relacionamentos com peso alto, então a métrica irá produzir um valor alto, se os relacionamentos tiverem um peso baixo, então o valor da métrica também será mais baixo.

Conforme apresentado no Capítulo 4, o cálculo do peso do relacionamento depende dos pesquisadores que estão envolvidos na conexão. Assim, se um dos pesquisadores se envolver em relacionamentos mais novos com outros pesquisadores e os relacionamentos mais antigos entre eles não forem reforçados, o peso da conexão entre eles vai diminuir de um ano para outro.

Por exemplo, suponha que em um determinado ano os pesquisadores A e B fossem ligados por muitos tipos de relacionamentos e que o pesquisador B tivesse alguns poucos relacionamentos com o pesquisador C. Se no ano seguinte, o pesquisador B criar novos relacionamentos com o pesquisador C e B não tiver novos relacionamentos com o pesquisador A, então o peso do relacionamento entre A e B irá diminuir, enquanto o peso entre B e C irá aumentar. Esse comportamento reflete como dois pesquisadores perdem seus contatos, pois o fluxo de informações entre eles reduz em comparação com outros pesquisadores na rede social. A nova métrica proposta vai refletir o comportamento anual de cada pesquisador, refletindo a evolução temporal da rede científica social.

Durante a análise dos relacionamentos novos foi observado que a métrica de Katz aplicada à rede social modelada neste trabalho apresentou bons resultados. Assim, pode-se concluir que a métrica de Katz reflete bem a evolução da rede social científica multi-relacional quando os relacionamentos possuem uma componente que reflete o tempo.

Existem casos, como era de se esperar, que a métrica não produz um valor alto, mas, apesar disso, um novo relacionamento surgiu. Analisando esses casos foi observado que alguns relacionamentos surgem ao acaso, sem um padrão pré-definido. Esse tipo de comportamento, por não ser padronizado, inviabiliza a criação de uma métrica para identificá-los.

Entretanto, existe um grupo de relacionamentos novos que não foi identificado pela métrica de Katz, mas que possui uma característica especial. Esse conjunto de relacionamentos é formado por pesquisadores que entraram recentemente na rede social e ainda possuem poucos relacionamentos.

A métrica de Katz calcula um valor baseando-se em todos os caminhos possíveis que existem entre os dois pesquisadores. O problema é que como um pesquisador recém chegado na rede social possui poucos relacionamentos o número de caminhos que chegam até ele também é menor e, conseqüentemente, o valor da métrica de Katz desse

pesquisador será baixo. Nesse caso, as chances da métrica de Katz acertar os relacionamentos de pessoas novatas são muito baixas.

Além disso, os pesquisadores mais novos são mais ativos na rede e estabelecem relacionamentos com pessoas que eles ainda não se relacionaram, enquanto que os pesquisadores mais antigos tendem a se relacionar novamente com pesquisadores que eles já estão relacionados. De uma maneira geral, os recém-chegados estão apenas começando a criar os seus laços e conhecendo novas pessoas e, portanto, estão abertos a novos relacionamentos. Essa análise só foi possível devido à análise evolutiva da rede social científica. Se essa análise fosse feita em uma rede estática não seria possível essas descobertas.

Para resolver esse problema, é proposta uma métrica composta para analisar a evolução dos relacionamentos e sugerir/prever novos relacionamentos. A proposta é construir uma métrica composta pelas três métricas mais utilizadas na literatura. O objetivo é construir uma métrica que combine a precisão da métrica de Katz com as análises do comportamento evolutivo da rede social científica. A métrica composta está definida na equação (24).

$$\textit{Métrica Composta} = \frac{\sum_{l=1}^{\infty} \beta^{l}.\#(\text{caminhos}_{v_i,v_j}^{<l>})}{(\#\Gamma(v_i)+\#\Gamma(v_j))-\#\{\Gamma(v_i)\cap\Gamma(v_j)\}+1} \tag{24}$$

A métrica composta é calculada dividindo o valor da métrica de Katz pela soma do grau dos nós menos o número de vizinhos comuns desses nós. É somado 1 no denominador para evitar uma divisão por zero.

A divisão da métrica de Katz pela soma dos graus dos pesquisadores tem o objetivo de produzir uma métrica que seja relativa ao grau dos nós. A subtração do número de vizinhos comuns é para permitir que os pesquisadores com muitos vizinhos em comuns tenham uma redução menor no valor da métrica de Katz.

Assim, a métrica composta irá produzir um valor relativo ao grau dos pesquisadores envolvidos no cálculo da métrica de Katz. Dois fatores contribuem para o bom resultado da métrica composta o primeiro deles é o peso do relacionamento que possui informações temporais, o segundo é que pesquisadores antigos com grau alto terão

uma redução na métrica de Katz maior que um pesquisador recém-chegado, que provavelmente está mais ativo na rede.

A Figura 49 ilustra a análise feita sobre a rede social científica para avaliar o comportamento evolutivo da mesma. A Figura 49(a) representa o ano T1 no qual os pesquisadores A e B estão relacionados.

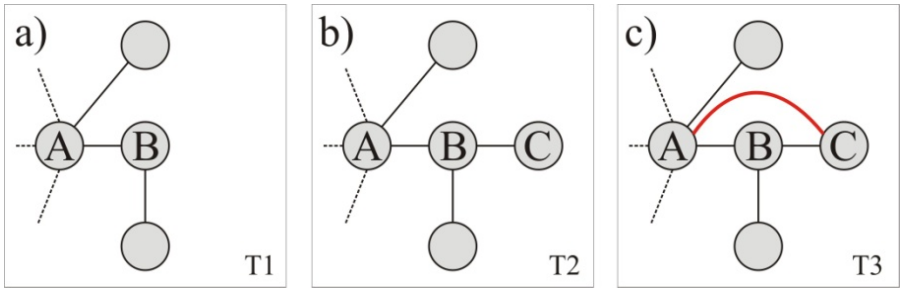


Figura 49 - Análise evolutiva dos pesquisadores novatos

No ano seguinte T2, representado na Figura 49(b), surge um novo pesquisador na rede através de um relacionamento com o pesquisador B. Como o pesquisador C não estava na rede no ano anterior não foi possível saber de antemão que o relacionamento iria surgir.

Já no ano T3 as métricas de Katz foram calculadas para todos os pesquisadores inclusive para o pesquisador C. Porém, a métrica de Katz do pesquisador A para outros pesquisadores foi muito maior que o valor obtido quando o cálculo foi feito entre os pesquisadores A e C. Isso se deve ao fato de só existir um caminho entre A e C, mas existirem vários caminhos entre A e os outros pesquisadores.

Utilizando apenas a métrica de Katz não foi possível prever o surgimento do relacionamento entre os pesquisadores A e C, representado em vermelho na Figura 49(c). Entretanto, utilizando a métrica composta os valores obtidos para o pesquisador A foram reduzidos e o valor da métrica entre A e C foi representativo quando comparado aos outros valores. Assim, com o uso da métrica composta foi possível prever esse novo relacionamento.

O principal objetivo deste trabalho com as técnicas de previsão de relacionamentos é proporcionar aos pesquisadores um módulo capaz de fazer boas sugestões de relacionamentos. Assim, a nossa ferramenta poderá ser utilizada, por

exemplo, para encontrar um novo membro para uma banca de defesa de tese, ou para criar um novo vínculo no caso de uma transferência de uma instituição para outra.

7.7 Cálculo da Métrica

O cálculo da métrica composta está diretamente relacionado ao cálculo das três métricas que a definem. O número de vizinhos comuns e o grau do nó são duas métricas simples de serem calculadas. Já a métrica de Katz é mais complexa e, por isso, ela está sendo analisada com mais detalhe nesta seção.

Embora por definição a métrica de Katz represente a soma do número de caminhos de tamanho *l* entre dois elementos, a grande maioria dos trabalhos de previsão de relacionamentos utiliza o peso do relacionamento no cálculo da métrica. Caso seja considerado apenas o número de caminhos a métrica não será capaz de diferenciar os pesquisadores que têm maior probabilidade de se relacionarem.

Caso não seja considerado o peso das conexões no cálculo das sugestões de novos relacionamentos, o aumento do número de relacionamentos entre pares de pesquisadores não é refletido na métrica de Katz. Por exemplo, suponha que em 2010 dois pesquisadores possuíssem 20 caminhos de tamanho 6 entre si. Em 2011 eles continuam possuindo os mesmos 20 caminhos de tamanho 6 entre si, porém, a força dos relacionamentos entre eles aumentou, ou seja, os vínculos entre esses dois pesquisadores foram reforçados por novos relacionamentos. Nesse caso, a métrica de Katz sem o peso terá o mesmo valor nos dois anos, pois o número de caminhos entre os pesquisadores se manteve o mesmo. Por outro lado, a métrica de Katz com peso será diferente nos dois anos sendo maior em 2011 do que em 2010.

A métrica de Katz pode ser calculada através da decomposição SVD ou através da análise do peso dos caminhos entre os pesquisadores. Nas próximas seções serão apresentadas essas duas formas de cálculo da métrica de Katz.

7.7.1 Decomposição SVD

Singular value decomposition (SVD) é um tipo de decomposição matricial que tem sido utilizada para extrair informações de grafos que modelam uma rede social. Essa técnica pode ser utilizada, por exemplo, em problemas de agrupamento e de previsão de relacionamentos.

Alguns pesquisadores utilizam essa técnica como sendo uma ferramenta de particionamento em grafo e como ferramenta de detecção de nós anômalos, que são trabalhos desenvolvidos para análise de redes escuras (ex.: redes terroristas) (Skillicorn, 2004; Xu, Zhang, Han and JieWang, 2006).

A decomposição SVD transforma os dados de maneira a converter correlação em proximidade (Acar, Dunlavy and Kolda, 2009; Boyd, Fitzgerald, Mahutga and Smith, 2010; Skillicorn, 2004; Xu, Zhang, Han and JieWang, 2006). A decomposição matricial é feita da seguinte maneira:

$$\overline{TR} = USV' \tag{25}$$

onde $\overline{TR}$ é a matriz *MxM* de fluxos máximos entre cada par de pesquisadores, definida no capítulo 4, *U* e *V* são matrizes ortogonais (V' é a matriz transposta de V), e S é uma matriz diagonal, na qual os valores da diagonal são chamados de *valores singulares* ($\sigma_1 > \sigma_2 > ... > \sigma_M > 0$). As linhas de U podem ser consideradas coordenadas dos pontos que correspondem aos objetos do conjunto de dados.

Geralmente, o tamanho das matrizes do lado direito é limitado para forçar que a decomposição represente os dados de uma maneira mais compacta. Assim, a decomposição truncada em *k* é dada por:

$$\overline{TR} \approx U_k S_k V_k'. \tag{26}$$

A decomposição SVD pode ser utilizada em diferentes tipos de problemas de análise do conjunto de dados. A aplicação mais comum é a redução da dimensionalidade do conjunto de dados. Essa é a aplicação mais comum da técnica de decomposição em valores singulares, pois ela possibilita a transformação de um conjunto de dados com alta dimensão (muitos atributos) em um conjunto de dados de dimensão menor, perdendo o mínimo de informação possível no decorrer do processo.

Um dos benefícios da redução da dimensionalidade do conjunto de dados é tornar possível a análise visual desse conjunto, caso seja adotado $k = 2$ ou $k = 3$. Assim, visualizando o gráfico do conjunto de dados é possível compreender no mínimo as estruturas mais significantes desse conjunto.

Outro problema que pode ser solucionado por técnicas de decomposição em valores singulares é o problema de agrupamento. Existem duas abordagens que podem ser tomadas nos problemas de agrupamento. A primeira é aplicar algum algoritmo de agrupamento na matriz de decomposição. A segunda é utilizar as propriedades da decomposição SVD em uma abordagem chamada *agrupamento espectral* (Kannan, Vempala and Vetta, 2000). Nessa abordagem, é feito o produto interno de cada elemento do conjunto de dados com os elementos da matriz de decomposição. Aqueles no qual o produto interno é menor que $1/2$ o grupo é definido. Por exemplo, se o produto interno entre um elemento e a terceira linha da matriz de decomposição for menor que $1/2$, então o grupo desse elemento é o grupo 3.

Outra aplicação da decomposição SVD é a classificação dos objetos através de suas correlações. Os trabalhos de previsão de relacionamentos desenvolvidos através de decomposição SVD utilizam as informações sobre a correlação entre os objetos para prever se eles irão se relacionar no futuro ou não.

Como dito anteriormente, cada linha da matriz *U* pode ser identificada como um ponto no espaço *k*-dimensional. Suponha que uma linha seja desenhada da origem até cada um desses pontos da matriz *U*. Então, o ângulo entre esses vetores revela a correlação entre esses pontos.

Dois pontos que sejam fortemente correlacionados positivamente terão vetores que estarão próximos entre si. O produto interno desses pontos terá um valor alto e positivo. Dois pontos que são fortemente correlacionados negativamente terão um produto interno alto e negativo. Dois pontos que não estão correlacionados terão um produto interno próximo de zero.

Outra maneira do produto interno entre os elementos ser próximo de zero é quando esses elementos estão muito próximos da origem. Assim, como o objetivo é identificar os elementos que estejam correlacionados positiva ou negativamente,

classificar os objetos considerando a distância desses pontos até a origem permite que os objetos mais interessantes sejam selecionados, ou seja, serão sempre selecionados objetos que realmente expressam suas correlações. Caso o produto interno seja próximo de zero significa que os elementos realmente não são correlacionados.

Essa aplicação da decomposição SVD pode ser utilizada na resolução de problemas de previsão de relacionamentos. Os objetos fortemente correlacionados possuem maior probabilidade de se relacionarem no futuro.

Embora seja uma técnica muito interessante para a análise de redes sociais, ela não foi adotada neste trabalho. Entretanto, o seu uso será avaliado em um trabalho futuro, no qual serão comparados os resultados obtidos e identificadas as vantagens e desvantagens dessa metodologia.

7.7.2 Soma do Peso dos Caminhos

Ao invés de utilizar decomposição SVD na resolução do problema de previsão de relacionamentos, neste trabalho foi calculado o fluxo máximo entre os elementos considerando o peso de todos os caminhos existentes entre eles. Para somar o peso de todos os caminhos entre os pares de pesquisadores foram feitos testes para definir qual o tamanho máximo dos caminhos deveria ser adotado. Nos testes foi feito $3 \leq l \leq 6$, onde l é o tamanho máximo do caminho entre os elementos, conforme definido na equação (24).

A maior dificuldade do uso da métrica de Katz é encontrar todos os caminhos com tamanho variando de 1 à l entre dois nós do grafo social. Para solucionar esse problema foi utilizado o algoritmo de Yen (Hershberger, Maxely and Suriz, 2007; Martins and Pascoal, 2003). Esse algoritmo foi desenvolvido para buscar todos os caminhos mais curtos que existem entre dois nós independentemente do comprimento do caminho, ou seja, o algoritmo lista todos os caminhos mais curtos de tamanho 1, 2, 3, e assim por diante.

Para ilustrar o resultado desse algoritmo considere que se deseja encontrar os caminhos existentes entre os nós s e t no grafo ilustrado na Figura 50. Existem vários caminhos diferentes no grafo representado nessa figura e o objetivo do algoritmo é

encontrar todos esses caminhos. O resultado obtido aplicando o algoritmo de Yen no grafo da Figura 50 está na Tabela 10.

Tabela 10 - Resultado algoritmo Yen

Caminho	Custo	Comprimento
(s, 4, t)	10	2
(s, 1, 3, t)	12	3
(s, 1, 2, 3, t)	12	4
(s, 4, 1, 3, t)	13	4
(s, 4, 1, 2, 3, t)	13	5
(s, 4, 2, 3, t)	13	4
(s, 4, 2, 1, 3, t)	15	5
(s, 2, 3, t)	16	3
(s, 1, 4, t)	17	3
(s, 2, 1, 3, t)	18	4
(s, 1, 2, 4, t)	19	4
(s, 1, 4, 2, 3, t)	20	5
(s, 1, 3, 2, 4, t)	23	5
(s, 2, 4, t)	23	3
(s, 2, 1, 4, t)	23	4
(s, 2, 4, 1, 3, t)	26	5
(s, 2, 3, 1, 4, t)	27	5

O caminho no grafo com menor custo possui comprimento 2 e custo 10 (s, 4, t) e o de maior custo possui comprimento 5 e custo 27 (s, 2, 3, 1, 4, t). Entretanto, observando a Tabela 10 vê-se que existem caminhos com comprimento menor e custo maior, como é o caso do caminho (s, 2, 3, t), e também existem caminhos com comprimento maior e custo menor, como o caminho (s, 4, 1, 2, 3, t). Isso ocorre porque o algoritmo se baseia no valor das arestas para listar os caminhos de mais baixo custo.

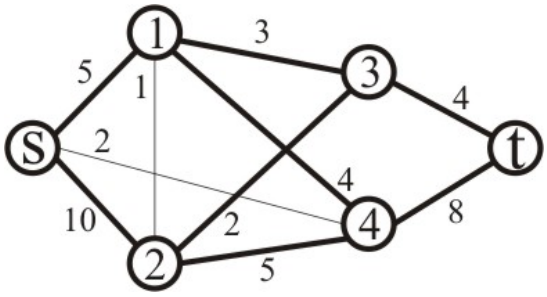


Figura 50 - *k*-caminhos mais curtos.

Porém, para calcular a métrica de Katz é necessário encontrar todos os caminhos de comprimento fixo entre dois nós, e não apenas os caminhos mais curtos, como é a proposta do algoritmo de Yen. Como mostra a equação da métrica de Katz é necessário encontrar todos os caminhos de comprimento 1, seguidos de todos os caminhos de comprimento 2, e assim por diante. Para atender a métrica de Katz os caminhos devem ser agrupados por seus comprimentos e não pelos custos.

Tabela 11 - Matriz de Adjacência

	s	**1**	**2**	**3**	**4**	**t**
s	-	1	1	0	1	0
1	1	-	1	1	1	0
2	1	1	-	1	1	0
3	0	1	1	-	0	1
4	1	1	1	0	-	1
t	0	0	0	1	1	-

Para solucionar essa questão foi utilizada a matriz de adjacência que representa a rede social científica multi-relacional modelada neste trabalho. Para ilustrar o funcionamento do algoritmo de Yen aplicando a matriz de adjacência do grafo, foi construída a matriz de adjacência do grafo da Figura 50. A Tabela 11 representa essa matriz. A matriz de adjacência é preenchida apenas por uns e zeros indicando apenas se dois elementos são vizinhos ou não.

Utilizar a matriz de adjacência na representação do grafo social significa que todas as arestas do grafo possuem peso igual a 1. Como as arestas possuem peso unitário, o custo do caminho será igual ao seu comprimento. Aplicando o algoritmo de Yen sobre

a matriz de adjacência do grafo da Figura 50 é obtida a lista de caminhos mais curtos, representada na Tabela 12.

Após obter uma lista de caminhos, semelhante a da tabela anterior, para a rede social científica modelada neste trabalho, a métrica composta, definida na equação (24), pode ser facilmente calculada. O último passo antes do cálculo da métrica é definir o custo de cada caminho obtido.

Como a proposta deste trabalho gira em torno da análise do fluxo máximo na rede social científica, o custo dos caminhos obtidos será definido pelo fluxo máximo de cada um deles. Para calcular o fluxo máximo todos os conceitos previamente definidos na etapa de modelagem da rede social científica deverão ser utilizados, ou seja, devem ser consideradas as perdas de informação nos caminhos mais longos.

Tabela 12 - Resultado algoritmo de Yen para matriz de adjacência

Caminho	Custo	Comprimento
(s, 4, t)	2	2
(s, 1, 3, t)	3	3
(s, 1, 4, t)	3	3
(s, 2, 4, t)	3	3
(s, 2, 3, t)	3	3
(s, 4, 1, 3, t)	4	4
(s, 2, 1, 4, t)	4	4
(s, 1, 2, 3, t)	4	4
(s, 4, 2, 3, t)	4	4
(s, 2, 1, 3, t)	4	4
(s, 1, 2, 4, t)	4	4
(s, 1, 3, 2, 4, t)	5	5
(s, 4, 1, 2, 3, t)	5	5
(s, 4, 2, 1, 3, t)	5	5
(s, 2, 3, 1, 4, t)	5	5
(s, 2, 4, 1, 3, t)	5	5
(s, 1, 4, 2, 3, t)	5	5

Para ilustrar esse processo, foram considerados apenas os caminhos de tamanho 3 encontrados pelo algoritmo de Yen para o grafo da Figura 50. A Tabela 13 apresenta os caminhos detalhando o peso de cada aresta e qual é o fluxo máximo de cada um deles. A última coluna representa o fluxo máximo após aplicar a resistência ao fluxo conforme definido no capítulo 4.

Tabela 13 - Análise dos caminhos de tamanho 3

Caminhos	Arestas do caminho	Peso da Aresta	Fluxo Máximo	Fluxo Máximo com Perda
(s, 1, 3, t)	s → 1	5	3	2,94
	1→ 3	3		
	3 → t	4		
(s, 1, 4, t)	s → 1	5	4	3,92
	1 → 4	4		
	4 → t	8		
(s, 2, 4, t)	s → 2	10	5	4,90
	2 → 4	5		
	4 → t	8		
(s, 2, 3, t)	s → 2	10	2	1,96
	2 → 3	2		
	3 → t	4		
TOTAL	----	----	14	13,72

Após calcular o fluxo máximo de cada caminho os mesmos são somados e multiplicados pelo parâmetro β conforme a equação (24). Os valores de β devem estar entre zero e um para que os caminhos de comprimento muito longo tenham uma influência menor no valor final da métrica de Katz e, conseqüentemente, na métrica composta.

Como na definição da equação da métrica de Katz o parâmetro β está elevado ao tamanho do caminho e $0 < \beta < 1$, a influência que os caminhos terão sobre o valor final da métrica depende diretamente do valor de β.

Se o valor de β for muito próximo de zero, a métrica de Katz apresenta resultados semelhantes à métrica de vizinhos comuns, já que os caminhos de comprimento maior que três terão uma influência insignificante no valor final da métrica. Por outro lado, se β for muito próximo de 1 os caminhos longos irão influenciar muito no valor da métrica podendo gerar distorções nas previsões e sugestões de novos relacionamentos.

O ajuste desses dois parâmetros não é trivial. Para encontrar a melhor definição para eles foram feitos vários testes utilizando os dados históricos da rede social científica modelada neste trabalho.

Como dito anteriormente, a métrica foi calculada considerando caminhos de comprimento máximo 6. Os experimentos e a definição dos valores dos parâmetros da métrica de Katz serão apresentados junto com a análise dos resultados obtidos no módulo de sugestão de relacionamentos no próximo capítulo.

Capítulo 8 – Estudo de Caso: Sugestão de Relacionamentos

Neste capítulo será apresentada a segunda etapa do estudo de caso deste trabalho, que é a sugestão de relacionamentos. Para tal, será analisada a evolução temporal da rede sociais científica avaliando o histórico de mais de 10 anos de relacionamentos a fim de obter uma boa qualidade na sugestão de novos relacionamentos para os pesquisadores.

Todos os resultados que serão apresentados neste capítulo foram obtidos através dos conceitos das métricas descritas no Capítulo 7. Além disso, será feita a análise da influência que a variação dos parâmetros da métrica Katz possui sobre as previsões de novos relacionamentos.

8.1 Introdução

Ao analisar o cotidiano de uma instituição científica, percebe-se que existem muitas outras informações úteis para os pesquisadores. Além dos grupos e das análises da rede social, que foram apresentados no capítulo 6, é interessante um pesquisador saber quem não está diretamente relacionado a ele, mas que, de alguma maneira, está "próximo" a ele.

Por exemplo, no caso de um pesquisador ir a uma conferência, quais dos pesquisadores presentes lá ele tem mais afinidade? Qual pesquisador poderia ser chamado para a banca de defesa de tese de um aluno? Perguntas como essas podem ser resolvidas com o módulo de sugestão de relacionamentos, ajudando assim no dia-a-dia dos pesquisadores.

O objetivo principal desse módulo não é prever novos relacionamentos, mas sim sugerir relacionamentos que sejam úteis para os pesquisadores. Entretanto, uma maneira de validar se o módulo está sugerindo relacionamentos com qualidade é através da análise de previsão de ligações.

Assim, neste capítulo será feita a análise das previsões de relacionamentos para cada ano do conjunto de dados. A idéia é prever os relacionamentos novos do ano analisando as informações do passado da rede social científica. Por exemplo, serão

previstos os relacionamentos para o ano de 2009 baseando-se em todos os relacionamentos da rede social que existiam antes de 2009. Como já são conhecidos os novos relacionamentos do ano de 2009 será possível avaliar o número de acertos obtidos pelo algoritmo de sugestão de relacionamentos.

Os experimentos foram desenvolvidos sempre com a preocupação de comparar os resultados obtidos pela métrica de Katz com os resultados obtidos pela métrica composta. Essa comparação foi feita para validar a métrica composta, pois, segundo a literatura (Liben-Nowell and Kleinberg, 2007), a métrica de Katz é a medida que produz os melhores resultados para os problemas de previsão de relacionamentos.

A métrica proposta neste trabalho é baseada na busca por todos os caminhos de tamanho máximo 6 (l = 1, ..., 6) existente entre todos os pares de pesquisadores da rede social. A configuração do tamanho máximo do caminho está na próxima seção. O valor associado a cada caminho foi calculado pelo algoritmo de fluxo máximo considerando o peso das relações entre os pesquisadores. Assim, a métrica irá refletir o poder de comunicação entre uma dupla de pesquisadores.

Conforme apresentado no capítulo de modelagem da rede social científica foram utilizadas três funções de penalização com base no ano do relacionamento para avaliar a que melhor reproduzia o comportamento da rede social. As funções que apresentaram resultados melhores para o problema de agrupamento foram a exponencial e a sigmóide. Assim, será feita neste capítulo uma comparação entre os resultados obtidos para a previsão de relacionamentos utilizando essas duas funções.

Os resultados obtidos com a análise da previsão de relacionamentos indicaram qual o melhor conjunto de parâmetros que deve ser utilizado no módulo de sugestão de relacionamentos. Esse módulo também está disponível na ferramenta de visualização da rede social, onde o pesquisador pode saber quais são as suas melhores sugestões de relacionamentos. Os detalhes sobre a visualização e as funcionalidades desse módulo estão no capítulo 10.

8.2 Medidas de Avaliação

A avaliação das técnicas de previsão de novas conexões em redes sociais é feita através de medidas estatísticas que medem a quantidade de acertos baseando-se no conjunto de dados.

Para avaliar a qualidade das recomendações foram utilizadas medidas padrões de avaliação, tais como exatidão, do inglês *accuracy*; precisão, do inglês *precision*; e *recall* (Huang, Chen and Zeng, 2001).

Para auxiliar a análise dessas medidas de avaliação foi construída a Tabela 14. Nessa tabela estão todos os casos possíveis de erros e acertos para o problema de previsão de relacionamentos. Os '*Positivos Verdadeiros*' representam todos os relacionamentos que foram previstos de existirem no ano seguinte e que realmente existiram. Da mesma forma, os '*Negativos Verdadeiros*' são os relacionamentos que foram previstos de não existirem no ano seguinte e que não existiram. Já os '*Positivos Falsos*' são os relacionamentos que foram previstos de ocorrerem no ano seguinte erroneamente. Os '*Negativos Falsos*' são os que foram previstos de não ocorrerem, mas que surgiram como novos relacionamentos no ano seguinte.

Tabela 14 - Possíveis resultados para a previsão de relacionamentos

Com Relacionamentos Novos	PV (Positivos Verdadeiros)	PF (Positivos Falsos)
Sem Relacionamentos Novos	NF (Negativos Falsos)	NV (Negativos Verdadeiros)

A *exatidão* é a proporção de resultados positivos em toda a população. Na avaliação da exatidão são levados em consideração tanto os acertos dos novos relacionamentos, quanto os acertos dos relacionamentos que não seriam criados. Com base na tabela definida anteriormente a fórmula da exatidão é dada por

$$exatidão = \frac{PV + NV}{PV + PF + NF + NV}, \tag{27}$$

onde *PV* e *NV* definem os casos de acertos, *NF* e *PF* definem os casos de erros.

A *precisão* é a quantidade de acertos obtidos para o número de tentativas, ou seja, é a taxa de acerto para o número de "chutes". Assim, a equação da precisão pode ser definida por

$$precisão = \frac{PV}{PV + PF}, \tag{28}$$

onde, como definido anteriormente, *PV* é o total de previstos positivos corretos e *PF* é o total de previsto positivo errado.

A diferença entre exatidão e precisão é que a exatidão é a proporção de todos os resultados corretos (positivos verdadeiros e negativos verdadeiros) por toda a população de previsões. Já a precisão é a proporção de todos os positivos verdadeiros por todos os resultados positivos.

Já o *recall* é a quantidade de acertos para o número de relacionamentos novos no ano. A sua equação é definida como segue

$$recall = \frac{PV}{PV + NF}, \tag{29}$$

onde a soma de PV e NF representa o total de relacionamentos novos do ano que está sendo feita a previsão.

Outra medida adotada é a *F-measure*, que é uma medida que combina a precisão e a cobertura (*recall*). Essa medida pode ser interpretada como uma média ponderada entre precisão e cobertura, que atinge o melhor valor em 1 e o pior em 0. A medida *F-measure* é dada por:

$$F = 2\frac{precisão * recall}{precisão + recall} \tag{30}$$

8.3 Configuração dos Parâmetros

A configuração dos parâmetros da métrica de Katz e, conseqüentemente, da métrica composta, é uma etapa delicada do módulo de sugestão de relacionamentos. Para facilitar a análise dos parâmetros, a equação dessa métrica está novamente definida na equação (31).

$$\textit{Métrica Composta} = \frac{\sum_{l=1}^{\infty} \beta^{l}.\#(\text{caminhos}_{v_i,v_j}^{<l>})}{(\#\Gamma(v_i) + \#\Gamma(v_j)) - \#\{\Gamma(v_i) \cap \Gamma(v_j)\} + 1} \quad (31)$$

Existem dois parâmetros que devem ser configurados na métrica composta: o tamanho máximo dos caminhos entre os pesquisadores e o parâmetro β. Por definição, o tamanho do caminho entre os pesquisadores não tem um limite superior e, assim, todos os caminhos possíveis entre dois elementos da rede social deveriam ser considerados no cálculo da métrica. Entretanto esse cálculo é inviável, considerando que as redes sociais possuem milhares de caminhos alternativos entre pares de nós. Nesta seção serão analisados os resultados obtidos pela variação desse parâmetro.

Outro parâmetro a ser configurado é o β. Observando a equação (31) é fácil ver que o objetivo do parâmetro β é dar um peso maior ou menor segundo o tamanho do caminho. O desafio é definir um valor para β que produza bons resultados na previsão de novos relacionamentos.

8.3.1 Parâmetro β

A configuração do parâmetro β é bastante sensível, pois, como dito anteriormente, caso seja adotado um valor muito baixo a métrica de Katz será aproximada à métrica de vizinhos comuns. Por outro lado, caso seja adotado um valor muito alto a influência dos caminhos longos será muito forte e, conseqüentemente, isso pode prejudicar a previsão de novos relacionamentos.

Alguns experimentos consideraram $\beta = 0,001$ (Acar, Dunlavy and Kolda, 2009), outros fizeram testes para $\beta = 0,05$, $\beta = 0,005$ ou $\beta = 0,0005$ (Huang, Li and Chen, 2005; Huang and Lin, 2009; Liben-Nowell and Kleinberg, 2007), e concluíram que as melhores previsões ocorreram para $\beta = 0,005$. Assim, com base nessas pesquisas observa-se que os melhores valores para β estão na ordem de 10^{-3}.

Entretanto, para avaliar os melhores valores de β para a previsão de relacionamentos na rede social científica multi-relacional foram feitos experimentos utilizando $\beta = 0,5$, $\beta = 0,05$, $\beta = 0,005$ e $\beta = 0,0005$. O objetivo é avaliar o

comportamento da métrica composta e da métrica de Katz com a variação dos valores do parâmetro beta.

Para avaliar as métricas com cada valor do parâmetro beta definido anteriormente foram feitos pontos de corte na base de dados nos anos de 2009 e de 2010. Assim, foi possível analisar a rede social até o ano de 2008 e prever os relacionamentos de 2009 e, da mesma maneira, analisar os relacionamentos da rede social até o ano de 2009 e prever os relacionamentos de 2010.

Tabela 15 - Análise da previsão de relacionamentos para 2009

Ano	Tamanho Caminho	Beta	Métrica Katz			Métrica Composta		
			Precisão	***Recall***	***F***	**Precisão**	***Recall***	***F***
2009	3	$\beta = 0,5$	0,0168	0,5111	0,0325	0,0175	0,5333	0,0339
		$\beta = 0,05$	0,0168	0,5111	0,0325	0,0175	0,5333	0,0339
		$\beta = 0,005$	0,0168	0,5111	0,0325	0,0175	0,5333	0,0339
		$\beta = 0,0005$	0,0168	0,5111	0,0325	0,0175	0,5333	0,0339
	4	$\beta = 0,5$	**0,0190**	**0,5778**	**0,0368**	**0,0197**	**0,6000**	**0,0381**
		$\beta = 0,05$	0,0182	0,5556	0,0352	0,0190	0,5778	0,0368
		$\beta = 0,005$	0,0168	0,5111	0,0325	0,0175	0,5333	0,0339
		$\beta = 0,0005$	0,0168	0,5111	0,0325	0,0175	0,5333	0,0339
	5	$\beta = 0,5$	0,0161	0,4889	**0,0312**	**0,0197**	**0,6000**	**0,0381**
		$\beta = 0,05$	0,0182	0,5556	0,0352	0,0190	0,5778	0,0368
		$\beta = 0,005$	0,0168	0,5111	0,0325	0,0175	0,5333	0,0339
		$\beta = 0,0005$	0,0168	0,5111	0,0325	0,0175	0,5333	0,0339
	6	$\beta = 0,5$	0,0153	0,4667	**0,0296**	**0,0197**	**0,6000**	**0,0381**
		$\beta = 0,05$	0,0182	0,5556	0,0352	0,0190	0,5778	0,0368
		$\beta = 0,005$	0,0168	0,5111	0,0325	0,0175	0,5333	0,0339
		$\beta = 0,0005$	0,0168	0,5111	0,0325	0,0175	0,5333	0,0339

Os resultados dessa primeira análise para a variação dos parâmetros da métrica de Katz estão na Tabela 15 e na Tabela 16. A Tabela 15 exibe os resultados das previsões de novos relacionamentos para o ano de 2009. Já na Tabela 16 são apresentados os resultados das previsões de relacionamentos para o ano de 2010. Nessas tabelas estão em negrito os melhores resultados obtidos pelas métricas. Para produzir esses resultados todos os dados da base de dados foram utilizados.

É importante observar que para valores de beta pequenos ($\beta = 0,005$ e $\beta = 0,0005$) os resultados são iguais independentemente do tamanho do caminho. Isso significa que o peso utilizado anula a influência que os caminhos de tamanho maiores poderiam ter na métrica. Esse fato foi observado nos dois anos que foram avaliados.

Observe que a métrica composta, proposta neste trabalho, apresenta resultados melhores que a métrica de Katz, que é considerada a melhor métrica de previsão de relacionamentos (Acar, Dunlavy and Kolda, 2009; Huang, Li and Chen, 2005; Huang and Lin, 2009; Liben-Nowell and Kleinberg, 2007), para o ano de 2009.

Além disso, a métrica composta apresenta uma estabilidade melhor com relação à variação dos parâmetros. Essa estabilidade pode ser verificada tanto na Tabela 15 quanto na Figura 51. Nessa figura estão os gráficos que representam a medida de *recall* das previsões do ano de 2009, obtidos através da métrica de Katz e da métrica composta.

Tabela 16 - Análise da previsão de relacionamentos para 2010

Ano	Tamanho Caminho	Beta	Métrica Katz			Métrica Composta		
			Precisão	***Recall***	***F***	**Precisão**	***Recall***	***F***
2010	3, 4, 5, 6	$\beta = 0,05$	**0,0015**	**0,5000**	**0,0030**	**0,0015**	**0,5000**	**0,0030**
		$\beta = 0,005$	**0,0015**	**0,5000**	**0,0030**	**0,0015**	**0,5000**	**0,0030**
		$\beta = 0,0005$	**0,0015**	**0,5000**	**0,0030**	**0,0015**	**0,5000**	**0,0030**
	3	$\beta = 0,5$	**0,0015**	**0,5000**	**0,0030**	**0,0015**	**0,5000**	**0,0030**
	4	$\beta = 0,5$	**0,0015**	**0,5000**	**0,0030**	**0,0015**	**0,5000**	**0,0030**
	5	$\beta = 0,5$	0,0007	0,2500	0,0014	0,0007	0,2500	0,0014
	6	$\beta = 0,5$	0,0000	0,0000	---	0,0007	0,2500	0,0014

Como pode ser visto nos gráficos da Figura 51, quando é adotado $\beta = 0,5$, a métrica de Katz apresenta uma previsão muito boa com caminhos de tamanho máximo 4, mas piora consideravelmente os resultados quando o tamanho máximo do caminho passa para 5 e 6. Por outro lado, a métrica composta apresenta resultados muito bons para $\beta = 0,05$ e $\beta = 0,5$, mesmo quando o tamanho máximo dos caminhos aumenta.

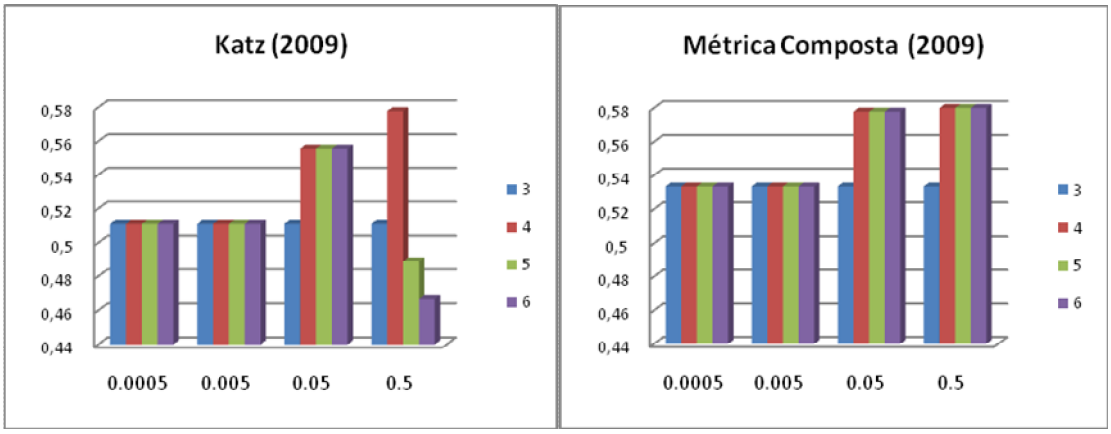


Figura 51 - Análise gráfica das previsões de relacionamentos para o ano de 2009

Analisando a Tabela 16 com mais detalhe observa-se que a métrica composta e a métrica de Katz apresentaram os mesmos resultados para as previsões de relacionamentos do ano de 2010. A métrica composta apresenta resultado melhor que a métrica de Katz quando o tamanho máximo do caminho é igual a 6. Nessa situação a métrica de Katz não foi capaz de acertar nenhuma previsão corretamente.

A Figura 52 representa graficamente a coluna *recall* da Tabela 16. É fácil ver nessa figura que para o ano de 2010 as duas métricas foram estáveis com $\beta = 0,05$, $\beta = 0,005$ e $\beta = 0,0005$. Para $\beta = 0,5$ as duas métricas apresentaram uma piora nos resultados para caminhos de tamanho máximo 5 e 6.

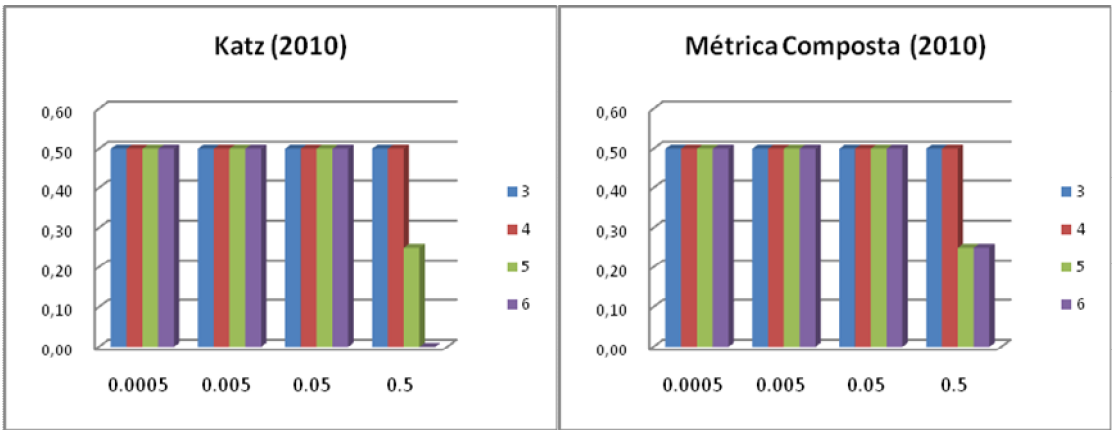


Figura 52 - Análise gráfica das previsões de relacionamentos para o ano de 2010

Observando os resultados obtidos anteriormente com a variação do parâmetro beta concluiu-se que os melhores resultados foram obtidos com beta mais próximo de 1. Com o intuito de avaliar o comportamento da métrica de Katz utilizando outra função peso com base no tamanho do caminho foi elaborada uma nova função α definida na equação (32). A curva descrita por essa função está ilustrada na Figura 53.

$$\alpha = 100l^{-\frac{1}{2}l} \tag{32}$$

Assim como na equação (31) a variável l da equação (32) é o tamanho do caminho. O objetivo dessa função é favorecer fortemente os relacionamentos diretos e os caminhos de comprimento menor e, por outro lado, penalizar os caminhos mais longos, mas permitindo que eles tenham uma influência na métrica de Katz.

Embora o objetivo das duas funções seja o mesmo, a função β^l decresce muito rapidamente e, conseqüentemente, a penalização sobre os caminhos mais longos é muito forte. Assim, para que os caminhos de tamanhos mais longos tenham influência na métrica é necessário que exista uma grande quantidade de caminhos longos.

O objetivo da função alfa é permitir que os caminhos mais longos também tenham alguma influência no valor final das métricas. Os resultados das previsões dos relacionamentos para os anos de 2009 e 2010 utilizando a função alfa estão na Tabela 17.

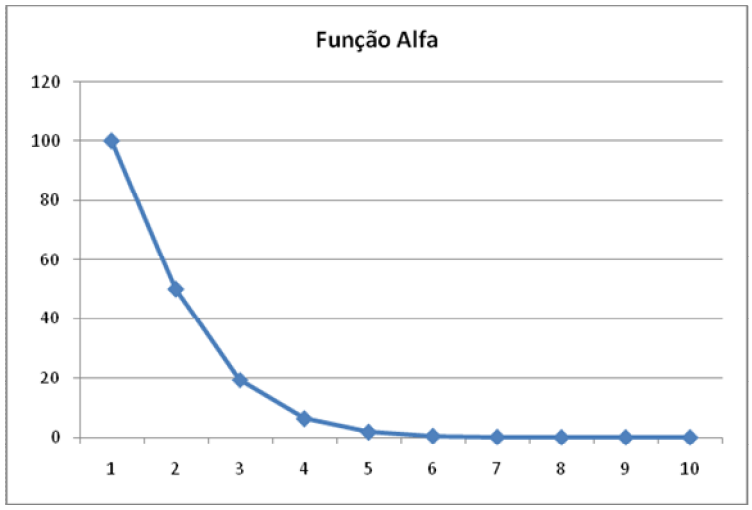


Figura 53 - Definição da função peso alfa.

Os resultados obtidos mostram mais uma vez que a métrica composta sempre produz resultados melhores ou iguais que a métrica de Katz. Comparando a Tabela 17 com a Tabela 15 e a Tabela 16 é possível ver que as métricas apresentaram resultados melhores utilizando a função alfa.

Tabela 17 - Análise da previsão de relacionamentos com o uso da função alfa

Ano	Tamanho Caminho	Métrica Katz			Métrica Composta		
		Precisão	*Recall*	*F*	Precisão	*Recall*	*F*
2009	3	0,0168	0,5111	0,0325	0,0175	0,5333	0,0339
	4	**0,0190**	**0,5778**	**0,0368**	**0,0204**	**0,6222**	**0,0395**
	5	0,0175	0,5333	0,0339	0,0197	0,6000	0,0381
	6	0,0160	0,4889	0,0310	0,0197	0,6000	0,0381
2010	3	**0,0014**	**0,5000**	**0,0028**	**0,0014**	**0,5000**	**0,0028**
	4	**0,0014**	**0,5000**	**0,0028**	**0,0014**	**0,5000**	**0,0028**
	5	0,0007	0,2500	0,0014	**0,0014**	**0,5000**	**0,0028**
	6	0,0007	0,2500	0,0014	0,0007	0,2500	0,0014

No ano de 2009 houve um aumento na precisão e no recall para todos os tamanhos dos caminhos. No ano de 2010, embora não tenha ocorrido aumento nas medidas de avaliação da métrica composta, o uso da função alfa permitiu que a métrica de Katz acertasse 25% (medida de *recall*) das previsões com o tamanho do caminho igual a 6.

A Figura 54 mostra os gráficos que representam a medida de *recall* da métrica de Katz e da métrica composta com o uso da função alfa e com o uso da função beta, para $\beta = 0,5$. Analisando os gráficos é possível ver que a função alfa apresenta resultados melhores ou iguais que a função beta tanto para a métrica de Katz quanto para a métrica composta.

Para consolidar a análise do parâmetro de penalização com base no tamanho do caminho foi feita uma comparação final entre a métrica composta e a métrica de Katz. O objetivo é comparar a métrica composta usando a função alfa com a métrica de Katz usando a função beta. Para essa comparação foi utilizado $\beta = 0,5$, já que esse valor produziu os melhores resultados para ambas as métricas. Os gráficos com essa comparação estão na Figura 55 e eles mostram que, como dito anteriormente, as previsões dos relacionamentos feitas pela métrica composta são sempre melhores ou iguais que a métrica de Katz.

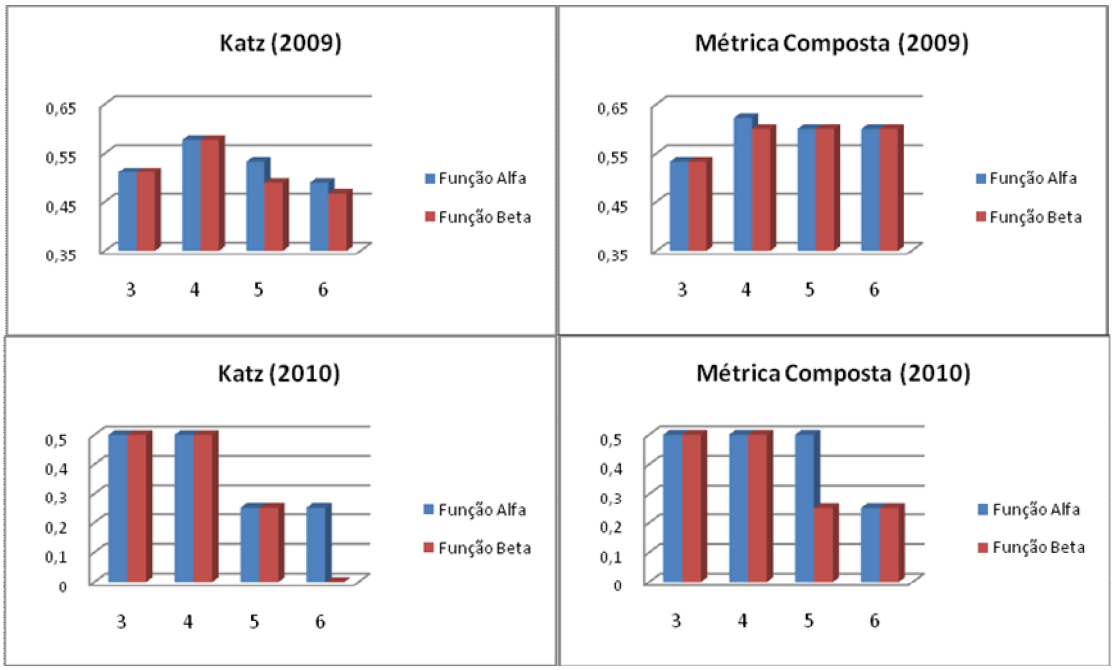


Figura 54 - Análise das funções alfa e beta

Comparando os resultados obtidos com a variação do parâmetro beta e com a função alfa vê-se que a nova função peso teve resultados muito melhores que os valores propostos na literatura. Essa melhora nas previsões de novos relacionamentos indica que os relacionamentos indiretos nas redes sociais científicas possuem boa influência no

nascimento de novos relacionamentos, ou seja, o fluxo de informação entre pesquisadores que estão a dois ou três níveis de distância pode ajudar para que um novo relacionamento seja criado.

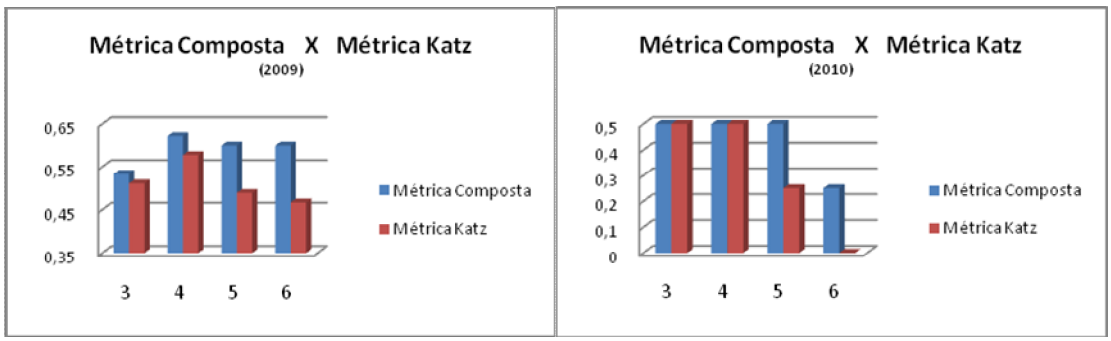


Figura 55 - Comparação entre a Métrica Composta e a Métrica de Katz para os anos de 2009 e 2010

8.3.2 Tamanho do Caminho

A definição do tamanho máximo dos caminhos entre os pesquisadores levou em consideração basicamente a influência que os caminhos teriam sobre o valor final das métricas e as análises feitas na seção anterior.

A Tabela 18 apresenta o total de caminhos necessários de um determinado tamanho para que eles tenham a mesma influência que um caminho de tamanho 1 para $\beta = 0{,}005$.

Tabela 18 - Análise do peso dos tamanhos dos caminhos

Tamanho Caminho (*l*)	β^l	Nº de Caminhos
1	0,005	1
2	0,000025	200
3	0,000000125	40.000
4	0,000000000625	8.000.000
5	0,000000000003125	1.600.000.000
6	0,000000000000015625	320.000.000.000

Nessa tabela é possível ver a relação da quantidade de caminhos de tamanhos 2, 3, 4, 5 e 6 necessária para que esses caminhos tenham o mesmo peso dos caminhos de tamanho 1. Por exemplo, 40.000 caminhos de tamanho 3 possuem o mesmo peso que 1

caminho de tamanho 1. Analisando essa tabela é possível compreender o motivo pelo qual a métrica de Katz apresenta valores constantes com tamanhos de caminhos grandes e beta com valores baixos.

A análise da Tabela 15, da Tabela 16 e da Tabela 18 mostra que os caminhos maiores possuem pequena ou nenhuma influência no valor final das métricas quando o valor de beta é muito baixo. Por outro lado, os resultados da previsão dos relacionamentos pioram quando o tamanho do caminho aumenta e a função de peso permite que os caminhos maiores influenciem nos resultados da métrica.

Além da piora nos resultados ou da pouca influência que os caminhos longos possuem no valor final da métrica, o custo computacional para calcular todos os caminhos entre dois elementos é alto. O cálculo da métrica de Katz e da métrica composta é inviável para caminhos muito grandes.

Avaliando os últimos resultados apresentados na Figura 55 vê-se que o tamanho de caminho máximo 4 produziu os melhores resultados para a métrica composta. No decorrer deste capítulo todas as análises serão feitas baseando-se no cálculo das métricas com tamanho máximo 4.

8.4 Função de Penalização: Exponencial X Sigmóide

No capítulo 4, onde a rede social científica multi-relacional foi modelada, foram feitos testes de funções de penalização com base no ano do relacionamento. O objetivo foi identificar a função que melhor representasse o comportamento evolutivo da rede social científica. Nos experimentos feitos nesse capítulo a função exponencial apresentou resultado melhor que a função sigmóide e a função potência. Entretanto, embora a função exponencial tenha sido melhor, a função sigmóide também apresentou resultados satisfatórios.

A função sigmóide busca reproduzir um comportamento específico da rede social científica que é o caso dos pesquisadores que publicam com uma freqüência de 2 em 2 anos ou 3 em 3 anos. Nesta seção serão feitas as análises dos resultados da previsão dos relacionamentos utilizando a função exponencial e a função sigmóide.

O objetivo desse estudo é verificar qual função de redução do peso dos relacionamentos produz melhor resultado para o módulo de sugestão de relacionamentos, mesmo o resultado do agrupamento sendo melhor com o uso da função exponencial do que com a função sigmóide.

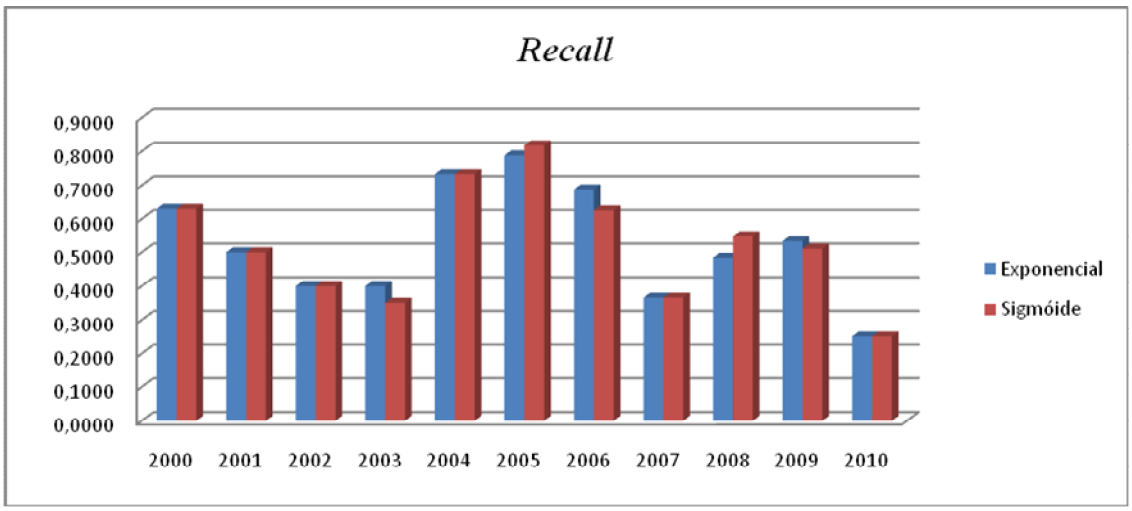


Figura 56 - Análise da previsão de relacionamento (Exponencial X Sigmóide)

Analisando o gráfico da Figura 56 foi observado que ambas as funções apresentaram resultados iguais nos anos 2000, 2001, 2002, 2004, 2007 e 2010. A função exponencial apresentou resultados melhores nos anos 2003, 2006 e 2009. Já a função sigmóide apresentou previsões melhores nos anos 2005 e 2008.

Mesmo a função sigmóide apresentando resultados melhores em alguns anos, a função exponencial foi melhor em um número de anos maior. Assim, como a função exponencial também apresentou resultados melhores no módulo de agrupamento, optou-se por utilizá-la no módulo de sugestão de relacionamentos.

Entretanto, o uso da função sigmóide deve ser considerado em estudos futuros, pois com o passar dos anos essa função pode se tornar mais adequada que a função exponencial.

8.5 Resultados

As sugestões de relacionamentos são feitas baseando-se em uma lista formada pelos valores da métrica composta. Uma lista é construída com todos os relacionamentos

possíveis de serem estabelecidos na rede social científica. Cada um desses relacionamentos possíveis recebe um valor que é obtido através da métrica composta. Essa lista é ordena em ordem decrescente pelo valor da métrica. Assim, os primeiros resultados indicam a maior probabilidade de novos relacionamentos.

Para avaliar o desempenho das recomendações, foi adotado o processo de recomendação topo-N. Nesse processo são selecionados N% dos primeiros elementos da lista, essa seleção é feita dessa forma para que sejam filtradas as melhores sugestões. Neste trabalho foram selecionados 10% de todos os novos relacionamentos possíveis na rede social científica.

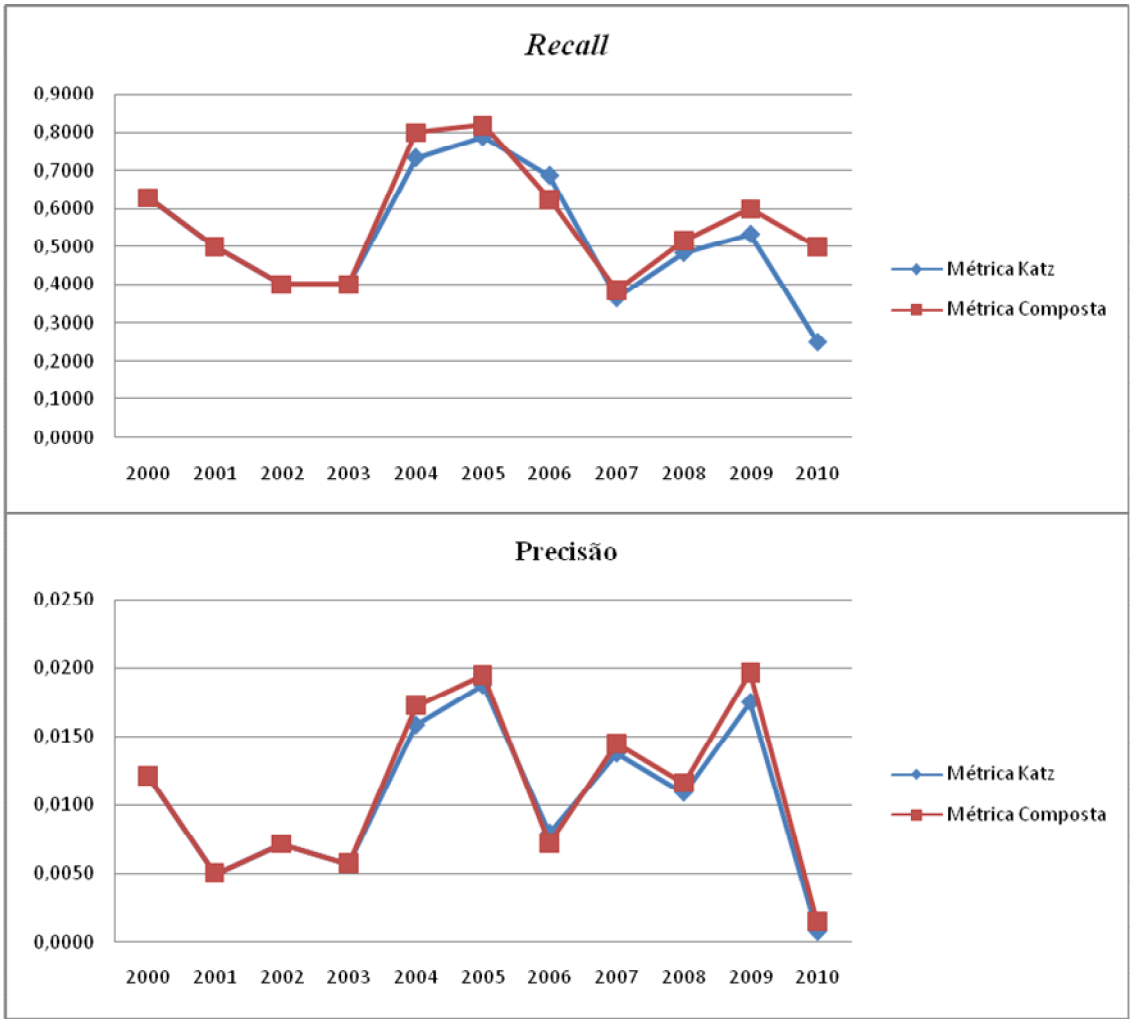


Figura 57 - Gráficos de Precisão e *Recall*.

Assim como nas análises feitas anteriormente, para avaliar a qualidade dos resultados do módulo de sugestão de relacionamentos foram utilizadas medidas padrões de avaliação de qualidade, tais como precisão e *recall* (Huang, Chen and Zeng, 2001). Todos os resultados apresentados nessa seção utilizaram a função alfa e tamanho de caminho máximo igual a 4 no cálculo da métrica de Katz e da métrica composta.

Os gráficos que ilustram essas medidas de qualidade são apresentados na Figura 57. Pode ser observado que em todos os anos, a Métrica Composta apresenta precisão e *recall* igual ou melhor que a métrica de Katz, com exceção do ano 2006 onde a métrica de Katz é melhor que a métrica proposta neste trabalho.

É importante observar que as previsões de novos relacionamentos das duas métricas para os anos 2000, 2001, 2002 e 2003 foram iguais. Com o passar dos anos a quantidade de relacionamentos aumenta consideravelmente e, com esse aumento, a influência dos relacionamentos de caminhos maiores também é maior, melhorando os resultados da métrica composta.

A Tabela 19 mostra os dados utilizados para construir os gráficos da Figura 57. Nessa tabela estão em negrito todos os resultados que foram diferentes quando comparados entre as duas métricas. As células em azul apresentam os valores quando a métrica composta foi melhor que a métrica de Katz, já as células em verde mostram os valores nos quais a métrica de Katz foi melhor que a métrica composta.

Com a ajuda desses gráficos e da tabela, observa-se que a métrica proposta para sugerir novos relacionamentos produziu bons resultados. A métrica proposta neste trabalho é composta, conforme apresentado no capítulo 7, por três métricas que possuem duas abordagens distintas no estudo de previsão de relacionamentos. O grau dos pesquisadores e o número de vizinhos comuns representam a abordagem que é baseada na vizinhança dos elementos. Já a métrica de Katz é uma abordagem baseada no caminho entre os pesquisadores.

Os resultados apresentados neste capítulo mostram que a união dessas duas abordagens utilizadas no estudo de previsão de relacionamentos pode melhorar significativamente o desempenho da previsão de relacionamentos em redes sociais científicas com múltiplos tipos de relacionamentos. Em geral, esses resultados indicam

que a previsão/sugestão de relacionamentos pode ser uma direção importante para melhorar o fluxo de conhecimento na rede científica social.

Tabela 19 - Valores encontrados para as Métricas de Katz e a Métrica Composta.

Ano	Métrica	Precisão	*Recall*	*F*
2000	Katz	0,0121	0,6296	0,0237
	Composta	0,0121	0,6296	0,0237
2001	Katz	0,0050	0,5000	0,0099
	Composta	0,0050	0,5000	0,0099
2002	Katz	0,0072	0,4000	0,0141
	Composta	0,0072	0,4000	0,0141
2003	Katz	0,0057	0,4000	0,0112
	Composta	0,0057	0,4000	0,0112
2004	Katz	**0,0158**	**0,7333**	**0,0309**
	Composta	**0,0173**	**0,8000**	**0,0339**
2005	Katz	**0,0188**	**0,7879**	**0,0367**
	Composta	**0,0195**	**0,8182**	**0,0381**
2006	Katz	**0,0080**	**0,6875**	**0,0158**
	Composta	**0,0072**	**0,6250**	**0,0142**
2007	Katz	**0,0138**	**0,3654**	**0,0266**
	Composta	**0,0145**	**0,3846**	**0,0279**
2008	Katz	**0,0109**	**0,4839**	**0,0213**
	Composta	**0,0116**	**0,5161**	**0,0227**
2009	Katz	**0,0175**	**0,5333**	**0,0339**
	Composta	**0,0197**	**0,6000**	**0,0381**
2010	Katz	**0,0007**	**0,2500**	**0,0014**
	Composta	**0,0015**	**0,5000**	**0,0030**

8.6 Validação

Diferentemente da análise dos resultados do módulo de agrupamento, o módulo de sugestão de relacionamentos não precisa ser validado por meio de questionários. A validação desse módulo foi feita através da análise dos anos para os quais os relacionamentos já são conhecidos.

Sabendo de antemão quais são os relacionamentos novos do ano seguinte foi possível saber a taxa de acerto da nova métrica proposta e validar os seus resultados. Os dados foram validados através da análise entre as métricas de Katz e a métrica composta, que é a métrica proposta neste trabalho.

A avaliação dos resultados da métrica proposta neste trabalho foi feita nas seções anteriores, onde foram feitas diversas previsões para comparar os resultados obtidos pela métrica composta e pela métrica de Katz. Nessa etapa foram analisados os anos de 2000 a 2010 e a métrica composta se mostrou igual ou melhor que a métrica de Katz na previsão de novos relacionamentos em todos os anos, exceto no ano de 2006.

Capítulo 9 – Visualização

Neste capítulo será apresentado o módulo de visualização para análise das redes sociais científicas multi-relacionais. O objetivo desse módulo é permitir que o usuário faça todas as análises apresentadas neste trabalho de forma simples.

Como dito na apresentação dos resultados dos módulos de agrupamento e de sugestão de relacionamentos, todas as imagens utilizadas neste trabalho para ilustrar e analisar a rede social científica multi-relacional foram extraídas do módulo de visualização que será apresentado no decorrer deste capítulo.

9.1 Introdução

Muitos trabalhos têm sido desenvolvidos para a resolução do problema de análise das redes sociais. Como o assunto é muito extenso e complexo, muitos autores não dão foco à visualização da rede social. Entretanto, acredita-se que a representação gráfica dos elementos permite que as pessoas tenham uma compreensão melhor do problema. Assim, a visualização é uma etapa extremamente importante na análise da rede social, pois a visualização permite que a pessoa analise fatores que não são possíveis de serem verificados sem a ajuda de componentes visuais.

Todo conjunto de dados que possui qualquer tipo de relação pode ser representado por um grafo. Como dito anteriormente, os nós do grafo representam os elementos e as arestas representam as relações entre esses elementos. Existem várias áreas nas quais a Visualização em Grafos pode ser aplicada, são elas: busca de arquivos em árvores, tipos específicos de grafos; redes sociais; mapeamento do histórico em Web sites; dentre outros.

Uma grande dificuldade da visualização em grafos é o tamanho do conjunto de dados que se deseja visualizar. Um grafo formado por grandes conjuntos de dados apresenta dificuldades de visibilidade e de usabilidade. Muitas vezes os nós e arestas se misturam e o operador tem dificuldades em distinguir quais são os reais relacionamentos

de cada objeto. A compreensão e a análise detalhada dos dados representados por um grafo é mais fácil quando o tamanho do grafo é menor.

Com o intuito de facilitar a visualização e a análise de grandes redes sociais foi adotado neste trabalho o conceito de visualização em níveis. Com essa abordagem o usuário é capaz de analisar partes menores da rede social sem a obrigatoriedade de visualizá-la integralmente.

Alguns trabalhos foram desenvolvidos com o objetivo de solucionar o problema de visualização das redes sociais. Entretanto, as redes sociais são estruturas com particularidades bastante específicas dependendo do tipo de problema que está sendo resolvido, o que torna a sua visualização também específica.

O objetivo deste trabalho é permitir que os conceitos envolvidos no módulo de visualização da rede social científica possam ser utilizados para qualquer outro tipo de problema, permitindo que o usuário faça diversas análises baseando-se na estrutura social da rede.

9.2 Visualização de Grafos

Existem várias bibliotecas de código aberto e outras comerciais desenvolvidas com o intuito de facilitar e auxiliar o desenvolvimento de ferramentas para a visualização de grafos. Cada uma dessas bibliotecas possui pontos positivos e pontos que precisam ser melhorados. Neste trabalho, foi feito um levantamento das principais bibliotecas que podem ser utilizadas no desenvolvimento de ferramentas para a análise de redes sociais com o intuito de identificar a que melhor atende ao propósito do módulo de visualização. Algumas dessas bibliotecas serão apresentadas a seguir.

O GINY (*Graph INterface librarY*) (http://csbi.sourceforge.net/index.html) é uma biblioteca de código aberto que fornece vários algoritmos de layout para grafos e é uma API bastante intuitiva. A API pública do GINY define apenas as interfaces Java, para que uma nova implementação possa ser desenvolvida sem grandes dificuldades.

Um dos focos principais desse projeto é facilitar a manipulação das interfaces para que o desenvolvimento da ferramenta para análise de grafos seja feita sem grandes esforços.

Entretanto, embora seja fácil trabalhar com essa API, a principal desvantagem dessa biblioteca são os seus componentes visuais, isto é, mesmo tendo muitas facilidades os componentes visuais não são muito nítidos, o que dificulta a análise das redes sociais científicas e de outros tipos de redes sociais.

O JUNG (*Java Universal Network/Graph*) (http://jung.sourceforge.net/) é uma biblioteca que fornece uma linguagem comum e extensível para a modelagem, análise e visualização de dados que podem ser representados como um grafo ou como uma rede. A arquitetura JUNG foi projetada para suportar uma grande variedade de representações de entidades e suas relações, tais como grafos direcionados e não direcionados, grafos com arestas paralelas, e hiper-grafos.

A distribuição atual de JUNG inclui implementações de uma série de algoritmos de teoria dos grafos, mineração de dados e análise de redes sociais, tais como rotinas para agrupamento, decomposição, otimização, geração de grafos aleatórios, análise estatística e cálculo de distâncias em rede, fluxos e medidas de importância (centralidade, PageRank, HITS, etc.). Essa biblioteca é muito completa e é uma boa opção para o desenvolvimento de ferramentas de visualização de redes sociais.

A Walrus (http://www.caida.org/tools/visualization/walrus/) é uma ferramenta para visualização interativa de grafos direcionados de grande escala no espaço tridimensional. Empregando geometria hiperbólica 3D para exibir gráficos em uma distorção olho de peixe (*fisheye-like*), fornecendo uma exibição que, simultaneamente, mostra detalhe local e o contexto global.

A principal desvantagem dessa biblioteca, que inviabiliza o seu uso neste trabalho, é que ela só trabalha com exibição de árvores. Assim, embora ela tenha todas as vantagens da visualização 3D, essa ferramenta possui uma grande restrição quanto aos tipos de grafos que podem ser exibidos. As redes sociais científicas não possuem a estrutura de árvores e, conseqüentemente, a Walrus não pode ser aplicada a elas.

Prefuse (*Information Visualization Toolkit*) (http://prefuse.org/) é um *toolkit* para a construção de uma interface interativa com o usuário para a visualização de dados estruturados e não estruturados. Isto inclui qualquer tipo de dados que pode ser representado como um conjunto de entidades (ou nós), possivelmente ligados por qualquer número de relacionamentos (ou arestas).

Exemplos de dados suportados pela Prefuse incluem hierarquias (organogramas, taxonomias, sistemas de arquivos), redes (redes de computadores, redes sociais, web sites) e até mesmo coleção de dados não-ligados (cronogramas, gráficos de dispersão).

A principal vantagem, além da interação com o usuário, é que essa ferramenta permite a visualização de agregações de dados. Assim, ela pode ser facilmente utilizada na visualização de problemas de agrupamentos.

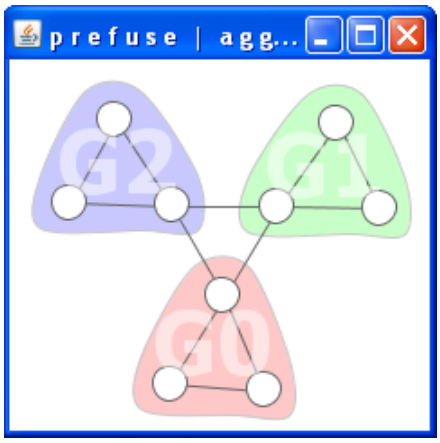


Figura 58 - Representação Visual de Agrupamentos

A Figura 58 ilustra um exemplo simples de problema de agrupamento produzido com auxílio da biblioteca Prefuse, no qual nove nós estão separados em três grupos distintos (G0, G1 e G2). A representação visual da biblioteca permite ao usuário uma rápida identificação de qual é o grupo de um determinado nó.

Avaliando as vantagens e desvantagens de cada biblioteca chegou-se à conclusão de que a Prefuse e a JUNG são as melhores opções para o desenvolvimento do módulo de visualização. Entretanto, optou-se por utilizar a Prefuse, pois essa é a que apresentou o melhor custo benefício para o desenvolvimento da proposta deste módulo de visualização e, além disso, essa biblioteca já havia sido utilizada no desenvolvimento de outros trabalhos (Monclar, 2007).

9.2.1 Conceitos de Visualização de Grafos

O problema de visualização pode ser resumido da seguinte maneira: “Dado um conjunto de nós com um conjunto de arestas (relacionamentos), calcule a posição dos nós e a curva que deverá ser desenhada para cada aresta” (Purchase, 1998).

Existem vários algoritmos para facilitar o desenho de um grafo, por exemplo: grafos em camadas, transformação em um grafo direcionado acíclico, exibir um grafo na forma planar, minimizar a área ocupada pelo layout do grafo, minimizar o número de curvas, minimizarem o número de arestas sobrepostas, etc. (Herman, 2000). Porém, esses algoritmos são em sua maioria muito complexos, sendo inviável o uso deles em aplicações cuja interação com o usuário é freqüente e o tempo de resposta da aplicação deve ser o mais rápido possível.

Como dito no capítulo 2, a definição do layout do grafo é um fator importante na visualização de redes sociais. Existem vários layouts para visualização, tais como representação em 3D e redução de arestas que se cruzam. Os vários modelos existentes na literatura foram definidos por Spritzer (2009) em sua dissertação de mestrado.

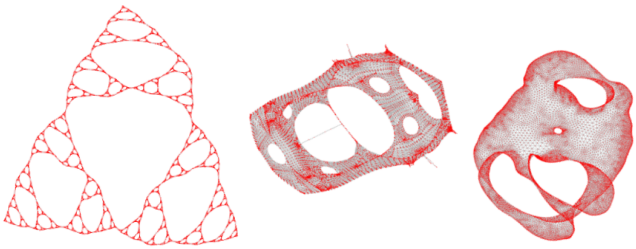


Figura 59 - Grafos gerados pelos algoritmos de Frishman e Tal (2007)

Frishman e Tal (2007) desenvolveram um algoritmo de layout para grafos baseado no layout de força dirigida (*force-directed*). Esse algoritmo apresenta o grafo em múltiplos níveis com o intuito de facilitar a visualização e análise do mesmo. Um exemplo de um grafo gerado por esse algoritmo está na Figura 59.

Na Figura 60 é apresentada outra técnica de definição de layout proposta por Davidson e Harel (1996). O algoritmo proposto por eles, embora exija bastante tempo de processamento, é interessante e produz bons resultados de visualização. A técnica deles

considera informações sobre os vértices, o comprimento das arestas e do número de arestas que se cruzam. Com base nessas informações o algoritmo produz visualizações semelhantes às apresentadas na Figura 60.

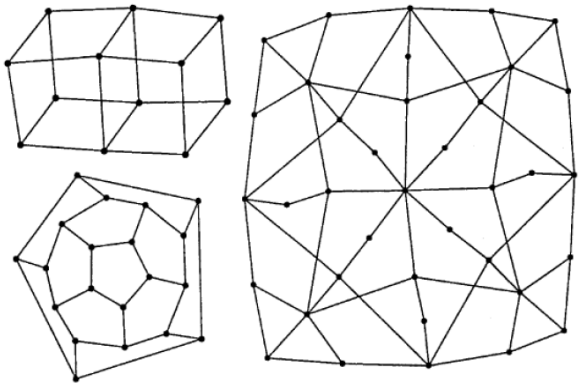


Figura 60 - Exemplos do algoritmo de Davidson e Harel (1996).

A biblioteca Prefuse, usada no desenvolvimento da ferramenta de visualização, utiliza o layout *force-directed* para construir as redes sociais. Assim, esse será o layout padrão adotado no desenvolvimento do módulo de visualização.

O módulo de visualização foi desenvolvido para dar suporte às análises feitas até o momento na rede social científica. Entretanto, muito ainda tem que ser feito com relação à definição do layout.

Como dito anteriormente, o tamanho do grafo é um fator que dificulta a visualização. Muitas vezes, embora o layout do grafo seja bem definido, o tamanho do conjunto de dados torna o uso do grafo inviável. Assim, embora o layout seja extremamente importante para a análise de uma rede social, o tamanho e a densidade do grafo dificultam a manipulação da rede por parte do usuário.

Além do problema com o tamanho dos conjuntos de dados, as técnicas de visualização também sofrem com a navegação feita pelo usuário através do grafo. O conceito de previsibilidade foi identificado como sendo um aspecto importante e necessário nos algoritmos que definem o layout dos grafos (Herman, Delest and Melancëon, 1998; North, 1995).

A previsibilidade em grafos significa que executar o mesmo algoritmo em grafos iguais ou similares não pode produzir representações visuais muito diferentes (Herman, 2000). Essa propriedade também é conhecida na literatura como a preservação do mapa mental do usuário (Misue, Eades, Lai and Sugiyama, 1995), ou seja, após o usuário definir um layout que facilite sua análise da rede, a mesma deveria ser sempre apresentada no formato previamente definido. Nos algoritmos clássicos de geração de layout esse conceito não é utilizado e são sempre geradas visões padrões dos grafos.

Em análise de redes com agregadores o ideal é que os elementos que pertencem aos mesmos grupos estejam sempre próximos independentemente dos relacionamentos que estão sendo exibidos. Dessa maneira, a análise dos agrupamentos fica visualmente mais fácil, pois os elementos que pertencem aos mesmos grupos são rapidamente identificados.

9.2.2 Trabalhos Relacionados

A visualização é uma forma bastante eficaz de análise das redes sociais, pois ela proporciona uma maneira natural de expressar a conectividade e promover um reconhecimento de padrões rápido por parte dos seres humanos. Por isso, a visualização e a análise das redes sociais estão atraindo grande interesse tanto das áreas sociais quanto das áreas de visualização da informação.

Freeman resume a história da visualização das redes sociais através de uma perspectiva sociológica (Freeman, 2000). Entretanto, embora seja uma ferramenta muito poderosa, a visualização envolve vários desafios (Battista, Tollis, Eades and Tamassia, 1999; Freeman, 2000; Grinstein, O'Connell, Laskowski, Plaisant, Scholtz and Whiting, 2006; Heer, Card and Landay, 2005; Henry, Fekete and McGuffin, 2007; Herman, 2000), alguns dos quais foram descritos anteriormente.

Existem algumas ferramentas desenvolvidas para ajudar os analistas a compreenderem melhor as redes sociais. Ferramentas como KrackPlot (Krackhardt, Blythe and McGrath, 1994), Pajet (Nooy, Mrvar and Batagelj, 2005), UCINET (Borgatti, Everett and Freeman, 2002), e *visone* (Brandes and Wagner, 2003) focam seus esforços em análises estatísticas e possuem interações limitadas na visualização.

Outros sistemas, como NetDraw (Borgatti, 2002) e Tom Sawyer (Sawyer, 2011) priorizam a visualização em seus trabalhos, mas não utilizaram muitos algoritmos estatísticos que são importantes para os analistas no estudo das redes sociais. O NetDraw é um aplicativo que permite a análise de redes sociais heterogêneas e de grande porte, com até dois tipos de nós diferentes. Uma lista de aplicativos utilizados para a análise das redes sociais pode ser encontrada em (Huisman and Duijn, 2005) e no site INSNA (http://www.sfu.ca/ insna/).

As pesquisas a partir da perspectiva da visualização da informação colocam mais esforço na representação e exploração visual das redes sociais. Vários métodos utilizados para desenhar grafos foram desenvolvidos para visualizar as redes sociais (Battista, Tollis, Eades and Tamassia, 1999; Herman, 2000; Jünger and Mutzel, 2004) e, conseqüentemente, são aplicados na análise das mesmas.

Perer e Shneiderman fizeram uma revisão completa dos projetos que focam seus desenvolvimentos em melhorar a interação e a forma como as redes sociais são exploradas (Perer and Shneiderman, 2006; Perer and Shneiderman, 2008). Em seus trabalhos eles buscam integrar as análises estatísticas com uma visualização amigável da rede social permitindo que o analista faça uma análise completa da rede (Perer and Shneiderman, 2008). Segundo eles, os trabalhos que envolvem essas duas abordagens são os de *Greenland*, que aumenta o diagrama de nó-ligação segundo uma estratégia estatística (Wong, Foote, Chin, Mackey and Perrine, 2006); e o de *NodeTrix* que usa uma abordagem híbrida do diagrama nó-ligação, que representa a estrutura da rede, e matrizes de adjacência, que destacam as comunidades (Henry, Fekete and McGuffin, 2007).

Muitas das técnicas de visualização desenvolvidas são úteis para a análise de redes sociais de pequeno ou médio porte. Quando se trata de redes sociais complexas e de grande porte o usuário enfrenta dificuldades no processo de análise ao visualizar toda a rede já no primeiro momento.

Para superar a complexidade visual dessas redes muito grandes, Abello apresenta algumas técnicas que permitem aos usuários encontrarem fatos sobre os atores e seus relacionamentos mais rapidamente através de uma navegação interativa na rede. Para tal ele utiliza diferentes níveis de abstração, a partir de um resumo da rede, representando

uma visão de grupos isolados, até chegar a uma visão mais detalhada (Abello, Korn and Finocchi, 2001; Abello and Korn., 2002).

Essas técnicas também permitem que os usuários tenham uma visão geral da rede enquanto trabalham no grupo que lhes interessa (Gansner, Koren and North., 2004; Ham and Wijk, 2004). Shen, Ma e Eliassi-Rad também utilizaram conceitos de abstração no desenvolvimento de seus trabalhos (Shen, Ma and Eliassi-Rad, 2006), onde, além da abstração estrutural, eles utilizaram informações semânticas para auxiliar o trabalho de análise visual das redes sociais heterogêneas de grande porte.

Clark Hu e Prodeep Racherla também usaram técnicas de redução da rede social para tornar possível a representação visual das redes de conhecimento, como as redes sociais acadêmicas e as redes de hospitalidade (Hu and Rachera, 2008). Segundo os autores a representação visual dessas redes de conhecimento contribui para uma melhor compreensão de como ocorrem as colaborações intelectuais em um domínio de conhecimento específico.

Mesmo que existam várias ferramentas voltadas para a análise das redes sociais, ainda são poucos os trabalhos que focam na visualização temporal dessas redes. Uma visualização temporal, ou dinâmica, é uma visualização que indica as mudanças na rede social em diferentes instantes de tempo.

As ferramentas TeCFlow (Gloor and Zhao, 2004) e rSONIA (Bender-deMol, Morris and Moody, 2007) são exemplos de ferramentas desenvolvidas para a análise das redes sociais e que permitem a visualização dinâmica da rede. O grande interesse em visualização de redes sociais dinâmicas gira em torno de como as redes se desenvolvem, evoluem e modificam no decorrer do tempo.

9.3 Framework para Análise Visual de Redes Sociais

O objetivo principal do framework de visualização é permitir que o usuário interaja e analise as redes sociais científicas sobre várias perspectivas diferentes. Assim, além das análises produzidas pelo método de mineração de dados e pela métrica composta, o usuário poderá analisar a rede social em níveis de visualizações diferentes e

compreender como os diferentes tipos de relacionamentos entre os elementos ocorrem em cada um desses níveis.

A ferramenta possui um painel, apresentado na Figura 61, onde o usuário pode alterar a forma como a rede social será exibida. A fim de facilitar a distribuição dos nós e a análise inicial, optou-se por exibir, em um primeiro momento, apenas os relacionamentos que compõe a árvore geradora mínima do grafo social. Assim, a construção da rede social é mais rápida e o usuário tem uma visão inicial mais limpa da rede.

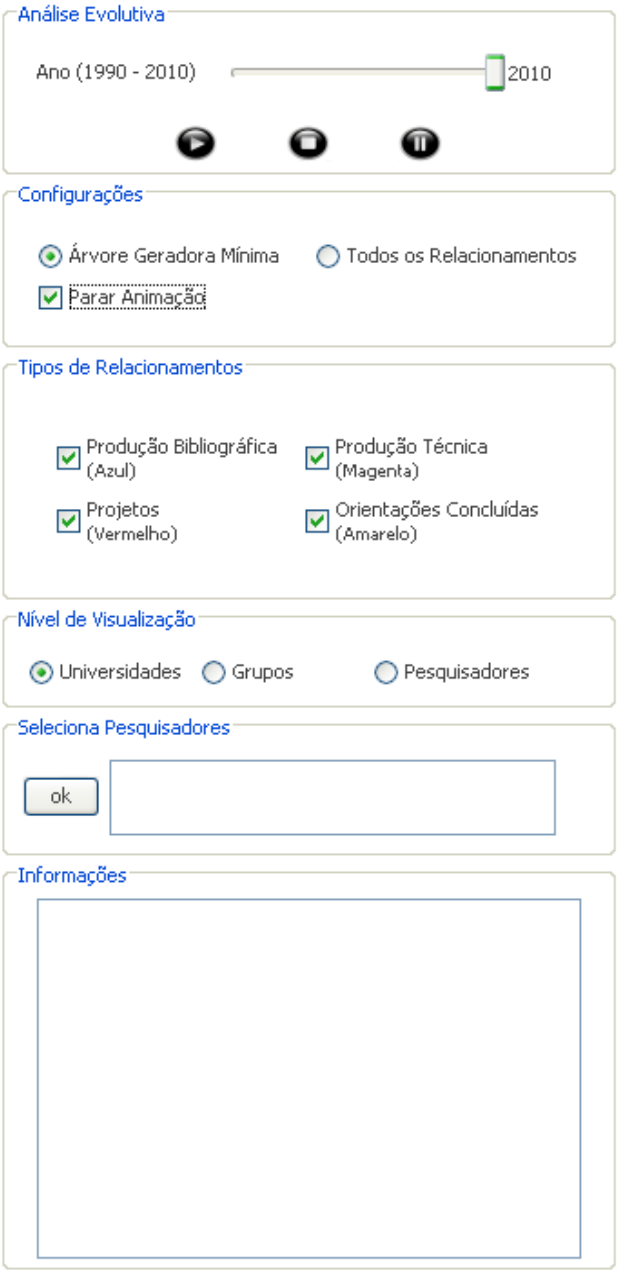


Figura 61 - Painel de Configuração

O primeiro controle disponível para o usuário nesse painel de configurações é a '*Análise Evolutiva*'. Através desse componente o usuário pode analisar a rede social ano a ano e identificar quando os pesquisadores e seus relacionamentos surgiram ao longo dos anos.

Esse componente permite que o usuário selecione um ano específico para a visualização ou que ele veja, como um filme, toda a evolução da rede através do botão *play*. Caso o usuário selecione um ano específico a rede social é apresentada exibindo apenas os pesquisadores e os relacionamentos existentes até o ano especificado. Se o usuário clicar no botão *play* é apresentada toda a evolução da rede social desde 1990 até o ano 2010.

A ferramenta destaca os nós e os relacionamentos novos do ano selecionado pelo usuário para permitir que o mesmo os identifique rapidamente com facilidade. A Figura 62 é um exemplo da visualização da rede social no ano de 2000.

As linhas mais grossas representam os relacionamentos novos e os elementos maiores representam os pesquisadores que entraram na rede social científica no ano 2000. Assim, com o auxílio do componente da análise evolutiva, é possível, por exemplo, saber o ano que um pesquisador entrou na rede social e com quem foi o seu primeiro relacionamento.

A grande vantagem desse componente, além da sua funcionalidade de permitir a análise do passado da rede social, é a velocidade com a qual a rede é atualizada. A resposta do framework quando o usuário interage com o componente mudando o ano de visualização é imediata, ou seja, não exige que o usuário espere para que a rede social seja recarregada, tornando a ferramenta realmente interativa.

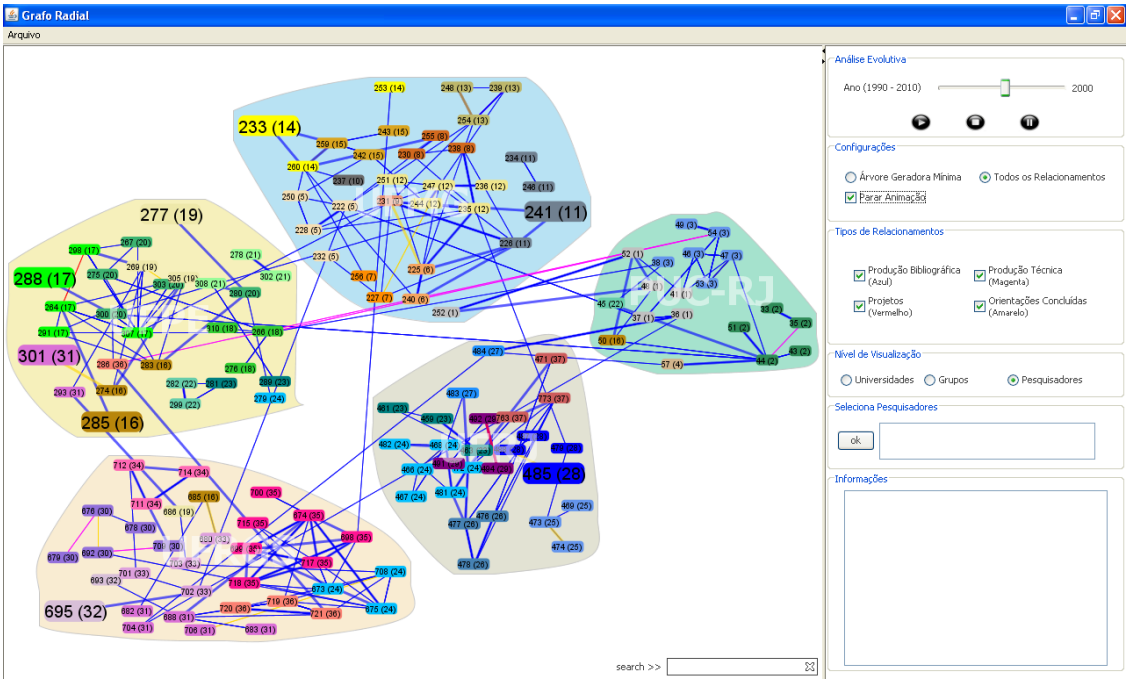


Figura 62 - Análise Evolutiva do ano 2000

O componente para análise evolutiva da rede social científica é uma grande contribuição deste trabalho. Embora existam trabalhos voltados para análise visual das

redes sociais, poucos se preocupam em permitir que o usuário acompanhe a evolução dos elementos ao longo do tempo.

A biblioteca Prefuse utiliza, dentre outros algoritmos, uma estratégia de atração e repulsão para estabilizar o posicionamento dos elementos da rede social, que é o layout *force-directed* definido anteriormente. As arestas do grafo funcionam como molas que puxam os elementos um de encontro ao outro.

Por outro lado, os elementos possuem massas que se repelem fazendo com que eles sejam mantidos distantes uns dos outros. Existe sempre uma força atraindo e outra repelindo os elementos, que estão sempre se acomodando buscando um melhor posicionamento. Com isso, a estabilidade total da rede social pode demorar muito ou até mesmo não ocorrer nunca.

Com o intuito de solucionar a questão apresentada anteriormente, o usuário tem a opção '*Parar Animação'* no componente de configurações. Através dessa ação, o usuário cancela todas as forças que atuam sobre os elementos da rede social e ele está livre para reposicionar os elementos da maneira que melhor lhe convém. Caso ele deseje que os elementos voltem a ser posicionados automaticamente, basta desmarcar a opção de parar a animação ou dar um duplo clique na área de apresentação da rede para que as forças voltem a atuar sobre os elementos da rede social.

Outra opção que o usuário possui é de visualizar apenas a árvore geradora mínima ou todos os relacionamentos da rede social científica. Sempre que o usuário iniciar a ferramenta de visualização, apenas a árvore geradora mínima é exibida, pois a visualização fica mais leve e os nós são posicionados mais rapidamente.

Como este trabalho foi desenvolvido para a análise de redes sociais multi-relacionais, a ferramenta de visualização também precisa dar suporte à análise dos diversos tipos de relacionamentos que compõem a rede. Assim, após optar por visualizar todos os relacionamentos, o usuário pode selecionar qual ou quais tipos de relacionamentos ele deseja visualizar, com isso a ferramenta permite que sejam analisadas apenas as estruturas relacionais desejadas.

No estudo de caso deste trabalho a rede social científica possui quatro tipos de relacionamentos: produção bibliográfica, produção técnica, participação em projetos e

orientações concluídas. A modelagem da rede social científica está descrita em detalhes nos capítulos anteriores.

Para especificar qual o tipo de relacionamento deve ser exibido, o usuário deve utilizar o componente '*Tipos de Relacionamentos*' do painel de configurações. Esse componente sempre virá com todas as opções previamente selecionadas, ou seja, todos os tipos de relacionamentos serão exibidos inicialmente.

A Figura 63 apresenta a mesma rede social com as quatro formações estruturais distintas. Em cada uma das imagens foi definido apenas um tipo de relacionamento. Pode ser observado que o posicionamento dos nós não se altera com a mudança da visualização do tipo de relacionamento. Dessa maneira, as únicas diferenças entre as redes são os relacionamentos entre os elementos, ficando mais fácil a análise das mudanças relacionais de cada objeto.

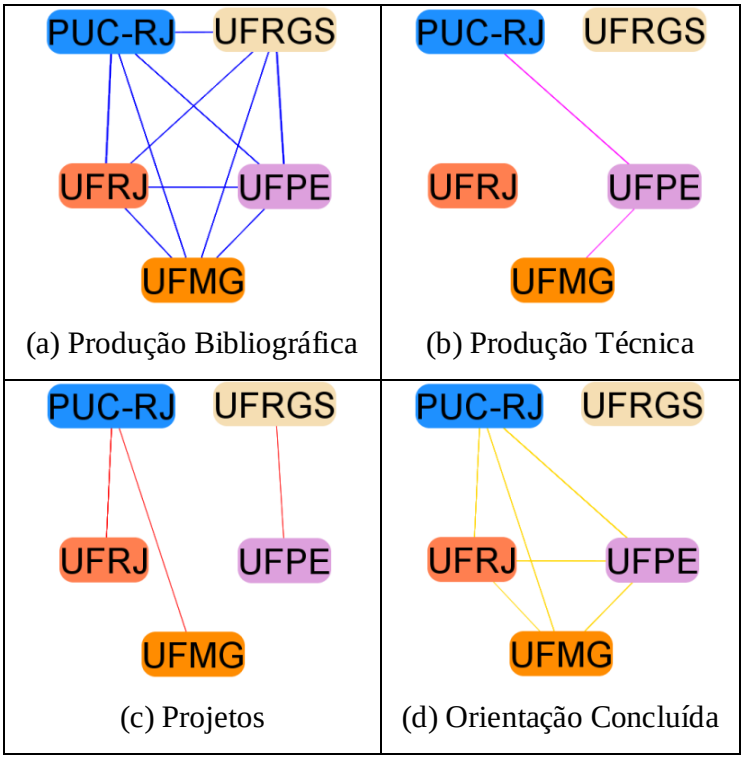


Figura 63 - Filtro por tipo de relacionamento.

Com o intuito de facilitar a análise dos múltiplos relacionamentos entre os pesquisadores, a ferramenta de visualização utiliza arestas com cores distintas para cada

tipo de relacionamento. Os relacionamentos de produção bibliográfica são apresentados em azul, os de produção técnica em magenta, os de projetos em vermelho, e os de orientações concluídas em amarelo.

Além de possuírem cores diferentes essas arestas possuem certo grau de transparência. Assim, a aresta terá uma cor composta pelas cores dos relacionamentos existentes entre os elementos. Por exemplo, caso os elementos possuam relacionamentos de produção bibliográfica e de orientações concluídas a aresta entre eles terá um tom esverdeado, que é a cor gerada pela composição do azul com o amarelo.

O componente da análise evolutiva pode ser utilizado em conjunto com o componente do tipo de relacionamento. Combinando esses dois elementos, o usuário é capaz de analisar a evolução de cada relacionamento individualmente. Sendo possível identificar, por exemplo, quando surgiu o primeiro relacionamento de projeto de um pesquisador; ou qual o padrão relacional de um grupo de pesquisadores; dentre outras análises.

A visualização da rede social foi desenvolvida em níveis nos quais as estruturas e as características sociais de cada um desses níveis são representadas independentemente. Com essa abordagem o usuário sempre possui uma visão geral da rede e ele poderá refiná-la em vários níveis de acordo com a análise que ele deseja realizar. Assim, a visualização da rede social pode ser detalhada até o ponto no qual o usuário irá analisar um nó específico da rede.

No caso das redes sociais científicas, que é o estudo de caso abordado neste trabalho, o usuário irá ter a opção de visualizar a rede social em três níveis distintos. No primeiro nível o usuário irá visualizar as instituições de ensino e os relacionamentos que existem entre elas. No segundo nível, a ferramenta exibe os agrupamentos produzidos pelo algoritmo de mineração de dados, ou seja, são exibidas as comunidades de pesquisas com interesses comuns. Finalmente, no nível três, o usuário irá visualizar todos os pesquisadores das cinco instituições de ensino analisadas neste trabalho.

No segundo e terceiro nível, os objetos são exibidos dentro de objetos agregadores que representam as instituições de ensino. Esse fato é importante para que o analista da rede nunca perca a referência de como está a interação entre as instituições e nem a qual instituição o grupo ou pesquisador pertence.

Para filtrar pelos diferentes níveis da rede social o usuário irá utilizar o componente '*Nível de Visualização'*. O primeiro nível de visualização é o das instituições de ensino. No nível um, o usuário tem a visão mais abstrata da rede social, sendo possível analisar como são os relacionamentos entre os agregadores de mais alto nível na rede.

Através dessa perspectiva mais global, é possível ver todos os relacionamentos entre as instituições de ensino, analisando o fluxo de conhecimento entre elas. A Figura 64 é um exemplo de visualização de nível 1. Nessa figura estão sendo exibidos todos os relacionamentos existentes entre as cinco universidades.

No segundo nível são exibidos os grupos gerados pelo método de mineração de dados. Neste nível de visualização é possível saber, por exemplo, como está a distribuição dos grupos dentro de cada instituição, como eles se relacionam, quantos grupos cada universidade possui, dentre outras análises.

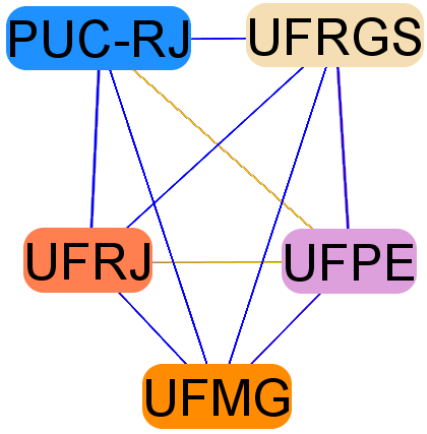


Figura 64 - Visualização Nível 1.

A Figura 65 ilustra o segundo nível de visualização. Nessa figura são apresentados todos os grupos contidos em cada universidade. Cada elemento da rede social representa um grupo dentro da instituição, para facilitar a identificação dos grupos cada nó possui o nome da instituição à qual ele pertence e o número do seu grupo, além disso, elementos que representam o mesmo grupo possuem cores iguais.

Vale observar que o mesmo grupo pode estar contido em universidades diferentes, como é o caso do grupo 1, que está presente na PUC-RJ e na UFMG, e do grupo 16, presente na UFRGS, UFPE e na PUC-RJ.

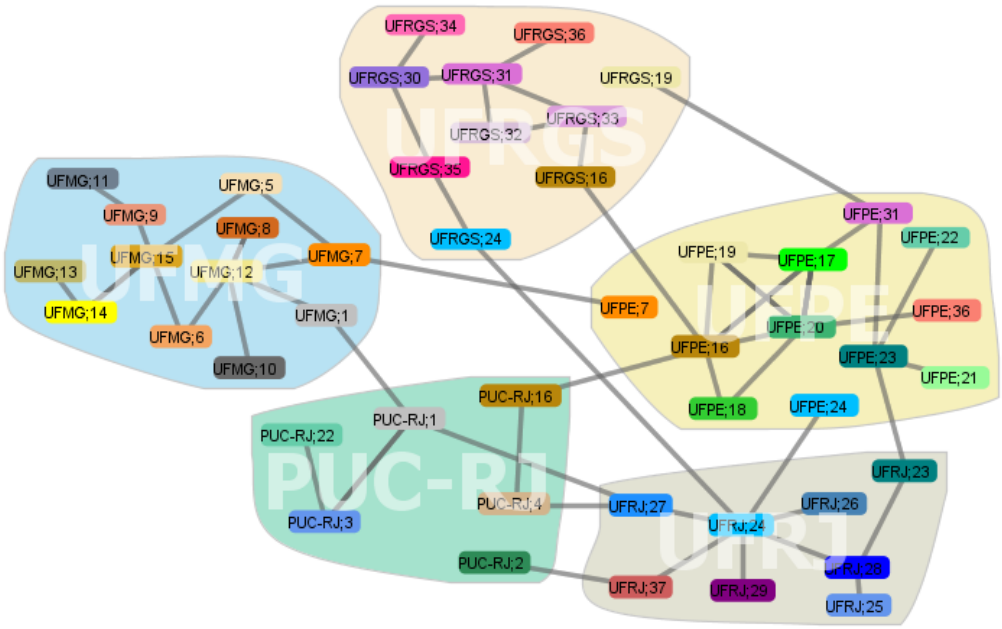


Figura 65 - Visualização Nível 2.

O terceiro nível é a camada mais baixa de visualização, na qual o usuário irá analisar a interação entre os pesquisadores e fazer uma análise local da rede social. Nesse nível o usuário é capaz de analisar os relacionamentos de cada pesquisador avaliando como eles colaboram entre si.

Através dessa funcionalidade o usuário pode identificar quais são os pesquisadores responsáveis pelos relacionamentos externos, quando esse relacionamento ocorreu, qual é o tipo de relacionamento externo, dentre outras análises que ajudam o usuário a compreender o funcionamento da rede social científica.

A Figura 66 apresenta a visualização da rede social científica no nível três. Nessa figura, cada retângulo representa um pesquisador e é composto por dois números. O primeiro representa um código único de identificação do pesquisador, já o segundo número, que está entre parênteses, representa o grupo ao qual o pesquisador pertence.

A fim de dar mais conforto ao usuário para que ele tenha uma resposta visual mais rápida com relação aos agrupamentos, a ferramenta utiliza cores iguais para os elementos que pertencem aos mesmos grupos.

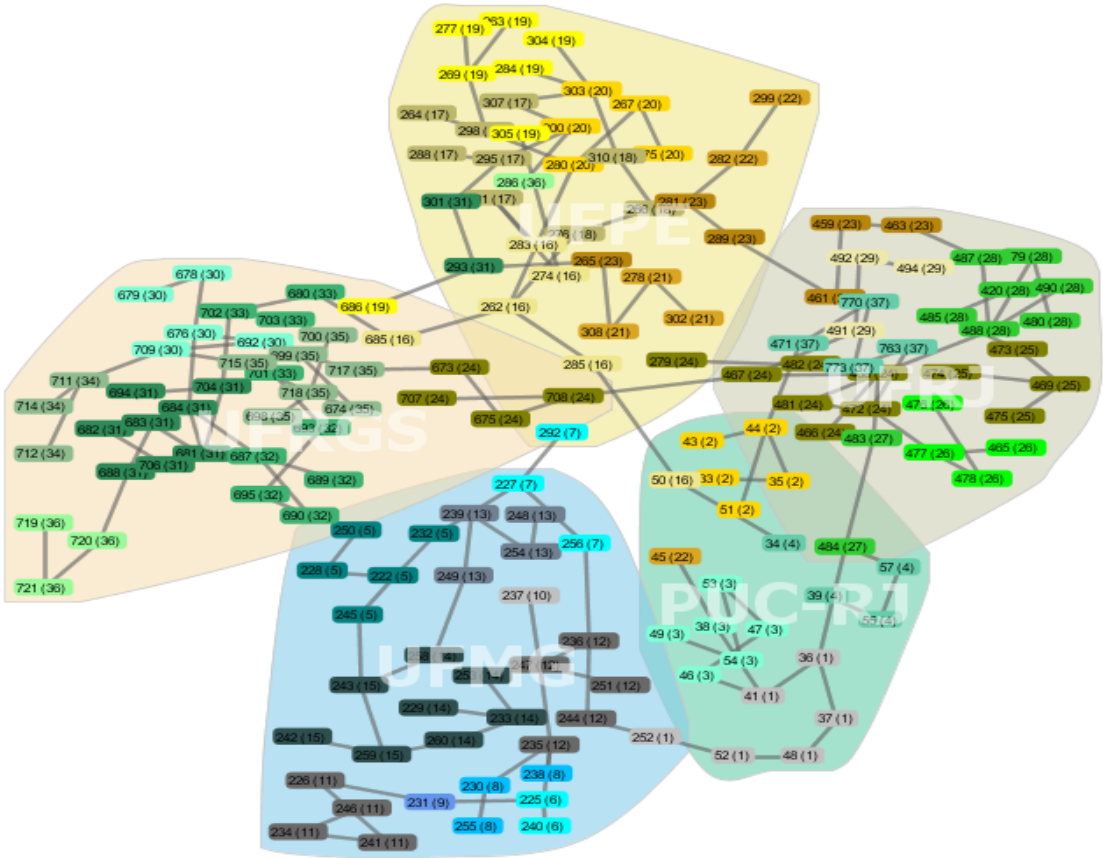


Figura 66 - Visualização Nível 3.

Nesse nível de visualização o usuário pode aplicar filtros na rede social para fazer uma análise voltada para pessoas específicas. Aplicando o filtro para um determinado pesquisador o usuário irá visualizar o nome e a foto de todos os outros pesquisadores diretamente relacionados ao pesquisador do filtro, todos os outros relacionamentos nos quais o pesquisador do filtro não estiver envolvido são escondidos.

Por uma questão de privacidade, as figuras deste trabalho não estão exibindo o nome e a foto dos pesquisadores. Cada pesquisador do conjunto de dados possui um número que o identifica de maneira única. Assim, esses números foram utilizados no lugar das fotos, garantindo a privacidade desses pesquisadores.

Com o uso desse componente que '*Seleciona Pesquisadores*' é possível acompanhar toda a vida acadêmica de uma pessoa. O usuário poderá verificar, por

exemplo, como esses relacionamentos surgiram ao longo do tempo e quais são os tipos de relacionamentos que o pesquisador possui.

A Figura 67 ilustra o uso do componente de filtro por pesquisadores. Nessa figura foi feito um filtro por cinco pesquisadores, os quais aparecem em destaque com tamanho maior que os outros pesquisadores.

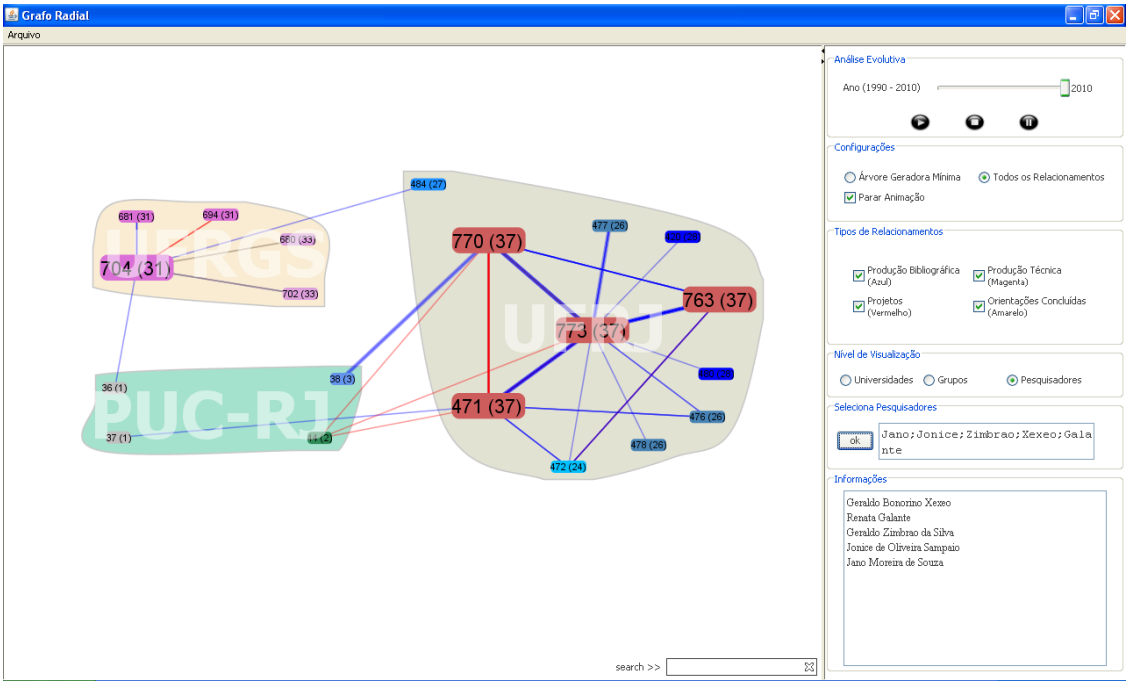


Figura 67 - Filtro por pesquisadores no Nível de visualização 3.

Embora a ferramenta seja composta por diversos componentes visuais que auxiliam e facilitam a análise da rede social científica, nem todas as propriedades dos objetos são possíveis de serem representadas visualmente. Por isso, a ferramenta possui um painel de '*informações*', no qual são exibidas as características específicas de cada elemento da rede.

Ao clicar sobre um elemento da rede (universidade, grupo ou pesquisador), o painel de informações é carregado com as principais características do elemento selecionado, tais como: o número de relacionamentos, relacionamentos externos e internos, tipos de relacionamentos que o elemento possui com outros elementos, dentre outros.

Vale lembrar que todas as configurações apresentadas anteriormente podem ser combinadas da maneira que melhor atender ao usuário. Assim, caso ele queira analisar a evolução dos relacionamentos de um pesquisador, grupo, ou universidade, basta que ele faça o filtro pelo elemento que ele deseja visualizar e manipule o componente da análise do tempo. Se em algum momento ele desejar visualizar apenas um tipo de relacionamento, basta ele desmarcar os tipos de relacionamentos que ele não deseja visualizar. A ferramenta foi desenvolvida para deixar o usuário livre para fazer suas escolhas e definir a melhor forma de analisar a rede social.

Embora existam alguns trabalhos voltados para a análise de redes sociais de grande porte, usando conceitos de abstração, não existe uma preocupação com a análise dos múltiplos relacionamentos e nem com a representação evolutiva da rede. Assim, acredita-se que essas sejam grandes contribuições deste trabalho no módulo de visualização.

9.4 Síntese dos Agrupamentos

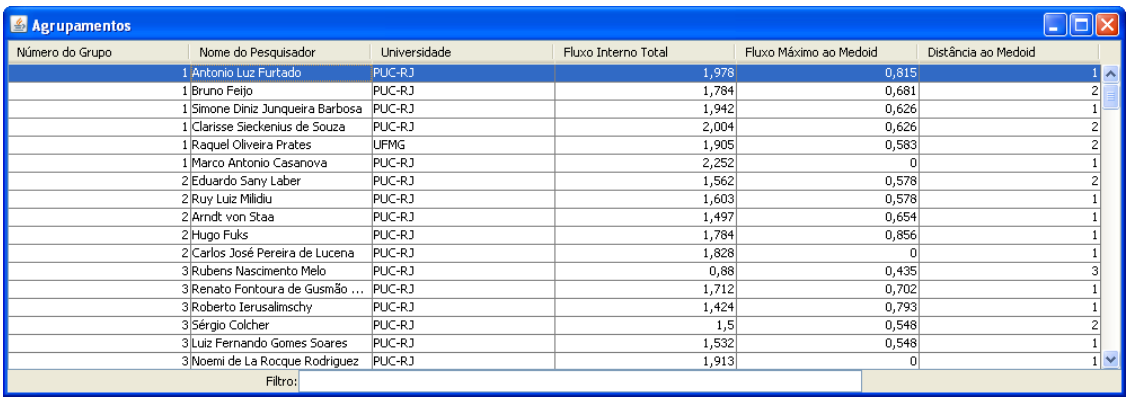
Todos os grupos identificados pela técnica de mineração de dados estão disponíveis visualmente na ferramenta conforme apresentado anteriormente. Na apresentação visual dos grupos é possível identificar os tipos relacionamentos que existem entre eles, se são relacionamentos externos ou internos, dentre outras informações que são possíveis de serem apresentadas visualmente.

Entretanto, existem várias medidas dos agrupamentos que foram utilizadas para a análise dos resultados apresentados no capítulo 6 que são inviáveis de serem exibidas visualmente. Para consolidar essa informação foi criada uma tabela que unifica todas as informações dos grupos.

Essa tabela pode ser acessada através do menu '*Arquivo → Agrupamentos*' da ferramenta de visualização ou através do atalho '*Ctrl + 2*'. A Figura 68 ilustra a tabela que é exibida ao se executar esse comando.

Nessa tabela a coluna 'Número do Grupo', como o próprio nome diz, indica o número do grupo definido pelo método de mineração de dados. As linhas da tabela que possuem o mesmo número representam o mesmo grupo. Esse número é o mesmo exibido

na rede social nos níveis de visualização dois e três para indicar a qual grupo o elemento pertence.



Número do Grupo	Nome do Pesquisador	Universidade	Fluxo Interno Total	Fluxo Máximo ao Medoid	Distância ao Medoid
1	Antonio Luz Furtado	PUC-RJ	1,978	0,815	1
1	Bruno Feijo	PUC-RJ	1,784	0,681	2
1	Simone Diniz Junqueira Barbosa	PUC-RJ	1,942	0,626	1
1	Clarisse Sieckenius de Souza	PUC-RJ	2,004	0,626	2
1	Raquel Oliveira Prates	UFMG	1,905	0,583	2
1	Marco Antonio Casanova	PUC-RJ	2,252	0	1
2	Eduardo Sany Laber	PUC-RJ	1,562	0,578	2
2	Ruy Luiz Milidiu	PUC-RJ	1,603	0,578	1
2	Arndt von Staa	PUC-RJ	1,497	0,654	1
2	Hugo Fuks	PUC-RJ	1,784	0,856	1
2	Carlos José Pereira de Lucena	PUC-RJ	1,828	0	1
3	Rubens Nascimento Melo	PUC-RJ	0,88	0,435	3
3	Renato Fontoura de Gusmão ...	PUC-RJ	1,712	0,702	1
3	Roberto Ierusalimschy	PUC-RJ	1,424	0,793	1
3	Sérgio Colcher	PUC-RJ	1,5	0,548	2
3	Luiz Fernando Gomes Soares	PUC-RJ	1,532	0,548	1
3	Noemi de La Rocque Rodriguez	PUC-RJ	1,913	0	1

Figura 68 – Tabela de Síntese dos Agrupamentos

A coluna ‘Nome do Pesquisador’ apresenta o nome do pesquisador que está contido no grupo. A coluna ‘Universidade’ exibe o nome da universidade do pesquisador. A coluna ‘Fluxo Interno Total’ indica a soma total do fluxo interno desse pesquisador, ou seja, indica o poder de comunicação interna do elemento no grupo.

A coluna ‘Fluxo Máximo ao Medoid’ indica o fluxo máximo entre o pesquisador e o medóide do grupo. O medóide é o elemento selecionado para representar o grupo, pois ele é o elemento que possui o maior poder de comunicação interno no grupo que ele pertence. Se o valor dessa coluna for igual a zero, então o pesquisador é o medóide do grupo.

Finalmente, a coluna ‘Distância ao Medoid’ representa o tamanho do caminho entre o pesquisador e o medóide. Caso o tamanho do caminho seja igual a 1 significa que os pesquisadores são diretamente relacionados.

Neste componente o usuário tem a opção de fazer filtros pelo número do grupo. Assim, caso ele deseje analisar os dados de um grupo específico basta ele utilizar o filtro disponível nesse componente.

9.5 Sugestão de Relacionamentos

Este componente permite que o usuário consulte quais são as principais sugestões de relacionamentos para um pesquisador ou um conjunto de pesquisadores. Para acessar esse módulo o usuário deve ir em '*Arquivo → Sugestões*' ou usar o atalho '*Ctrl + 3*'.

Através desse módulo o usuário terá acesso a uma tela na qual ele pode preencher o nome dos pesquisadores para os quais ele deseja saber as melhores opções de relacionamentos. Ao preencher o nome dos pesquisadores a ferramenta irá listar todas as opções de relacionamentos para os pesquisadores em ordem decrescente com relação ao valor da métrica composta.

Nessa tabela são apresentadas tanto as opções de novos relacionamentos quanto o nível de interação entre os pesquisadores quando o relacionamento já foi estabelecido. O formato da tabela apresentada pela ferramenta de visualização está na Figura 69.

Na tabela de sugestão de relacionamentos são apresentadas oito colunas. A primeira coluna apresenta o '*rank*' das previsões de relacionamentos do pesquisador. Assim, o usuário consegue identificar facilmente em qual posição está uma determinada sugestão de relacionamento.

A segunda e terceira colunas apresentam o nome completo dos pesquisadores para os quais foram feitas as sugestões de relacionamentos e o nome da universidade desses pesquisadores, respectivamente.

Figura 69 - Tabela com as sugestões de Relacionamentos

Na quarta coluna está o nome do pesquisador para o qual foi analisado o nível de interação ou a possibilidade de um novo relacionamento com o pesquisador definido pelo

usuário. A quinta coluna apresenta o nome da instituição do pesquisador para o qual está sendo sugerido o novo relacionamento.

O valor da métrica composta está na quinta coluna. Esse valor representa o nível de interação entre os pesquisadores. Baseando-se nesse valor que os resultados sobre as previsões de relacionamentos do capítulo 8 foram desenvolvidos.

A sexta coluna indica se os pesquisadores já estão relacionados ou não. Caso eles já estejam relacionados o valor da métrica indica como está o fluxo de comunicação total entre os pesquisadores. Com base nesse valor é possível saber se os pesquisadores tendem a se relacionar novamente ou não. Por outro lado, caso eles não estejam relacionados, o valor da métrica irá indicar a possibilidade dos pesquisadores estabelecerem um novo relacionamento.

A última coluna indica se o relacionamento entre os pesquisadores é externo ou não. Essa coluna é interessante para o caso de escolha de membros externos para banca de defesa de tese, pois é possível saber, através dessa coluna, quais são sugestões de relacionamentos externos.

Para facilitar a manipulação dos dados dessa tabela foram adicionados quatro filtros de pesquisa. Através deles o usuário pode filtrar os dados da tabela da maneira que lhe for mais interessante. Por exemplo, caso o usuário queira saber quais são as sugestões de relacionamento entre um pesquisador para os pesquisadores da PUC-RJ, basta que ele coloque esse filtro e serão exibidas na tabela apenas as sugestões de relacionamentos com a PUC-RJ.

Da mesma maneira, caso o usuário queira visualizar apenas as sugestões de relacionamentos para pesquisadores externos ou para pesquisadores não diretamente relacionados, basta que os filtros sejam preenchidos para que a tabela exiba apenas os resultados que são do seu interesse.

Assim, a ferramenta permite que o usuário manipule os dados da maneira que melhor lhe atender. O objetivo é que ele seja capaz de extrair o máximo de informações possíveis do módulo de sugestão de relacionamentos e, conseqüentemente, estabelecer novos relacionamentos que melhorem o fluxo de conhecimento da rede social científica modelada neste trabalho.

Com o desenvolvimento da ferramenta de visualização foi obtido um framework que dá suporte a vários tipos de análises em redes sociais científicas. Através das propostas apresentadas neste trabalho é possível analisar as comunidades científicas, os múltiplos relacionamentos, a evolução temporal da rede social e tudo isso contando com o apoio visual dado pela ferramenta.

Capítulo 10 – Considerações Finais

A internet possibilitou que a barreira da distância fosse quebrada, de forma que os limites físicos que antes existiam não existem mais. Com isso, a análise das redes sociais está se tornando uma ferramenta cada vez mais importante no momento em que as organizações estão se tornando cada vez menos limitadas.

Grande parte das organizações pode ser modelada como redes sociais, nas quais os tipos de relacionamentos e elementos envolvidos podem ser diversificados. Atualmente, existem vários tipos de redes sociais que auxiliam desde empresas a analisarem o mercado financeiro, até o governo de determinados países no combate ao terrorismo.

Neste trabalho foi modelada uma rede social científica multi-relacional, onde os elementos dessa rede representam pesquisadores de instituições de ensino brasileiras e as ligações são diferentes tipos de relacionamentos científicos. Nessa modelagem foram abordados diversos conceitos envolvidos na análise das redes sociais científicas, tais como: idade do relacionamento, perda de informação na transferência de conhecimento, tipos de relacionamentos diferentes, etc.

Um dos objetivos da análise dessa rede social científica multi-relacional é compreender como ocorre o fluxo de conhecimento entre as instituições de ensino. Para tal foi adotado um algoritmo de agrupamento por fluxo máximo, através do qual foi possível identificar as comunidades científicas brasileiras e analisar como essas comunidades trocam seus conhecimentos.

Outro fator importante deste trabalho é a análise evolutiva da rede social científica. Através dessa análise foi possível elaborar uma nova métrica, definida por métrica composta, para previsão/sugestão de novos relacionamentos, que produziu resultados melhores que a métrica de Katz, considerada pela literatura a melhor métrica para o problema de previsão de ligações em redes sociais.

Com essa métrica foi desenvolvido um módulo de sugestão de relacionamentos com o intuito de auxiliar os pesquisadores a estabelecerem novos relacionamentos.

Quando novos relacionamentos são criados o fluxo de conhecimento na rede social científica aumenta.

Melhorando a comunicação em uma Rede Social Científica como um todo, a quantidade e a qualidade das interações entre especialistas de diferentes áreas também melhora. A melhora pode se dar, por exemplo, através da união de especialistas de uma área específica que começam a trabalhar de maneira cooperativa. Como resultado dessa interação, mais dados de alta qualidade pode ser gerado melhorando ainda mais a comunicação na rede social.

Nas próximas seções serão apresentadas as considerações finais das etapas do desenvolvimento deste trabalho. Em cada uma delas serão sugeridos trabalhos futuros que podem dar continuidade às pesquisas iniciadas neste trabalho.

10.1 Modelagem da Rede Social

Neste trabalho foi desenvolvida uma modelagem para a análise de redes sociais científicas multi-relacionais. Embora a modelagem tenha sido aplicada a uma rede social científica, a mesma foi desenvolvida de maneira abstrata para que ela possa ser aplicada a qualquer tipo de rede social.

A modelagem proposta neste trabalho considera diversos aspectos das redes sociais, tais como: tipos de relacionamentos diferentes, pesos diferenciados para cada tipo de relacionamento, inclusão do fator tempo como um dos critérios de peso do relacionamento, perda de informação na transferência de conhecimento entre os elementos, dentre outros.

As análises realizadas sobre a rede social científica segundo o modelo proposto neste trabalho mostraram a validade desse modelo. Segundo os resultados obtidos a modelagem da rede social científica multi-relacional está bem próxima da realidade das instituições de ensino.

Um dos trabalhos futuros na área de modelagem das redes sociais científicas é modelar a rede social da base de dados DBLP. Essa base de dados possui apenas dois tipos de relacionamentos científicos: co-autoria e citação. O tempo também é um fator importante na análise dos dados dessa rede social e, além disso, essa base de dados é

bastante conhecida no meio científico e pode ser modelada segundo os padrões propostos neste trabalho. Após a rede social científica formada pelos dados da base DBLP ser modelada a técnica de agrupamento por fluxo máximo pode ser aplicada para que as comunidades científicas sejam encontradas.

Como a modelagem pode ser aplicada a qualquer tipo de rede social, um dos trabalhos futuros é modelar outros tipos de redes sociais nas quais o fator tempo tenha um impacto importante na análise da rede. Uma das propostas futuras é criar um modelo de redes escuras, que simule, por exemplo, redes terroristas, e verificar se o modelo da rede social científica pode ser aplicado nesses tipos de redes.

10.2 Módulo de Agrupamento

Neste trabalho foi utilizado um método de agrupamento por fluxo máximo para identificar as comunidades de pesquisa na rede social científica brasileira multi-relacional. Os resultados permitiram que fosse feita uma análise detalhada dessa rede social. Vários aspectos da rede social foram analisados, tais como: relacionamentos fortes e fracos; pesquisadores que desempenham um papel centralizador na rede social, evolução temporal dos relacionamentos, etc.

Para o estudo dos agrupamentos foram utilizadas cinco instituições de ensino brasileiras, três delas classificadas como nível 7 e as outras duas classificadas como nível 6 pela Capes. Os grupos dessas cinco instituições foram identificados e analisados sobre aspectos diferentes. Através dessas análises foi possível identificar como ocorre a comunicação entre essas instituições de ensino. Curiosamente, as instituições de nível 7 se comunicam mais com as de nível 6 do que entre si, demonstrando certa competição entre elas.

Um dos trabalhos futuros para o módulo de agrupamento é adicionar as outras instituições de ensino com nível 3, 4 e 5 no conjunto de dados para analisar o fluxo de conhecimento entre as instituições brasileiras de todos os níveis. O objetivo é avaliar se a comunicação entre essas instituições segue o mesmo padrão de comunicação encontrado entre as instituições de nível 6 e 7.

Neste trabalho foi feita uma análise superficial dos centralizadores de conhecimento. Nessa análise foi possível identificar pesquisadores que possuem perfis de relacionamento diferenciados. Concluiu-se que embora a maioria dos pesquisadores possua um perfil de relacionamento interno, existem pesquisadores com perfil de relacionamento externo e interno/externo.

Em trabalhos futuros, serão feitos estudos mais detalhados sobre os centralizadores de conhecimento, com o intuito de verificar se eles são *pontes* ou se existe algum pesquisador classificado como *gargalo* na rede. Os pesquisadores ponte representam aqueles que interligam grupos ou instituições de uma maneira colaborativa. Já os gargalos seriam os pesquisadores que, por não terem bons relacionamentos com outros grupos, dificultam a transferência de conhecimento.

Com essas novas análises sobre os centralizadores de conhecimento será possível avaliar o impacto dessas pessoas em uma instituição educacional, e, finalmente, as ações que devem ser tomadas tanto para reconhecer essas pessoas quanto as mudanças que devem ser feitas para aliviar os gargalos. Essas análises devem ser realizadas em conjunto com os trabalhos de equilíbrio em redes sociais (Monclar, 2007), pois esses trabalhos já possuem os conceitos estruturais da rede social.

10.3 Módulo de Sugestão de Relacionamentos

Neste trabalho foi feita uma análise detalhada da previsão de relacionamentos em redes sociais científicas. Nessa análise foram estudadas duas métricas, a métrica de Katz e a métrica composta, que é a métrica proposta neste trabalho. Os parâmetros dessas métricas foram analisados para identificar quais valores produzem as melhores previsões de relacionamentos.

Com a análise da evolução temporal da rede social foi possível propor uma nova métrica para sugerir novos relacionamentos em Redes Sociais Científicas. Com o uso da métrica composta o resultado da previsão de relacionamentos foi mais bem sucedido do que quando se utiliza a métrica Katz, que é a métrica mais utilizada na literatura.

O principal objetivo do módulo de sugestão de relacionamentos é aumentar o fluxo de conhecimento e melhorar a troca de informações na rede social científica como um todo.

Com o intuito de analisar o comportamento da métrica composta em outras redes sociais, em um trabalho futuro, a mesma será aplicada na rede social formada pelos dados da base DBLP. O objetivo é avaliar a métrica composta em uma rede científica de grande porte.

Além disso, pretende-se aplicar a métrica proposta neste trabalho em outros tipos de redes sociais. O objetivo é verificar se os diferentes tipos de redes sociais apresentam comportamentos semelhantes ou se outras estratégias devem ser adotadas para prever/sugerir novos relacionamentos.

Nesse trabalho futuro, a métrica composta poderá ser aplicada no estudo de redes terroristas onde as informações estão escondidas na rede social. O objetivo é descobrir se a métrica composta irá apresentar o mesmo comportamento identificado nas redes sociais científicas.

Outro estudo interessante é aplicar a métrica composta nas redes sociais comerciais, nas quais o interesse de descoberta de novos relacionamentos não é de colaboração e sim de competição. Nesse tipo de rede social é extremamente importante esconder as informações da sua empresa e tentar descobrir em que os seus concorrentes estão trabalhando.

10.4 Módulo de Visualização

O módulo de visualização desenvolvido neste trabalho tem o intuito de facilitar as análises da rede social científica multi-relacional. Esse módulo foi desenvolvido com o auxílio da biblioteca Prefuse, que é uma biblioteca desenvolvida em Java para auxiliar os estudos das redes sociais.

Esse módulo permite que o usuário faça análises evolutivas na rede social; avalie os níveis da rede social; analise cada tipo de relacionamento independentemente; faça pesquisas e filtros por pesquisadores para ter uma visão local da rede social; dentre outras análises visuais.

Um dos trabalhos futuros deste módulo de visualização é adicionar na rede social as arestas que representam as sugestões de novos relacionamentos. Para tal, o usuário deverá consultar as sugestões de relacionamentos para um pesquisador ou grupo de pesquisadores; informar quantos relacionamentos ele deseja que sejam sugeridos; e a ferramenta irá exibir as sugestões de relacionamentos na área de visualização da rede.

Outro fator que deve ser trabalhado futuramente é a definição do layout da rede social científica. O objetivo é definir um layout que minimize as arestas que se cruzem e, conseqüentemente, tornar a visualização da rede social ainda mais nítida.

Dessa maneira o usuário será capaz de avaliar visualmente todos os relacionamentos envolvidos na análise da sugestão dos relacionamentos. Visualizando os novos relacionamentos ele será capaz de saber se é uma boa sugestão para ele ou não, dados os caminhos que estão envolvidos na análise da métrica composta.

O objetivo dos trabalhos futuros sugeridos neste capítulo é aprimorar ainda mais o framework de análise de redes sociais científicas multi-relacionais desenvolvido neste trabalho. Assim, ele poderá ser utilizado sobre vários aspectos e auxiliar os pesquisadores das instituições de ensino a estabelecerem novos vínculos, melhorando o desenvolvimento de suas pesquisas e o fluxo de conhecimento na rede social científica.

A ferramenta desenvolvida neste trabalho é uma ferramenta poderosa para a análise de redes sociais multi-relacionais em instituições científicas que pode ser facilmente adaptada para análise da rede social de outras organizações.

Referências Bibliográficas

Abello, J., Korn, J. and Finocchi, I.: 2001, 'Graph Sketches.' *Proc. IEEE Symp. Information Visualization*, 67.

Abello, J. and Korn., J.: 2002, 'MGV: A System for Visualizing Massive Multidigraphs.' *IEEE Trans. Visualization and Computer Graphics* **vol. 8, no. 1**, 21-38.

Acar, E., Dunlavy, D. M. and Kolda, T. G.: 2009. 'Link Prediction on Evolving Data Using Matrix and Tensor Factorizations', in, *IEEE International Conference on Data Mining Workshops*, pp. 262-269.

Agrawal, R., Rajagopalan, S., Srikant, R. and Xu, Y.: 2003. 'Mining newsgroups using networks arising from social behavior', in, *Proceedings of 12th International World Wide Web Conference*.

Aldenderfer, M. S. and Blashfield, R. K.: 1984, *Cluster analysis*. Beverly Hills, CA: Sage.

Assunção, R. M., Neves, M. C., Câmara, G. and Freitas, C. C.: 2006, 'Efficient regionalization techniques for socioeconomic geographical units using minimum spanning trees.' *International Journal of Geographical Information Science* **v. 20**, pages 797–811.

Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A. and Vicsek, T.: 2002, 'Evolution of the social network of scientific collaborations.' *Physica A 311*, pp. 590-614.

Battista, G. D., Tollis, I. G., Eades, P. and Tamassia, R.: 1999, *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall.

Bender-deMol, S., Morris, M. and Moody, J.: 2007, 'Prototype Packages for Managing and Animating Longitudinal Network Data: dynamicnetwork and rSoNIA.' *Journal of Statistical Software* **vol. 24, no.7**, 1-36.

Bezdek, C. J. and Pal, R. N.: 1998, 'Some New Indexes of Cluster Validity.' *IEEE Transations on systems, man, and cybernetics – Part B: Cybernetics* **v. 28**.

Bird, C., Gourley, A., Devanbu, P., Gertz, M. and Swaminathan, A.: 2006. 'Mining email social networks', in, *Proceedings of the 2006 international workshop on Mining software repositories*, pp. 137-143.

Boaventura, P. O.: 1996, *Grafos: teoria, modelos, algoritmos*. Edgard Blücher LTDA.

Borgatti, S. P.: 2002, 'NetDraw Software for Network Visualization.' *Analytic Technologies: Lexington, KY*.

Borgatti, S. P., Everett, M. G. and Freeman, L. C.: 2002, 'Ucinet for Windows: Software for Social Network Analysis.' *Harvard, MA: Analytic Technologies*.

Boyd, J. P., Fitzgerald, W. J., Mahutga, M. C. and Smith, D. A.: 2010, 'Computing continuous core/periphery structures for social relations data with MINRES/SVD.' *Journal of Social Networks* **vol. 32**, 125-137.

Brandes, U. and Wagner, D.: 2003, 'Visone - Analysis and Visualization of Social Networks In Graph Drawing Software.' *M. Junger and P. Mutzel. Springer-Verlag*.

Cai, D., Shao, Z., He, X., Yan, X. and Han, J.: 2005. 'Community mining from multi-relational networks', in, *Proceedings of the 2005 European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05)*, Porto, Portugal.

Campbell, C., Maglio, P., Cozzi, A. and Dom, B.: 2003. 'Expertise Identification using Email Communications', in, *Proceedings of the twelfth international conference on Information and knowledge management*, New Orleans.

Carley, K.: 2003. 'Dynamic Network Analysis', in, *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, Committee on Human Factors, National Research Council, National Research Council, Washington, DC, pp. 133–145.

Carpenter, T., Karakostas, G. and Shallcross, D.: 2002, 'Practical Issues and Algorithms for Analyzing Terrorist Networks.' *Telcordia Technologies*.

Chen, F., Farahat, A. and Brants, T.: 2004. 'Multiple Similarity Measures and Source-Pair Information in Story Link Detection', in, *Proceedings of the Human Language Technology Conference, North American Chapter of the Association for Computational Linguistics Annual Meeting*, Boston.

Clauset, A., Moore, C. and Newman, M. E. J.: 2008, 'Hierarchical structure and the prediction of missing links in networks.' *Nature* **vol. 453**, 98-101.

Coffman, T., Greenblatt, S. and Marcus, S.: 2004, 'Graph-Based Technologies for Intelligence Analysis.' *Communications of the ACM* **vol. 47 no. 3**, 45-47.

Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C.: 2001, 'Introduction to Algorithms (Second ed.).' **MIT Press and McGraw–Hill. pp. 651–664. ISBN 0-262-03293-7**.

Cross, R. L., Parker, A. and Cross, R.: 2004, *The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations*. Harvard Business Press.

Davidson, R. and Harel, D.: 1996, 'Drawing graphs nicely using simulated annealing.' *ACM Trans. Graph* **vol. 15, no. 4**, 301-331.

Desikan, P. and Srivastava, J.: 2004. 'Mining temporally evolving graphs', in, *Proceedings of the sixth WEBKDD workshop in conjunction with the 10th ACM SIGKDD conference*, Seattle.

Dietrich, J., Dietrich, J. and Wright, J.: 2008, 'Using social networking and semantic web technology in software engineering – Use cases, patterns, and a case study.' *Journal of Systems and Software* **Volume 81**, Pages 2183-2193.

Edmonds, J. and Karp, R. M.: 1972, 'Theoretical improvements in algorithmic efficiency for network flow problems.' *Journal of the ACM 19 (2): 248–264. DOI:10.1145/321694.321699*.

Farrel, S., Campbell, C. and Mayagmar, S.: 2005. 'Relescope: An Experiment in Accelerating Relationships', in, *Conference on Human Factors in Computing Systems*, Portland.

Fellman, P. V. and Wright, R.: 2004. 'Modelling Terrorist Networks - Complex Systems at the Mid-Range', in, *Joint Complexity Conference*, London School of Economics.

Flake, G. W., Lawrence, S. and Giles, C. L.: 2000. 'Efficient identification of Web communities', in, *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD-2000).*

Ford, L. R. and Fulkerson, D. R.: 1956, 'Maximal flow through a network.' *Canadian Journal of Mathematics 8: 399–404*.

Freeman, I. C.: 2000, 'Visualizing Social Networks.' *Journal of Social Structure* **vol.1 no. 1**.

Freeman, L.: 1979. 'Centrality in social networks: Conceptual clarifications', in, *Social Networks, 1:215-239*, Vol. Social Networks, 1:215-239.

Frigge, M., Hoaglin, D. C. and Iglewicz, B.: 1989, 'Some Implementations of the Boxplot.' *The American Statistician* **Vol. 43, No. 1**, 50-54.

Frishman, Y. and Tal, A.: 2007, 'Multi-Level Graph Layout on the GPU.' *Visualization and Computer Graphics, IEEE Transactions on* **vol. 13, no. 6**, 1310-1319.

Gansner, E., Koren, Y. and North., S.: 2004, 'Topological Fisheye Views for Visualizing Large Graphs.' *Proc. IEEE Symp. Information Visualization*, 175-182.

Gibson, D., Kleinberg, J. and Raghavan, P.: 1998. 'Inferring Web communities from link topology', in, *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*.

Gloor, P. A. and Zhao, Y.: 2004. 'TeCFlow – A Temporal Communication Flow Visualizer for Social Network Analysis', in, *ACM CSCW Workshop on Social Networks. ACM CSCW Conference*.

Golbeck, J.: 2005. 'Semantic Web Interaction through Trust Network Recommender Systems End User Semantic Web Interaction Workshop ', in, *4th International Semantic Web Conference*.

Golbeck, J. and Hendler, J.: 2004. 'Reputation Network Analysis for Email Filtering', in, *Proceedings of the First Conference on Email and Anti-Spam*.

Goldenberg, A., Kubica, J., Komarek, P., Moore, A. and Schneider, J.: 2003. 'A Comparison of Statistical and Machine Learning Algorithms on the Task of Link Completion', in, *Proceedings of the KDD Workshop on Link Analysis for Detecting Complex Behavior*.

Grinstein, G., O'Connell, T., Laskowski, S., Plaisant, C., Scholtz, J. and Whiting, M.: 2006. 'The VAST 2006 Contest: A tale of Alderwood', in, *Proc. IEEE Symp. on Visual Analytics Science and Technology*.

Ham, F. V. and Wijk, J. J. V.: 2004, 'Interactive Visualization of Small World Graphs.' *Proc. IEEE Symp. Information Visualization*, 199-206.

Han, J. and Kamber, M.: 2006, *Data Mining: Concepts and techniques*. USA, Morgan Kaufmann Publishers.

Han, L. and Yan, H.: 2011, 'BSN: An automatic generation algorithm of social network data.' *Journal of Systems and Software* **Volume 84**, 1261-1269

Hansen, D., Shneiderman, B. and Smith, M. A.: 2010, *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. Morgan Kaufmann.

Heer, J., Card, S. K. and Landay, J. A.: 2005. 'Prefuse: A Toolkit for Interactive Information Visualization', in, *Proc. ACM Conf. on Human Factors in Computing Systems*.

Henry, N., Fekete, J.-D. and McGuffin, M. J.: 2007, 'NodeTrix: a Hybrid Visualization of Social Networks.' *IEEE Transactions on Visualization and Computer Graphics* **vol. 13, no. 6**, 1302-1309.

Henziger, M.: 2000, 'Link Analysis in Web Information Retrieval.' *IEEE Data Engineering Bulletin* **vol. 23**, 3 – 8.

Herman, I.: 2000, 'Graph Visualization and Navigation in Information Visualization: A Survey.' *IEEE Transactions on Visualization and Computer Graphics* **Vol. 6, no. 1**.

Herman, I., Delest, M. and Melancëon, G.: 1998, 'Tree Visualization and Navigation Clues for Information Visualization.' *Computer Graphics Forum* **vol. 17, no. 2**, 153-165.

Hershberger, J., Maxely, M. and Suriz, S.: 2007, 'Finding the k Shortest Simple Paths: A New Algorithm and its Implementation.' *Journal ACM Transactions on Algorithms* **vol. 3**.

Holme, P.: 2003, 'Network dynamics of ongoing social relationships.' *Europhys. Lett.* **vol. 64**, 427-433.

Holme, P., Edling, C. and Liljeros, F.: 2004, 'Structure and Time Evolution of an Internet Dating Community.' *Social Networks* **no. 26**, 155-174.

Hu, C. and Rachera, P.: 2008, 'Visual Representation of Knowledge Networks: A Social Network Analysis of Hospitality Research Domain.' *Int'l J. Hospitality Manage.* **vol. 27**, 302-312.

Huang, Liben-Nowell, D. and Kleinberg, J.: 2003. 'The link prediction problem for social networks', in, *Proceedings of the twelfth international conference on information and knowledge management*, pp. pp. 556-559.

Huang, Taskar, B., Abbeel, P., Wong, M.-F. and Koller, D.: 2003. 'Label and Link Prediction in Relational Data', in, *Proceedings of the IJCAI Workshop on Learning Statistical Models from Relational Data*.

Huang, Z., Chen, H. and Zeng, D.: 2001, 'Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering.' *ACM 1073-0516/01/0300-0034*.

Huang, Z., Li, X. and Chen, H.: 2005, 'Link Prediction Approach to Collaborative Filtering.' *JCDL*, 141-142.

Huang, Z. and Lin, D. K. J.: 2009, 'The time-series link prediction problem with applications in communication surveillance.' *INFORMS J. Computing* **vol. 21**, 286-303.

Huisman, M. and Duijn, M. A. J. V.: 2005, 'Software for Social Network Analysis.' *Models and Methods in Social Network Analysis, P.J. Carrington, J. Scott and S. Wasserman, eds. Cambridge University Press,* 270-316.

Ichise, R., Takeda, H. and Ueyama, K.: 2005. 'Community Mining Tool using Bibliography Data', in, *Proceedings of the Ninth International Conference on Information Visualisation*, IEEE.

Jackson, M. O.: 2010, *Social and Economic Networks*. Princeton University Press.

Jain, A. K. and Dubes, R. C.: 1988, *Algorithms for Clustering Data*. Prentice Hall, Michigan State University.

Jung, J. J., Juszczyszyn, K. and Nguyen, N. T.: 2007, 'Centrality Measurement on Semantically Multiplex Social Networks: Divide-and-Conquer Approach.' *International Journal of Intelligent Information and Database Systems* **Vol. 1, No. 3/4**, 277-292.

Jünger, M. and Mutzel, P.: 2004, *Graph Drawing Software*. Springer.

Kannan, R., Vempala, S. and Vetta, A.: 2000. 'On clusterings: Good, bad and spectral.' in, *Proceedings of the 41st Foundations of Computer Science (FOCS '00),*, pp. 367.

Knoke, D. and Song, Y.: 2008, *GSocial Network Analysis*. 2nd ed. Series: Quantitative applications in the social sciences, **154**.

Komarek, P.: 2004, *Logistic Regression for Data Mining and High-Dimensional Classification*. Carnegie Mellon University.

Krackhardt, D., Blythe, J. and McGrath, C.: 1994, 'KrackPlot 3.0: An Improved Network Drawing Program.' *Connections* **vol. 17, no. 2**, 53-55.

Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A.: 1999. 'Trawling the Web for emerging cyber communities', in, *Proceedings of The 8th International World Wide Web Conference*.

LATTES: 2008. 'LATTES', in, http://lattes.cnpq.br/eng/.

Leskovec, J., Kleinberg, J. and Faloutsos, C.: 2005. 'Graphs Over Time: Densification Laws, Shrinking Diameters and Possible Explanations', in, *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*.

Liben-Nowell, D. and Kleinberg, J.: 2007, 'The Link Prediction problem for social networks.' *Journal of the American Society for Information Science and Technology* **vol. 58**, 1019–1031.

Lim, M., Negnevitsky, M. and Hartnett, J.: 2005. 'Artificial Intelligence Applications for Analysis of E-mail Communication Activities', in, *Proceedings International Conference On Artificial Intelligence In Science And Technology, p.p. 109-113*.

Lü, L. and Zhou, T.: 2009. 'Role of weak ties in link prediction of complex networks', in, *In Proceeding of the 1st ACM international Workshop on Complex Networks Meet information & Knowledge Management*, Hong Kong.

Martins, E. Q. V. and Pascoal, M. M. B.: 2003, 'A new implementation ofYen's ranking loopless paths algorithm.' *4OR – Quarterly Journal of the Belgian, French and Italian Operations Research Societies* **vol. 1**, 121–133.

McMahon, S., Miller, K. and Drake, J.: 2001, 'Networking Tips for Social Scientists and Ecologists.' *Journal of Science* **vol. 293**, 1604-1605.

Milgram, S.: 1974, *Obedience to authority: an experimental view*. Harper and Row, New York.

Misue, K., Eades, P., Lai, W. and Sugiyama, K.: 1995, 'Layout Adjustment and the Mental Map.' *J. Visual Languages and Computing* **vol. 6**, 183-210.

Monclar, R. S., et al: 2007. 'A New Approach to Balance Social Networks', in, *Proceedings of UK Social Network Conference*, pp. p. 141-142.

Newman, M. E. J.: 2000. 'Who are the best connected scientists? A study of scientific co-authorship networks', in, SFI Working Paper 00-12-64, Santa Fe.

Newman, M. E. J.: 2001a. 'Scientific collaboration networks. I. Network construction and fundamental results', in, Physical Review E, vol. 64, no. 1, pp. 016 131+.

Newman, M. E. J.: 2001b. 'Scientific collaboration networks. II. shortest paths, weighted networks, and centrality', in, Physical Review E, vol. 64, no. 1, pp. 016 132+.

Newman, M. E. J.: 2001c, 'The structure of scientific collaboration networks.' *Proceedings of the National Academy of Science* **USA 98, 404-409**.

Newman, M. E. J.: 2004a. 'Co-authorship networks and patterns of scientific collaboration', in, *Proceedings of the National Academy of Sciences, 101: 5200-5205*.

Newman, M. E. J.: 2004b, 'Detecting community structure in networks.' *European Physical Journal B, 38:321-330*.

Nooy, W. d., Mrvar, A. and Batagelj, V.: 2005, *Exploratory Social Network Analysis with Pajek*. Cambridge University Press.

North, S.: 1995, 'Incremental Layout in DynaDAG.' *Proc. Symp. Graph Drawing GD '95*, 409-418.

Oliveira, J., Souza, J. M., Miranda, R. and Rodrigues, S.: 2006. 'GCC: A Knowledge Management Environment for Research Centers and Universities', in, *Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, 3841/2006: 652-667.

Opsahla, T., Agneessensb, F. and Skvoretzc, J.: 2010, 'Node centrality in weighted networks: Generalizing degree and shortest paths.' *Social Networks* **vol. 32**, 245–251.

Pakhira, M. K., Bandyopadhyay, S. and Maulik, U.: 2004, 'Validity index for crisp and fuzzy clusters.' *Pattern Recognition*.

Perer, A. and Shneiderman, B.: 2006, 'Balancing Systematic and Flexible Exploration of Social Networks.' *IEEE Trans. on Visualization and Computer Graphics* **vol. 12, no. 5**, 693-700.

Perer, A. and Shneiderman, B.: 2008. 'Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis', in, *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*.

Pioch, N., Barlos, F., Fournelle, C. and Stephenson, T.: 2005. 'A Link and Group Analysis Toolkit (LGAT) for Intelligence Analysis', in, https://analysis.mitre.org/proceedings/Final_Papers_Files/348_Camera_Ready_Paper.pdf.

Popescul, A. and Ungar, L. H.: 2003. 'Structural Logistic Regression for Link Analysis', in, *Proceedings. of KDD Workshop on Multi-Relational Data Mining*.

Popescul, A. and Ungar, L. H.: 2004. 'Cluster-based Concept Invention for Statistical Relational Learning', in, *Proceedings of Conference Knowledge Discovery and Data Mining (KDD-2004)*, pp. 22-25.

Potgieter, A., April, K. A., Cooke, R. J. E. and Osunmakinde, I. O.: 2009, 'Temporality in Link Prediction: Understanding Social Complexity.' *Journal Article E:CO* **vol. 11**.

Pozo, M. D., Manuel, C., González-Arangüena, E. and Owen, G.: 2011, 'Centrality in directed social networks. A game theoretic approach.' *Social Networks* **vol. 33**, 191-200.

Purchase, H. C.: 1998. 'Which Aesthetic Has the Greatest Effect on Human Understanding?' in, *Proc. Symp. Graph Drawing GD '97*, pp. 248-261.

Raab, J. and Milward, H.: 2003, 'Dark Networks as Problems.' *Journal of Public Administration Research and Theory, vol. 13, no. 4. pp 413-439*.

Rattigan, M. J. and Jensen, D.: 2005. 'The Case for Anomalous Link Detection', in, *Proceedings of the 4th international workshop on multi-relational mining*, pp. 69-74.

Ryan, S. and O'Connor, R. V.: 2009, 'Development of a team measure for tacit knowledge in software development teams.' *Journal of Systems and Software* **Volume 82**, 229-240

Sawyer, T.: 2011, 'Visual Interaction Methods for Clustered Graphs.' *Australian Research Council (ARC)*, 11-13.

Schwartz, M. F. and Wood, D. C. M.: 1993, 'Discovering shared interests using graph analysis.' *Communications of the ACM, 36(8):78–89*.

Shen, Z., Ma, K.-L. and Eliassi-Rad, T.: 2006, 'Visual Analysis of Large Heterogeneous Social Networks by Semantic and Structural Abstraction.' *IEEE Transactions on Visualization and Computer Graphics* **vol. 16, no. 6**, 1427-1439.

Skillicorn, D. B.: 2004. 'Social Network Analysis via Matrix Decompositions: al Qaeda', in, Technical report, School of Computing, Queen's University, Canada.

Spritzer, A. S.: 2009, *MagnetViz: Design and Evaluation of a Physics-based Interaction Techinique for Graph Visualization*. Universidade Federal do Rio Grande do Sul (UFRGS).

Ströele, V., Oliveira, J., Zimbrao, G. and Souza, J. M.: 2009. 'Mining and Analyzing Multirelational Social Networks', in, *2009 International Conference on Social Computing (SocialCom09)*, Proceedings of International Conference on Social Computing (IEEE CS), Vancouver.

Ströele, V., Silva, R., Oliveira, J., Souza, J. M. and Zimbrao, G.: 2009. 'Mining and Analyzing Organizational Social Networks for Collaborative Design', in, *13th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2009)*, Proceedings of the 13th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2009), Santiago.

Ströele, V., Silva, R., Souza, M., Oliveira, J., Mello, C. E., Souza, J. M. and Zimbrão, G.: 2008. 'Mining and Analyzing Organizational Social Networks Using Minimum Spanning Tree', in, *16th International Conference on Cooperative Information Systems (CoopIS 2008)*, Monterrey.

Ströele, V., Silva, R., Souza, M. F. d., Mello, C. E. R. d., Souza, J. M., Zimbrao, G. and Oliveira, J.: 2011, 'Identifying Workgroups in Brazilian Scientific Social Networks.' *Journal of Universal Computer Science* **Vol. 17, No. 14**, 1951-1970.

Ströele, V., Zimbrão, G. and Souza, J. M.: 2011. 'Evaluating Knowledge Flow in Multirelational Scientific Social Networks', in, *The 15th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2011)*, Lausanne, pp. p. 516-523.

Ströele, V., Zimbrão, G. and Souza, J. M.: 2012, 'Modeling, Mining and Analysis of Multi-Relational Scientific Social Network.' *Journal of Universal Computer Science* **(Aceito para o *special issue* de Análise de Redes Sociais do Journal)**.

Taskar, B., Wong, M.-F., Abbeel, P. and Koller, D.: 2004. 'Link prediction in relational data', in, *Proceedings of Neural Information Processing Systems*, pp. 13-18.

Tyler, J. R., Wilkinson, D. M. and Huberman, B. A.: 2003. 'Email as Spectroscopy: Automated Discovery of community Structure Within Organizations', in, *Proceedings of the First International Conference on Communities and Technologies*, M. Huysman,E. Wenger, V. Wulf

Valente, T. W. and Fujimoto, K.: 2010, 'Bridging: Locating critical connectors in a network.' *Social Networks* **32**, 212–220.

Wasserman, S. and Faust, K.: 1994. 'Social Network Analysis: Methods and Applications', in, Cambridge, UK, Cambridge University Press.

Wong, P. C., Foote, H., Chin, G., Mackey, P. and Perrine, K.: 2006, 'Graph Signatures for Visual Analytics.' *IEEE Trans. on Visualization and Computer Graphics* **vol. 12, no. 6**, 1399-1413.

Xie, X. L. and Beni, G.: 1991, 'A validity measure for fuzzy clustering.' *IEEE Transations on Pattern Analisys and Machine Intelligence* **v. 12**, 841-847.

Xu, S., Zhang, J., Han, D. and JieWang: 2006, 'Singular value decomposition based data distortion strategy for privacy protection.' *Journal of Knowledge and Information Systems* **vol. 10(3)**, 383–397.

Zadeh, N.: 1972, 'Theoretical Efficiency of the Edmonds-Karp Algorithm for Computing Maximal Flows.' *Journal of the ACM 19 (1): 184-192. ISSN:0004-5411*.

Zanette, D.: 2002, 'Dynamics of rumor propagation on small-world networks.' *Physical review E* **vol. 65, no. 4**.

Zhou, D. and Scholkopf, B.: 2004. 'A regularization framework for learning from graph data', in, *Proceedings of Workshop on Statistical Relational Learning at International Conference on Machine Learning*, Banff.

Zhu, J.: 2003, *Mining Web Site Link Structures for Adaptive Web Site Navigation and Search*. University of Ulster.

Zhuge, H. and Guo, W.: 2007, 'Virtual knowledge service market—For effective knowledge flow within knowledge grid.' *Journal of Systems and Software* **Volume 80**, 1833-1842.

Anexo A

Questionário para análise dos resultados obtidos pelo método de agrupamento por fluxo máximo.

Qual é a sua instituição? _____

Quais são as suas linhas de pesquisa?

1)_____ 4)_____

2)_____ 5)_____

3)_____

O senhor trabalha com pesquisadores de outras linhas ou laboratórios do seu departamento?

( ) não ( ) sim. Quais linhas?

_____

_____

_____

O senhor costuma publicar resultados de pesquisas com professores de outras universidades?

( ) sim ( ) não

Com que freqüência você publica com professores externos?

( ) nunca ( ) raramente ( ) de vez em quando ( ) sempre

Cite o nome de 5 professores externos (e suas instituições) com os quais você mais publica classificando-os de acordo com a quantidade de publicações feitas em conjunto.

1)_____ 4)_____

2)_____ 5)_____

3)_____

E quanto aos professores da própria instituição? O senhor costuma publicar resultados de pesquisas com eles?

( ) sim ( ) não

Com que freqüência você publica com professores internos?

( ) nunca ( ) raramente ( ) de vez em quando ( ) sempre

Cite o nome de 5 professores internos com os quais você mais publica classificando-os de acordo com a quantidade de publicações feitas em conjunto.

1)_____ 4)_____

2)_____ 5)_____

3)_____

Em termos de publicações, como o senhor classificaria o relacionamento da sua instituição com as instituições abaixo em termos de trabalhos conjuntos e publicações?

UFRJ	( ) inexistente	( ) bom	( ) médio	( )excelente
UFPE	( ) inexistente	( ) bom	( ) médio	( )excelente
UFMG	( ) inexistente	( ) bom	( ) médio	( )excelente
PUC-RJ	( ) inexistente	( ) bom	( ) médio	( )excelente
UFRGS	( ) inexistente	( ) bom	( ) médio	( )excelente

O senhor acredita que a classificação da sua instituição na CAPES possui está relacionada com o nível dos relacionamentos externos que a sua instituição possui?

( ) sim ( ) não ( ) talvez

Muito obrigado pela colaboração!