

Universidade Federal da Paraíba
Centro de Informática
Programa de Pós-Graduação em Informática

Metodologia para Análise de Relevância de Publicações através
de Rede de Citações

Moacir Lopes de Mendonça Junior

Dissertação submetida à Coordenação do Curso de Pós-Graduação em
Informática da Universidade Federal da Paraíba como parte dos requisi-
tos necessários para obtenção do grau de Mestre em Informática.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Computação Distribuída

Alisson Vasconcelos de Brito

João Pessoa, Paraíba, Brasil

©Moacir Lopes de Mendonça Junior, 26 de Fevereiro de 2015

M539m Mendonça Junior, Moacir Lopes de.
Metodologia para análise de relevância de publicações
através de rede de citações / Moacir Lopes de Mendonça
Junior.-- João Pessoa, 2015.
87f. : il.
Orientador: Alisson Vasconcelos Brito
Dissertação (Mestrado) – UFPB/CI
1. Informática. 2. Redes sociais – análise. 3. Redes de
citações – relevância – análise.

UFPB/BC

CDU: 004(043)

Ata da Sessão Pública de Defesa de Dissertação de
Mestrado de **Moacir Lopes de Mendonça Júnior**,
candidato ao título de Mestre em Informática na
Área de Sistemas de Computação, realizada em 13
de fevereiro de 2015.

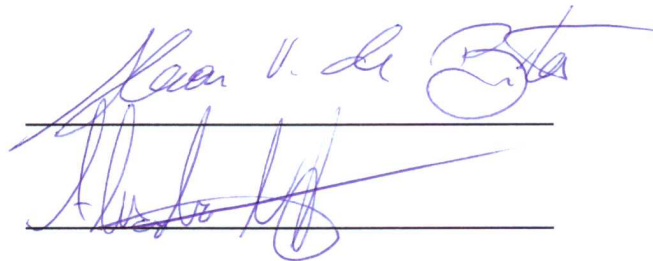
Ao décimo terceiro dia do mês de fevereiro do ano de dois mil e quinze, às quatorze horas, no laboratório 2 da Escola Superior de Redes - Universidade Federal da Paraíba, reuniram-se os membros da Banca Examinadora constituída para examinar o candidato ao grau de Mestre em Informática, na área de “*Sistemas de Computação*”, na linha de pesquisa “*Sinais, Sistemas e Gráficos*”, a **Sr. Moacir Lopes de Mendonça Júnior**. A comissão examinadora foi composta pelos professores doutores: **Alisson Vasconcelos de Brito (PPGI-UFPB)**, orientador e presidente da Banca, **Alexandre Nóbrega Duarte (PPGI-UFPB)**, examinador interno, **Francisco Petrônio Alencar de Medeiros**, examinador externo à Instituição. Dando início aos trabalhos, o professor Alisson Vasconcelos de Brito cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra ao candidato para que o mesmo fizesse, oralmente, a exposição do trabalho de dissertação intitulado “*Metodologia para Análise de Relevância de Publicações através de Rede de Citações*”. Concluída a exposição, o candidato foi arguido pela Banca Examinadora que emitiu o seguinte parecer: “*Aprovado*”. Assim sendo, deve a Universidade Federal da Paraíba expedir o respectivo diploma de Mestre em Informática na forma da lei e, para constar, eu, Tadeá Maria da Silva, técnica em secretariado, lavrei a presente ata que vai assinada por mim e pelos membros da Banca Examinadora. João Pessoa, 13 de fevereiro de 2015.


Tadeá Maria da Silva

Prof. Dr. Alisson Vasconcelos de Brito
Orientador (PPGI-UFPB)

Prof. Dr. Alexandre Nóbrega Duarte
Examinador Interno (PPGI-UFPB)

Prof. Dr. Francisco Petrônio A. de Medeiros
Examinador Externo à Instituição





Resumo

A pesquisa científica é definida como um conjunto de atividades que têm por objetivo a descoberta de novos conhecimentos de interesse para a solução de problemas existentes. O conhecimento científico não se resume na descoberta de fatos e leis novas, mas também em sua publicação. Trata-se de obter e comunicar resultados através de formas de produção científica. Visando compreender o complexo sistema de produção científica inúmeras técnicas de análise vêm sendo aplicadas em repositórios digitais. Essas técnicas objetivam a identificação de produções científicas relevantes baseadas em indicadores. Neste sentido o presente trabalho de pesquisa tem como objetivo criar uma metodologia que facilite identificar publicações que são relevantes a determinadas áreas de conhecimento. Para isto utilizamos três contextos, o primeiro sobre a área de análise de redes sociais, o segundo sobre os anais dos 20 anos do *webmedia* e por último sobre área de sistemas embarcados. Como fonte de estudo nossa ferramenta de software conseguiu diversos resultados referentes às áreas de conhecimento, entre eles a evolução de publicações para os veículos de publicação mais relevantes, a classificação das comunidades pertencentes a rede, e entre outros resultados.

Palavras-chave: Análise de Redes Sociais, Redes de Citações, Relevância, Análise.

Abstract

Scientific research is defined as a set of activities that are aimed at the discovery of new knowledge relevant to the solution of existing problems. Scientific knowledge is not just the discovery of new facts and laws, but also in its publication. This is to obtain and communicate results through forms of scientific production. To understand the complex scientific production system numerous analytical techniques have been applied in digital repositories. These techniques aim to identify relevant scientific productions based on indicators. In this sense the present work aims to create a methodology that facilitate the identification of publications that are relevant to specific areas of expertise. For this we used three contexts, the first on the area of social network analysis, the second on the records of 20 years from webmedia and finally on area of embedded systems. As a source of study our software tool achieved several results for the areas of knowledge between them we can mention the development of publications for the most relevant publication vehicles and the classification of network communities.

Keywords: Social Network Analysis, Citation Network, Relevance, Analysis

Agradecimentos

Agradeço primeiramente a Deus pela minha existência, pelas bênçãos que me tem concedido ao longo de toda minha vida, por guiar-me e acompanhar-me diariamente e, por incentivar-me na conclusão de mais uma etapa.

A minha mãe Valéria Marques Lopes de Mendonça, pelo exemplo de dignidade e bravura dados a cada dia, sem você, nada disso estaria acontecendo. Amo você!

A minha avó Maria José de Brito (In Memoriam), não foram em vão todos os ensinamentos! Para sempre estarão guardados em meu coração. Eternamente grato!

A todos integrantes do Grupo de Trabalho LabSNA os quais me acompanharam durante grande parte do meu trabalho de mestrado. Em especial agradeço ao professor Alexandre Duarte e aos alunas Jarbele e Ayslânya pelos momentos de estudo, pela compreensão e por todo apoio e ajuda nesta jornada.

A meus amigos Danyllo Wagner, André Assad e Taína Jesus agradeço por toda ajuda e tempo disponíveis na revisão deste trabalho .

Aos professores, Andrei Formiga, Álvaro Medeiros, Lucídio Formiga pela paciência e competência na transmissão dos conhecimentos.

Ao meu orientador e amigo Alisson Vasconcelos de Brito, pelo exemplo de profissionalismo, pela paciência, apoio e contribuição para realização deste trabalho, Muito Obrigado!

Conteúdo

1	Introdução	1
1.1	Motivação	2
1.2	Metodologia	3
1.3	Objetivos	3
1.4	Questões de Pesquisa	4
1.5	Contribuição	7
1.6	Estrutura da Dissertação	7
2	Fundamentação Teórica	9
2.1	Publicações científicas	9
2.1.1	Definições e Termos Gerais	9
2.1.2	Repositórios digitais de publicações	10
2.2	Redes Sociais	10
2.2.1	Definições e Termos Gerais	10
2.2.2	Teoria dos grafos	11
2.2.3	Rede de citações	13
2.2.4	Análise de Redes Sociais	14
2.3	Métricas para análise de redes sociais	15
2.3.1	Centralidade de Grau	15
2.3.2	Centralidade de Intermediação	16
2.3.3	<i>Page Rank</i>	17
2.3.4	Modularidade	18
2.4	Propriedades da Rede	20
2.4.1	Componente Gigante	20

2.4.2	Clique	20
3	Trabalhos relacionados	22
3.1	<i>Coauthorship Networks and Patterns of Scientific Collaboration</i>	22
3.2	<i>SUCUPIRA: A system for Information extraction of the Lattes Platform to identify academic social networks</i>	23
3.3	Minerando publicações científicas para análise da colaboração em comunidades de pesquisa	23
3.4	Um Estudo sobre os Impactos dos Relacionamentos Sociais na Avaliação do Mérito Científico	24
3.5	Análise da rede de coautoria do simpósio brasileiro de bancos de dados . .	24
3.6	<i>Characterizing and Predicting Community Members from Evolutionary and Heterogeneous Networks</i>	25
3.7	Considerações	25
4	Procedimentos Metodológicos	27
4.1	Processo para criação e análise da rede de citações	27
4.2	Criação do suporte automatizado	29
4.3	Ferramentas e tecnologias	31
4.3.1	Java	31
4.3.2	Gephi	32
4.3.3	Gephi toolkit	32
5	Estudos de caso	34
5.1	Análise de redes sociais	34
5.1.1	Análise quantitativa da área	35
5.1.2	Análise da Rede de Citações	40
5.1.3	Considerações	43
5.2	WebMedia	43
5.2.1	Análise dos Anais do <i>WebMedia</i>	45
5.2.2	Considerações	48
5.3	Sistemas Embarcados	49

5.3.1	Análise quantitativa da área	49
5.3.2	Análise da Rede de Citações	54
5.3.3	Classificação das Comunidades	56
5.3.4	Considerações	62
5.4	Publicações	62
6	Considerações Finais	64
6.1	Ameaças à validade	64
6.2	Discussão	66
6.3	Trabalhos Futuros	67
	Bibliografia	74

Lista de Símbolos

API Application Programming Interface

HTML HyperText Markup Language

GEXF Graph Exchange XML Format

JAR Java ARchive

ARS Análise de Redes Sociais

Lista de Figuras

2.1	Representação semi-abstrata das 7 pontes da Prússia	12
2.2	Grafo representando as 7 pontes da Prússia	12
2.3	Grafo representando uma rede de Citações	14
2.4	Equação da Centralidade de Grau	15
2.5	Cálculo do grau de saída (a) e grau de entrada (b).	16
2.6	Nós com maior grau de intermediação	17
2.7	Equação da Centralidade de Intermediação	17
2.8	Equação do <i>Page Rank</i>	18
2.9	Grafo representando 3 comunidades determinadas pela modularidade . . .	19
2.10	Rede com três cliques	21
4.1	Processo Metodológico	28
4.2	Paper Crawler	30
4.3	Gephi toolkit	33
5.1	Número de Publicações ao longo dos anos	36
5.2	Meios de Publicação Mais Relevantes	37
5.3	Número de Publicações ao longo dos anos para a área de ARS	38
5.4	Meios de publicação mais relevantes	39
5.5	Distribuição de publicações por citação	41
5.6	Evolução da rede de coautorias do <i>WebMedia</i>	45
5.7	Componente Gigante da rede de 1995 a 2012.	46
5.8	Relação entre autores e artigos	48
5.9	Crescimento de publicações e autores do <i>WebMedia</i>	48
5.10	Evolução ao longo dos anos	50

5.11 Meios de Publicação	51
5.12 Evolução de Publicações dos Meios de Publicação	52
5.13 Tendência dos meios de publicação mais relevantes	53
5.14 Rede classificada por Grau de entrada	54
5.15 Distribuição de publicações	55
5.16 Rede classificada por Page Rank	55
5.17 5 maiores comunidades	57
5.18 Evolução de Publicações das 5 maiores comunidades	58
5.19 Tendência das comunidades mais relevantes	59

Lista de Tabelas

1.1	Hipóteses	5
4.1	Tecnologias utilizadas na construção da arquitetura	29
5.1	Publicações mais relevantes	40
5.2	Quantidade de Citações das publicações mais relevantes	42
5.3	<i>Page Rank</i> das publicações mais relevantes	42
5.4	Estatísticas referentes a rede de coautoria	47
5.5	Publicações mais relevantes	56
5.6	5 maiores comunidades	56
5.7	Média das Métricas para as 5 comunidades mais relevantes	57
5.8	5 <i>keywords</i> mais relevantes para comunidade A	60
5.9	5 <i>keywords</i> mais relevantes para comunidade B	60
5.10	5 <i>keywords</i> mais relevantes para comunidade C	60
5.11	5 <i>keywords</i> mais relevantes para comunidade D	60
5.12	5 <i>keywords</i> mais relevantes para comunidade E	60
5.13	5 publicações mais relevantes da comunidade A - <i>Costs</i>	61
5.14	5 publicações mais relevantes da comunidade B - <i>Energy Consumption</i>	61
5.15	5 publicações mais relevantes da comunidade C - <i>Computer architecture</i>	61
5.16	5 publicações mais relevantes da comunidade D - <i>Real time systems</i>	61
5.17	5 publicações mais relevantes da comunidade E - <i>Processor scheduling</i>	62

Capítulo 1

Introdução

A pesquisa científica é definida como um conjunto de atividades que têm por objetivo a descoberta de novos conhecimentos de interesse para a solução de problemas existentes. Ela é crucial para que ocorra o desenvolvimento, oferecendo novos subsídios que irão permitir indicar soluções adequadas para os problemas detectados, pela utilização de novas técnicas desenvolvidas.

Embora a realização da pesquisa científica seja muito importante, igualmente necessária é a sistematização e a divulgação destes conhecimentos produzidos. Assim, o conhecimento científico não se resume na descoberta de fatos e leis novas, mas também em sua publicação. Trata-se de obter e comunicar resultados através de artigos científicos, relatórios técnicos ou outras formas de produção científica. Muito esforço tem sido realizado pela comunidade científica em criar e manter diversos repositórios digitais para o armazenamento da produção científica. Entre esses repositórios podemos citar o *IEEEExplore*, o *ACM Digital Library* e *Science Direct*.

Visando compreender o complexo sistema de produção científica, inúmeras técnicas de análise vêm sendo aplicadas em repositórios digitais. É importante mencionar que os repositórios digitais armazenam as produções científicas de forma estruturada, facilitando dessa forma a aplicação de técnicas (semi-)automatizadas para recuperação de informação. Essas técnicas objetivam prioritariamente a identificação de produções científicas relevantes baseadas em indicadores quantitativos. Entre elas podemos citar estudos de produtividade científica como Procópio, Laender e Moro (2011) e o de Newman (2004) que analisam uma série de estatísticas e indicadores para verificar a produtividade científica de determinada

área ou evento, tais como a média de artigos por autor, média de artigos por edição, média de coautores por artigo, entre outros. Adicionalmente, podemos citar estudos sobre análise de impacto como Guedes e Duarte (2013), Cervi, Galante e Oliveira (2011) e Patsopoulos, Anatalatos e Ioannidis (2005) onde estes procuram analisar a relevância de publicações, citações ou autores para determinada comunidade científica, evento ou área de conhecimento.

Baseado nas técnicas de análise do estado da arte, pesquisadores que desejam obter informações eficazes a respeito de publicações científicas, não conseguem obter ou mesmo conseguem de forma limitada. Como consequência, a identificação de conteúdo relevante (e.g. autores ou publicações científicas relevantes em um determinado tema) é comprometida.

1.1 Motivação

No início da presente pesquisa, um mapeamento sistemático sobre a área de Análise de Redes Sociais foi realizado com objetivo de identificar métodos e técnicas para construção da pesquisa. Durante o desenvolvimento do mapeamento sistemático, encontrou-se uma série de problemas para obtenção e análise dos dados das publicações associadas à área. Nós devemos mencionar que essa atividade pode ser considerada um processo não repetível, já que depende da avaliação humana e como dependendo da experiência de cada pesquisador que está participando de uma pesquisa destas, o resultado final pode influenciado, tornando este propenso a erros e atualmente não suportada por técnicas (semi-)automatizadas. Isso torna a identificação de conteúdo relevante (e.g. através de mapeamento sistemático ou revisão sistemática) custosa em termos de tempo e de difícil replicabilidade por parte de outros pesquisadores que desejam realizar essa atividade. Em face a essa problemática, propomos a criação de uma metodologia com suporte automatizado que permita a identificação de conteúdo relevante, relacionado a publicações científicas armazenadas em repositórios de dados digitais. Assim, nossa metodologia apresenta um processo formal de extração e análise dos dados, baseado em técnicas de análise de redes sociais, tornando a identificação de conteúdo relevante menos sujeito a subjetividade da análise humana. Finalmente, acreditamos que nossa metodologia servirá de suporte a qualquer pesquisador que tenha interesse em obter de modo rápido e preciso informações relevantes sobre uma determinada área de pesquisa.

1.2 Metodologia

A nossa questão geral de pesquisa é “Pode o uso de técnicas de análise de redes sociais auxiliarem a mineração de conteúdo relevante associado a publicações científicas?”. Nesse contexto temos o interesse de analisar as publicações científicas a partir da criação de uma rede de citações. Para suportar a nossa questão de pesquisa, alguns procedimentos metodológicos foram aplicados no contexto dessa pesquisa. Inicialmente, realizamos uma revisão da literatura com objetivo de identificar métodos e técnicas para construção de redes de citações e identificação publicações relevantes. Em seguida, buscamos na literatura atual formas de análise da rede de citações baseadas em técnicas de análise de redes sociais. Após a realização desses dois passos, identificamos um modo formal de extração e análise de dados de uma rede de citações. Nosso próximo passo foi a definição de um suporte automatizado para criação e análise de uma rede de citações. Finalmente, nossa metodologia foi avaliada e validada através dos seguintes estudos de caso: (i) análise da relevância de publicações e meios de publicação da área de análise de redes sociais, (ii) identificação de pesquisadores mais relevantes no contexto do *webmedia* e (iii) identificação da relevância de publicações, meios de publicação e comunidades da área de conhecimento de Sistemas Embarcados.

1.3 Objetivos

O presente estudo tem como principal objetivo a criação de uma metodologia para análise de relevância de publicações através de rede de citações utilizando técnicas de análise de redes sociais. Para atingirmos o objetivo geral de pesquisa, diversas atividades são de fundamental importância e são as seguintes:

- Realização de uma revisão da literatura sobre redes sociais e teoria dos grafos, com objetivo de identificar formas de representação e construção de redes de citações;
- Identificação de métricas de análise de redes sociais relevantes visando a análise dos componentes de uma rede de citações. As métricas de principal interesse dessa pesquisa são aquelas que fornecem suporte a identificação de: (i) grau de relevância de determinada publicação, (ii) autores de publicação relevantes, (iii) comunidades científicas de uma determinada área, (iv) classificação de comunidades (v) meios de

publicação relevantes;

- Definição de um processo formal para criação e análise de uma rede de citações no contexto de uma análise de relevância de publicações;
- Desenvolvimento de suporte automatizado para criação e análise de uma rede de citações no contexto de uma análise de relevância de publicações;
- Definição de um *framework* de avaliação e validação da metodologia para análise de relevância de publicações através da realização de estudos de caso em repositórios de publicações científicas;

1.4 Questões de Pesquisa

Com o objetivo de fornecer suporte a nossa questão geral de pesquisa (Seção 1.2), no contexto da presente pesquisa foram definidos duas atividades principais: (i) validar o processo formal para criação e análise de uma rede de citações no contexto de análise de publicações relevantes; e (ii) validar o suporte automatizado para criação e análise de uma rede de citações.

A seguir apresentaremos alguns conceitos que são de suma importância para definição das hipóteses associadas a presente pesquisa. De acordo com Newman (2010), uma rede social pode ser definida como um conjunto de unidades (e.g. pessoas, artigos, computadores) interligadas de acordo com algum critério. As redes sociais podem ser modeladas através de grafos onde as unidades (bem como seus atributos) são representadas através dos nós e as ligações entre as unidades são representadas através de arestas. Assim, no contexto de publicações científicas podemos criar as chamadas redes de citações. Nas redes de citações os nós são formados pelos artigos científicos e seus atributos (e.g. nome do autor, meio de publicação, ano de publicação) e as arestas unem os artigos através do critério de citação. Por exemplo, se um artigo A cita um determinado artigo B em suas referencias, deverá ser criada uma aresta ligando o artigo A ao Artigo B.

Assim, no contexto da presente pesquisa definimos que é possível: (i) criar uma rede de citações a partir dos conceitos de redes sociais; (ii) avaliar aspectos quantitativos e qualitativos associados aos componentes da rede de citações através do uso de métricas de redes

sociais. Nesse sentido, definimos cinco hipóteses (H) para atingir os objetivos acima mencionados. Todas as hipóteses associadas ao presente trabalho estão descritas na Tabela 1.1 a seguir.

Tabela 1.1: Hipóteses

H	Descrição
H1	A metodologia baseada em redes sociais identifica de forma eficaz meios de publicação que são relevantes para uma determinada área de pesquisa
H2	A metodologia baseada em redes sociais identifica de forma eficaz publicações científicas relevantes no contexto de uma determinada área de pesquisa
H3	As métricas de centralidade são bons indicadores para identificação de publicações científicas relevantes no contexto de uma determinada área de pesquisa
H4	As métricas de centralidade são bons indicadores para análise de evolução de comunidades científicas no contexto de um determinado evento
H5	A metodologia baseada em redes sociais identifica e classifica de forma eficaz comunidades de rede relevantes no contexto de uma determinada área de pesquisa

A **primeira hipótese (H1)** foi definida pois supomos que a partir da criação da rede de citações (utilizando redes sociais) e sua posterior análise (utilizando métricas de redes sociais), podemos identificar meios de publicação relevantes no contexto de uma determinada área de pesquisa. A seguir demonstramos como verificar a veracidade de H1:

Verificação da H1 - Para verificar se a primeira hipótese é verdadeira vamos classificar os meios de publicação utilizando o número de publicações (i.e grau). Essa métrica utiliza apenas aspectos quantitativos no cálculo dos seus valores. Caso através do número de publicações consigamos apresentar uma classificação dos meios de publicação, temos que a primeira hipótese é verdadeira.

A **segunda hipótese (H2)** foi definida pois supomos que a partir da criação da rede de citações (utilizando redes sociais) e sua posterior análise (utilizando métricas de redes sociais), podemos identificar publicações científicas relevantes no contexto de uma determinada área de pesquisa. A **terceira hipótese (H3)** foi definida de modo complementar a segunda hipótese. Aqui supomos que as métricas de centralidade são mais adequadas para identificação de publicações científicas relevantes, pois estas utilizam aspectos qualitativos para o cálculo

dos valores associados a métrica. A seguir demonstramos como verificar a veracidade de H2 e H3:

Verificação da H2 - Para verificar se a segunda hipótese é verdadeira vamos criar a rede de citações e classificar os nós (i.e. publicação científicas) utilizando métricas de redes sociais: Grau de Entrada e *Page Rank*. A primeira métrica utiliza apenas aspectos quantitativos no cálculo dos seus valores. Em contraste, a segunda métrica utiliza aspectos qualitativos no cálculo dos seus valores. Caso possamos criar a classificação das publicações científicas utilizando as duas métricas, temos que a segunda hipótese é verdadeira.

Verificação da H3 - Para verificar se a terceira hipótese é verdadeira vamos classificar as publicações científicas utilizando duas métricas de redes sociais: Grau de entrada e *Page Rank*. Em seguida, iremos analisar as duas métricas. Caso os resultados nós dê que o *page rank* ofereça melhores resultados, temos que a terceira hipótese é verdadeira.

A **quarta hipótese (H4)** foi definida, pois supomos que as métricas de centralidade são bons indicadores para análise da evolução de comunidades científicas de eventos ou periódicos, pois estas utilizam aspectos qualitativos para o cálculo dos valores associados a métrica. No contexto de comunidades científicas os aspectos qualitativos podem ser considerados melhores indicadores para definição da relevância. A seguir demonstramos como verificar a veracidade de H4:

Verificação da H4 - Para verificar se a quarta hipótese é verdadeira vamos criar a rede de coautoria e classificar os nós (i.e. autores) utilizando métricas e propriedades de redes sociais: Grau, Centralidade de Intermediação e Componente Gigante. Caso possamos analisar a evolução da produção científica da comunidade científica em questão utilizando tais métricas, temos que a quarta hipótese é verdadeira.

A **quinta hipótese (H5)** foi definida pois supomos que a partir da criação da rede de citações (utilizando redes sociais) e sua posterior análise (utilizando métricas de redes sociais), podemos identificar e nomear comunidades relevantes dentro da rede no contexto de uma determinada área de pesquisa. A seguir demonstramos como verificar a veracidade de H5:

Verificação da H5 - Para verificar se a quarta hipótese é verdadeira vamos criar a rede de citações e particionar os nós (i.e. publicações) utilizando métricas de redes sociais: Modularidade. Em seguida, iremos identificar e nomear as comunidades relevantes da rede, utilizando como fonte as palavras-chave de cada publicação referentes a essas comunidades.

Caso possamos classificar as comunidades da rede utilizando tal métrica, temos que a quinta hipótese é verdadeira.

1.5 Contribuição

A presente pesquisa traz como principais contribuições: (i) a definição de um processo formal para criação e análise de redes de citações, (ii) a criação de suporte automatizado para criação e análise da rede de citações e (iii) definição de um *framework* de avaliação da metodologia para análise de relevância de publicações científicas através de redes de citações.

Através da definição do processo formal para criação e análise de uma rede de citações podemos tornar os processos como o mapeamento sistemático ou revisão sistemática repetíveis, menos propenso a erros e menos dependente da avaliação humana para identificação de conteúdo relevante. A partir da criação do suporte automatizado para criação e análise da rede de citações, esperamos que os resultados possam ser obtidos em um menor espaço de tempo e com menos esforço em termos da necessidade da avaliação humana. Finalmente, através da definição do *framework* de avaliação da metodologia para análise de relevância de publicações científicas esperamos prover para a comunidade científica meios adequados para avaliação do processo formal e o suporte automatizado para criação e análise de redes de citações.

Com a realização dessa pesquisa pretendemos avançar o estado da arte relativo a identificação de conteúdo relevante associado a publicações científicas. Por fim, a discussão do processo formal de criação da rede de citação, do seu suporte automatizado, das tecnologias e ferramentas envolvidas na sua construção possam estimular pesquisas científicas para o desenvolvimento de novos meios e métodos que possam ser utilizados para este fim, bem como podendo ser apontadas melhorias nos que foram utilizados durante este trabalho.

1.6 Estrutura da Dissertação

Este trabalho encontra-se dividido em seis capítulos. O Capítulo 1 apresenta o trabalho de maneira geral descrevendo sua motivação, objetivos, contribuição e estrutura da dissertação. O capítulo 2 apresenta os conceitos e definições relacionados a publicações científicas, redes

sociais e teoria dos grafos, bem como a análise de redes sociais através de métricas e propriedades da rede. Esses conceitos são empregados para a modelagem da rede de citações. No Capítulo 3 descreve-se os trabalhos relacionados utilizados para embasar esta pesquisa. No Capítulo 4 descreve a metodologia empregada para identificação de conteúdo relevante associado a publicações científicas. O Capítulo 5 apresenta os estudos de caso realizados no contexto da presente pesquisa. Tais estudos de caso fornecem suporte para avaliar as hipóteses bem como responder as questões de pesquisa do presente trabalho. Finalmente, no Capítulo 6, elencamos as ameaças a validade, as considerações finais do estudo bem como apontamento para trabalhos futuros na área de pesquisa.

Capítulo 2

Fundamentação Teórica

O presente capítulo pretende modelar o referencial teórico necessário para o entendimento da pesquisa. Por ser um assunto amplo e de domínio de várias áreas de conhecimento, procuramos deixar explícito apenas os principais conceitos utilizados. Temos particular interesse em definir como criar uma rede de citações a partir dos conceitos de redes sociais. Adicionalmente, descreveremos os meios para avaliação dos aspectos quantitativos associados aos componentes da rede de citações através do uso de métricas de redes sociais.

Iniciamos o capítulo apresentando os principais conceitos sobre publicações científicas e os principais repositórios digitais (Seção 2.1). Em seguida apresentaremos os conceitos associados a redes sociais e teoria dos grafos, bem como a forma de criação de redes de citações (Seção 2.2). Já na seção Métricas (Seção 2.3) descrevemos as principais métricas de análise de redes sociais que foram utilizadas para análise da relevância das publicações e definição das comunidades científicas na rede de citações. Finalmente, na seção Propriedades da Rede (Seção 2.4) definimos propriedades utilizadas durante a geração da rede: componente gigante e clique.

2.1 Publicações científicas

2.1.1 Definições e Termos Gerais

O processo do conhecimento científico estabelece duas etapas essenciais. A primeira é a produção do conhecimento, onde este envolve a elaboração do problema, a seleção de uma

metodologia adequada, a condução da investigação e a interpretação dos resultados obtidos. A segunda é a divulgação do conhecimento, que é realizada por através da publicação em algum veículo de publicação. Ao ser enviado a um periódico, o manuscrito é submetido à revisão, mas sua aprovação só se dá caso atenda critérios mínimos de excelência (ABREU-RODRIGUES, 2009).

O número de artigos submetidos a publicação vem crescendo no país, isto se deve à mudanças que foram realizadas no sistema de avaliação de programas de pós-graduação adotado pela CAPES, onde este passou a priorizar a produtividade. Assim, quanto maior o número de citações, maior a influência do conhecimento publicado e, consequentemente, maior a sua relevância (ABREU-RODRIGUES, 2009).

2.1.2 Repositórios digitais de publicações

Os repositórios digitais de pesquisa são bases de dados disponíveis na Web que reúnem de maneira estruturada a produção científica. A partir disso, oferecem uma série de benefícios tanto para os pesquisadores quanto às instituições ou sociedades científicas proporcionando uma maior visibilidade aos resultados de pesquisas e possibilitando a preservação da memória científica. Estes repositórios pode ser considerados elementos de uma rede ou infraestrutura informacional de um país ou de um domínio institucional destinados a garantir a guarda, preservação a longo prazo e, fundamentalmente, o livre acesso à produção científica (CAMARGO; VIDOTTI, 2006).

2.2 Redes Sociais

2.2.1 Definições e Termos Gerais

Existe uma abundância de redes pelo mundo, os exemplos incluem desde a Internet (rede mundial de computadores), redes sociais online, conexões entre quaisquer indivíduos até redes de conexão comercial, redes metabólicas, redes de distribuição tais como energia elétrica, serviços telefônicos, serviços de correio, entre outros (NEWMAN, 2003). Uma rede, em sua forma mais simples, pode ser definida como um conjunto de pontos que às vezes chamaremos de nós, interligados por conexões denominadas arestas (NEWMAN, 2003). Esta

morfologia permite identificar as relações entre pares de nós que podem desde páginas web, pessoas, computadores, até mesmo cidades, países ou aeroportos.

Um tipo específico de rede é de extremo interesse ao presente trabalho: rede social, que é definida como qualquer conjunto limitado de unidades sociais interligadas (BERKOWITZ, 1982). Esta definição destaca três características importantes de redes sociais: Primeiro, as redes têm limites, isto é, existe algum critério para determinar a associação na rede. O segundo elemento chave da definição é "conexão" em redes sociais. Para fazer parte de uma rede social, cada membro deve ter ligações reais ou potenciais para pelo menos outro membro da rede. Esses links podem ser diretos ou indiretos. O terceiro aspecto chave da presente definição é uma unidade social, ou seja, cada componente da rede é um elemento distinto do grupo.

A noção de rede social e métodos para análise de redes sociais têm atraído consideravelmente a comunidade das ciências sociais e comportamentais a fim de analisar os relacionamentos entre entidades sociais e os padrões e implicações desses relacionamentos (WASSERMAN; FAUST, 1994). Para estudar como os comportamentos ou as opiniões dos indivíduos dependem das estruturas nas quais eles se inserem, as unidades de análise deixam de ser os atributos individuais (classe, sexo, idade), passando a ser o conjunto de relações que os indivíduos estabelecem através de suas interações uns com os outros através da rede (MARTELETO, 1994).

2.2.2 Teoria dos grafos

Diversas situações do mundo real podem ser convenientemente modeladas através de uma rede que são representadas através dos grafos. Esta estrutura é composta por um conjunto de pontos, juntamente com linhas que unem alguns pares de pontos. No caso específico das redes sociais temos que os pontos podem representar pessoas e as linhas podem representar uma ligação de amizade, ou ainda os pontos podem ser centros de comunicação e as linhas podem representar o meio de comunicação entre eles (Mika, 2007). Segundo Crilly (2007), a origem da teoria dos grafos moderna se deu da solução de um problema encontrado em Königsberg, uma cidade da Prússia oriental famosa por suas 7 pontes que atravessam o rio Pregel, onde ela e suas pontes são relacionadas ao famoso matemático Leonhard Euler. Na Figura 2.1 observamos que a ilha no meio do rio foi chamada de I e as margens do rio de A,

B e C.

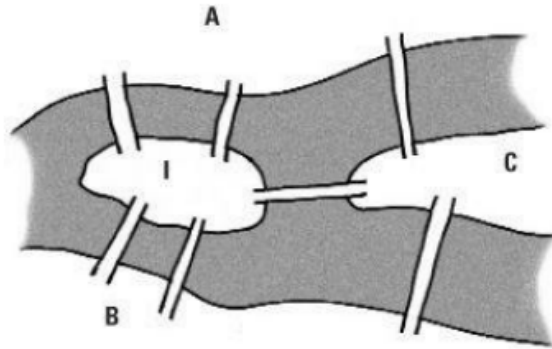


Figura 2.1: Representação semi-abstrata das 7 pontes da Prússia
Fonte: (CRILLY, 2007)

Em 1735, Euler apresentou uma solução a Academia Russa onde queria provar se era possível atravessar cada ponte do rio apenas uma vez. Em seu teorema o passo chave foi tornar a Figura 2.1 mais abstrata, obtendo como resultado um grafo de linhas e pontos, onde o terreno é representado por pontos e as pontes por linhas, ilustrado na Figura 2.2.

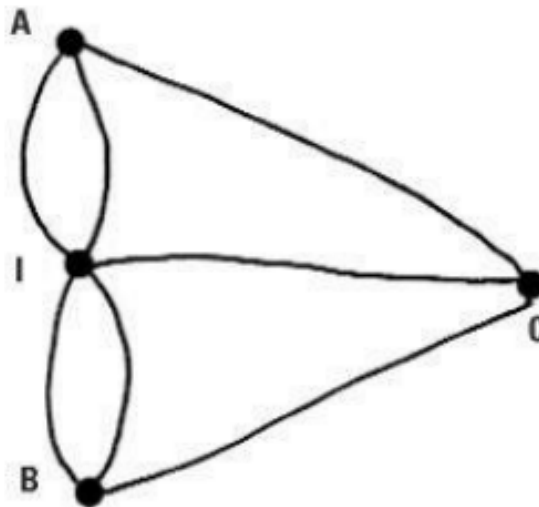


Figura 2.2: Grafo representando as 7 pontes da Prússia
Fonte: (CRILLY, 2007)

Euler fez uma observação sobre a caminhada de sucesso. Excluindo o início e o final da caminhada, toda vez que uma ponte é atravessada até uma porção de terra deve ser possível sair desta na ponte que não foi atravessada anteriormente. Traduzindo este pensamento em uma figura abstrata representado na Figura 2.2, podemos dizer que o encontro de linhas com

um ponto deve ocorrer em pares. Excluindo dois pontos que representam o início e término da caminhada, as pontes podem ser atravessadas se, somente se, cada ponto tem um número par de linhas incidindo nelas. O número de linhas incidindo no ponto é chamado de “grau” do ponto.

O Teorema de Euler afirma que as pontes de uma cidade podem ser atravessadas exatamente uma vez se, excluindo, no máximo, dois, todos os pontos têm mesmo grau.

Logo, um grafo pode ser considerado como um conjunto de pontos, denominados de vértices interligados por linhas, denominadas de arestas (SÁ; ROCHA, 2010). Segundo Ramos, Neto e Vega (2009) um grafo é um par ordenado (V, A) , onde V é o conjunto de vértices deste grafo e A é o conjunto de arcos que denotam uma relação binária com V . Onde os arcos indicam associação entre os vértices.

A simplicidade de tal conceito torna claro que ele permite modelar situações concretas muito variadas de que são exemplo as redes de transporte, de comunicações, de citações, de coautoria, etc. Mas apesar desta simplicidade, o grafo se torna um modelo interessante, pois permite transportar informação suficiente para o esclarecimento ou resolução de uma situação modelada utilizando resultados e algoritmos da teoria dos grafos (SÁ; ROCHA, 2010).

2.2.3 Rede de citações

De acordo com Newman (2010), uma rede social pode ser definida como um conjunto de unidades (e.g. pessoas, artigos, computadores) interligadas de acordo com algum critério. As redes sociais podem ser modeladas através de grafos onde as unidades (bem como seus atributos) são representadas através dos nós e as ligações entre as unidades são representadas através de arestas. Assim, no contexto de publicações científicas podemos criar as chamadas redes de citações. Nas redes de citações os nós são formados pelos artigos científicos e seus atributos (e.g. título, nome do autor, meio de publicação, ano de publicação) e as arestas unem os artigos através do critério de citação. Por exemplo, se um artigo A cita um determinado artigo B em suas referencias, deverá ser criada uma aresta ligando o artigo A ao Artigo B. Na Figura 2.3 podemos observar um exemplo de uma rede de citações.

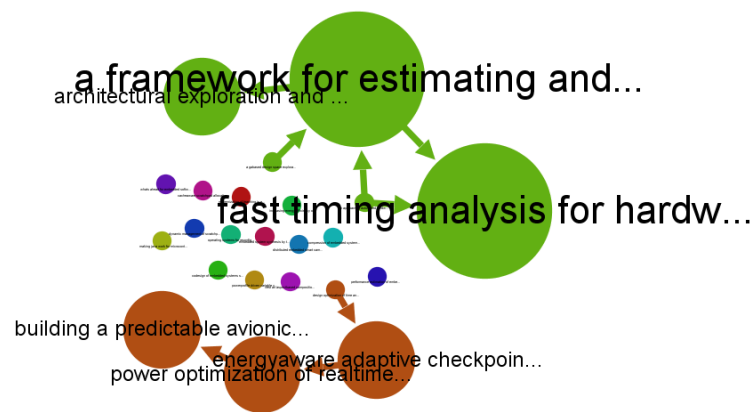


Figura 2.3: Grafo representando uma rede de Citações

Há muitas razões do porque de uma publicação citar outra, para apontar informações que pode ser interessante ao leitor, dar credito a um trabalho, para indicar influência no trabalho atual, ou para discordar do conteúdo do artigo. Então geralmente se uma publicação cita outra é por este conteúdo é relevante para a publicação atual (NEWMAN, 2010).

2.2.4 Análise de Redes Sociais

A Análise de Redes Sociais (ARS) é baseada na descrição formal das redes através de estruturas denominadas grafos. Através da ARS, é possível compreender e acompanhar de forma mais eficaz a disseminação de informações e a interação entre as pessoas que compõem a rede. Tal acompanhamento pode auxiliar na identificação de conectores de redes isoladas, melhorar a atuação de atores críticos, identificar a fragilidade da rede em relação à comunicação dos membros, dentre outras. Ela pode ser facilmente aplicada a uma vasta gama de unidades sociais, como indivíduos por exemplo, como é o caso das redes de apoio social. Mas eles também podem ser agências de serviços sociais, instituições sociais nas comunidades locais, ou nações na economia global. Em uma profissão diversa como o trabalho social, análise de redes sociais tem aplicações diretas para o estudo da prática clínica, a análise da política social, organização comunitária e gestão organizacional (WASSERMAN; FAUST, 1994).

De acordo com Scott e Carrington (2011), a análise de redes sociais tem como ponto de partida a premissa de que a vida social é criada principalmente por suas relações e os

padrões formados por elas. Redes sociais são formalmente definidas como um conjunto de nós que são atados por um ou mais tipos de relações. A Definição de que nós incluir em uma análise de rede frequentemente representa um desafio precoce. Um estudioso ao desejar analisar pesquisadores médicos que estudam doenças do coração tem que levar em conta que pessoas devem ser consideradas como pesquisadores dessa área, pois isso pode ser traiçoeiro. Especialmente porque muitos analistas de rede evitam abordagens baseadas em grupos para a compreensão do mundo social. Após identificar os membros da rede, eles devem identificar as relações entre os nós. Estas podem incluir colaborações, amizades, laços comerciais, links web, citações, fluxo de recursos, troca de apoio social ou qualquer outra ligação possível entre essas unidades especiais (SCOTT; CARRINGTON, 2011).

2.3 Métricas para análise de redes sociais

Nas seções seguintes serão descritas algumas métricas para estudo de redes que são de interesse do presente trabalho. Além das métricas supracitadas, outras foram propostas e podem ser aplicadas em face às características da rede a ser analisada. Uma lista mais extensa de tais métricas pode ser encontrada no trabalho de Newman (2003).

2.3.1 Centralidade de Grau

A centralidade de grau é definida como o número de ligações que incidem sobre o nó. A centralidade de grau de um vértice v_i , para um dado grafo $G := (V, A)$ com $|V|$ vértices e $|A|$ arestas está definido na Figura 2.4

$$C_D(v_i) = \deg(v_i) = \sum_{j=1}^g x_{ij} = \sum_{j=1}^g x_{ji}$$

Figura 2.4: Equação da Centralidade de Grau
Fonte: (NEWMAN, 2010)

Na equação da Figura 2.4, poderão ser utilizados os valores das linhas i de uma matriz de conectividade, ou das colunas j da mesma matriz, para o cálculo do grau de centralidade de um nó. O valor x_{ij} ou x_{ji} , indica posição de linha/coluna (ij) ou coluna/linha (ji) de

uma matriz de conectividade, onde esta é uma matriz binária de uma rede direcionada, em que as linhas representam os nós de partida e as colunas, os nós de chegada. O valor x igual a 1, indica que existe um enlace entre os nós i e j . A soma dos valores de x_{ij} ou x_{ji} indica o valor do grau do nó. O valor g representa o número total de nós ou de linhas/colunas da matriz de adjacências.

Em uma rede direcionada, são definidas duas medidas distintas de centralidade de grau, o grau de entrada e o grau de saída. Onde, o grau de entrada é a contagem do número de arestas direcionadas para o nó e o grau de saída é o número de arestas que saem do nó para outros. Quando as arestas estão associadas a alguns aspetos positivos, como amizade ou colaboração, o grau de entrada é muitas vezes interpretado como uma forma de popularidade, e o grau de saída como sociável (NEWMAN, 2010). Considerando o grau de entrada, $dI(v_i)$, e o grau de saída, $dO(v_i)$, de acordo com a direção dos arcos que chegam ou partem do nó. Podemos obter distintamente o grau de saída e o grau de entrada conforme as equações da Figura 2.3.1 respectivamente.

$$C'_{Do}(v_i) = \frac{dO(v_i)}{g-1} \quad C'_{Di}(v_i) = \frac{dI(v_i)}{g-1}$$

(a)
(b)

Figura 2.5: Cálculo do grau de saída (a) e grau de entrada (b).

Cabe a ressalva de que os cálculos para a métrica centralidade de grau descritos nas equações da Figura 2.3.1 se encontram devidamente normalizados. A Normalização serve para que sejam utilizados valores em uma determinada faixa (geralmente 0 e 1) visando uma maior padronização dos valores obtidos.

2.3.2 Centralidade de Intermediação

Segundo Freeman (1977), em sua pesquisa, a centralidade de intermediação é uma métrica utilizada para quantificar o controle de um ser humano sobre a comunicação entre outros numa rede social. Foi observado que vértices da rede que estão num caminho mais curto escolhido aleatoriamente entre dois vértices também escolhidos aleatoriamente, irão possuir uma elevada intermediação. Portanto, a centralidade de intermediação mede a quantidade de vezes que um nó age como ponte ao longo do caminho mais curto entre dois outros nós. Exemplificando, na Figura 2.6 os vértices A, B, C e D são os que possuem maior grau de

centralidade de intermediação.

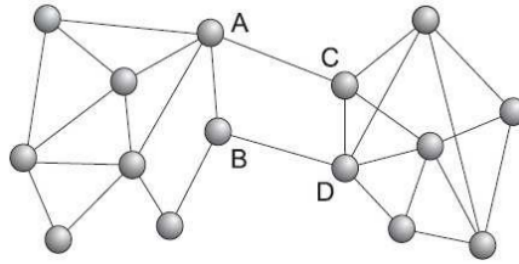


Figura 2.6: Nós com maior grau de intermediação

Para realizar o cálculo da intermediação de um vértice v num grafo $G := (V, E)$ com V vértices, temos que calcular os caminhos mais curtos entre cada par de vértices (s, t) e determinar a fração de caminhos mais curtos que passam através do vértice em questão (neste caso, vértice v) e por último somar esta fração de todos os pares de vértices (s, t) (BRANDES, 2001). Estando representado o cálculo da centralidade de intermediação na Figura 2.7.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Figura 2.7: Equação da Centralidade de Intermediação
Fonte: (BRANDES, 2001)

Onde, σ_{st} é o número total de caminhos curtos desde o nó s ao nó t e $\sigma_{st}(v)$ é o número desses caminhos que passam por v (BRANDES, 2001).

2.3.3 Page Rank

O *Page Rank* é um algoritmo inicialmente proposto por Brin e Page (1998) para ordenar resultados de busca do Google que gera um peso numérico para cada nó, assim podendo estimar sua importância em relação ao grafo. O entendimento por trás do *Page Rank* é que uma página Web é importante se existem muitas páginas apontando para ela ou se existem páginas importantes apontando para ela. A equação que calcula o *Page Rank* (PR) de um nó i , $PR(i)$, é definida na figura 2.8.

$$PR(i) = (1-d) + d \sum_{v \in S(i)} \frac{PR(v)}{N_v}$$

Figura 2.8: Equação do *Page Rank*

Fonte: (BRIN; PAGE, 1998)

Onde $S(i)$ é o conjunto de páginas que apontam para i , N_v denomina o número de arestas que saem do nodo v , e o parâmetro d é um fator que pode ter valor entre 0 e 1.

2.3.4 Modularidade

Vamos supor que temos a estrutura de uma rede qualquer e desejamos verificar se existe alguma divisão natural de seus vértices na forma de grupos ou comunidades onde estes não se sobrepõem e podem ser de qualquer tamanho. Como saber se existe uma melhor divisão da rede em apenas duas comunidades? Talvez o mais obvio seria resolver isto observando as divisões dos vértices em dois grupos, de modo a minimizar o número de arestas entre os grupos. Esta abordagem é frequentemente utilizada na literatura de particionamento de grafos, no entanto, o problema da estrutura de comunidade difere da de particionamento de grafos, no que se refere ao conhecimento antecipado do tamanho das comunidades. O problema é que a simples contagem de arestas não é uma boa maneira de quantificar o conceito de estrutura de comunidade. A boa divisão de uma rede em comunidades não é apenas aquela em que há poucas arestas entre as comunidades; é um em que há menos do que arestas esperadas entre as comunidades (NEWMAN, 2006).

Muitos problemas cientificamente importantes podem ser representados e empiricamente estudados usando redes. Por exemplo, padrões biológicos e sociais, a *World Wide Web*, redes metabólicas, redes neurais, etc., são problemas do mundo real que podem ser matematicamente representados e topologicamente estudados para revelar características estruturais inesperadas (NEWMAN, 2006). Muitas destas redes possuem uma certa comunidade que tem importância substancial na construção de um entendimento sobre a dinâmica da rede. Por exemplo, uma comunidade social intimamente ligada implicará em uma taxa mais rápida de transmissão de informação ou rumor entre elas do que uma comunidade vagamente ligada. Entretanto, se a rede é representada por um número de nós individuais conectados

por *links* que significam um certo grau de interação entre os nós, comunidades são definidas como grupos de nós interconectados densamente que somente são esparsamente conectadas com o resto da rede. Por isso, pode ser imperativo identificar as comunidades em redes já que estas podem ter propriedades muito diferentes, tais como grau de nó, coeficiente de agrupamento, centralidade de intermediação, centralidade, entre outros (NEWMAN, 2007). A modularidade é uma dessas medidas, que quando maximizada, leva ao aparecimento das comunidades em uma determinada rede.

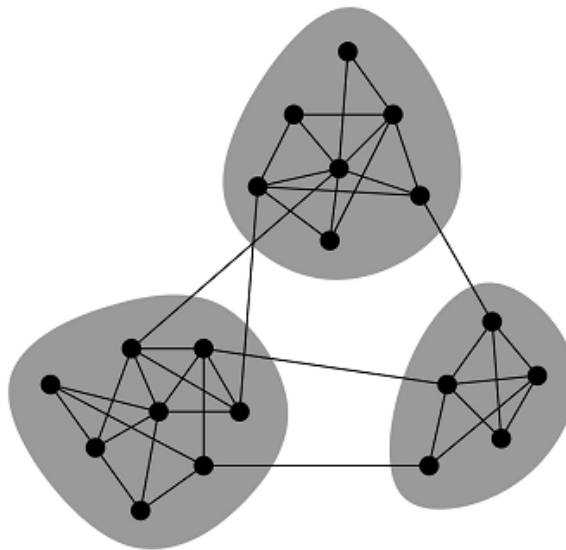


Figura 2.9: Grafo representando 3 comunidades determinadas pela modularidade
Fonte: (NEWMAN, 2006)

Segundo Newman (2006) a modularidade é uma medida de estrutura de redes ou grafos. Foi designada para medir a força de divisão da rede em módulos (grupos ou comunidades). Redes com alta modularidade têm conexões densas entre os nós dentro de comunidades, mas ligações esparsas entre nós de diferentes comunidades. A modularidade é frequentemente usada em métodos de otimização para a detecção de estrutura da comunidade em redes. A modularidade é uma constante multiplicativa, o número de arestas que caem dentro dos grupos menos o número esperado de uma rede com arestas equivalentes colocadas aleatoriamente. A modularidade pode ser positiva ou negativa, com valores positivos, indica a possível presença de estrutura da comunidade. Assim, podemos procurar a estrutura da comunidade precisamente por olhar para as divisões de uma rede que têm valores positivos, e preferencialmente amplos, de modularidade.

2.4 Propriedades da Rede

2.4.1 Componente Gigante

Na teoria de redes, um componente gigante é um componente conectado de um dado grafo aleatório que contém uma fração constante de todos os vértices do gráfico. Em Componentes gigantes cada possível aresta ligando pares de um dado conjunto de n vértices está presente, independentemente de outras arestas, com uma probabilidade p . Neste modelo, se $p \leq \frac{1-\epsilon}{n}$ para qualquer constante $\epsilon > 0$, então com alta probabilidade de todos os componentes ligados do grafo tem tamanho $O(\log n)$, e não há nenhum componente gigante. Entretanto, para $p \geq \frac{1+\epsilon}{n}$ existe com alta probabilidade um único componente gigante, com todos os outros componente tendo tamanho $O(\log n)$. Para $p = \frac{1}{n}$, intermediário entre essas duas possibilidades, o número de vértices no maior componente do grafo é com alta probabilidade proporcional ao $n^{2/3}$ (BOLLOBÁS, 2001).

2.4.2 Clique

Uma clique em um grafo não-direcionado é um subconjunto dos seus vértices tais que a cada dois vértices no subconjunto estão ligados por uma aresta. Cliques são um dos conceitos básicos da teoria dos grafos e são utilizados diversos outros problemas matemáticos e construções de grafos. Cliques também tem sido estudado na ciência da computação, por exemplo, a tarefa de encontrar a existência de um clique de um determinado tamanho em um grafo é um problema NP-Completo, mas apesar deste resultado, muitos algoritmos para encontrar cliques foram estudados (ALBA, 1973).

O termo “clique” vem de Luce e Perry (1949), que usaram subgrafos completos em redes sociais para modelar cliques de pessoas; Isto é, grupos de pessoas que conheciam umas às outras. Uma clique em grafo não direcionado $G = (V, E)$ é um subconjunto do conjunto de vértices $C \subseteq V$, tal que, para cada dois vértices em C , exista uma ponta de ligação entre os dois.

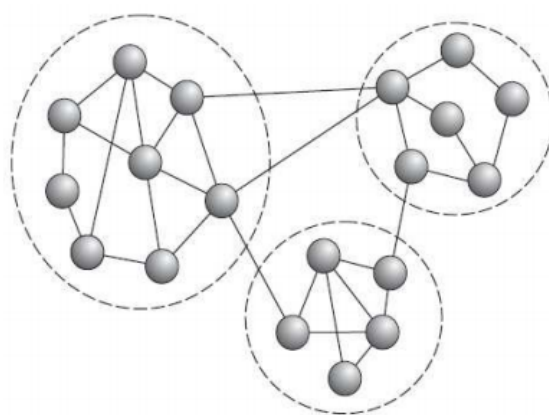


Figura 2.10: Rede com três cliques

A Figura 2.10 mostra um exemplo de uma rede com três cliques. Os cliques podem representar uma instituição, um subgrupo específico e mesmo identificar a movimentação em torno de um determinado problema (MARTELETO, 2001). É nos cliques que existe uma densidade maior de comunicação, ou seja, é mais eficiente compartilhar informações dentro de um grupo (FREITAS; PEREIRA, 2005).

Capítulo 3

Trabalhos relacionados

O objetivo deste capítulo é apresentar os trabalhos relacionados direta ou indiretamente ao tema de pesquisa. Temos o interesse principal de analisar de que modo redes de citações estão sendo utilizadas com objetivo de fornecer meios para identificação de publicações relevantes. É importante mencionar que os trabalhos relacionados citados foram objeto de uma revisão da literatura. Na medida do nosso conhecimento, os trabalhos descritos seguir são os principais trabalhos relacionados ao tema de pesquisa do presente trabalho.

3.1 *Coauthorship Networks and Patterns of Scientific Collaboration*

O artigo de Newman (2004), apresentou um estudo das redes de coautoria de diferentes áreas do conhecimento, dentre elas Biomedicina, Física e Matemática, entre os anos de 1995 a 1999. O autor destacou a importância de se estudar a rede como um sistema dinâmico. Os resultados revelaram ambas similaridades e diferenças entre as diferentes áreas. Cientistas de Biomedicina tendem ter significativamente mais coautores que matemáticos e físicos, um resultado que reflete o trabalho intensivo, direção predominantemente experimental da biologia atual. Outras diferenças são menos facilmente explicadas, como em biologia, por exemplo, é muito menos provável do que em matemática que dois dos coautores de um também será coautores de um outro, um resultado que ainda não recebeu uma explicação clara.

Esse tipo rede com diversas áreas de conhecimento permite analisar aspectos de relação entre as áreas, podendo assim comparar quantitativamente os dados de produção científica de cada área, por exemplo a distribuição do número de coautores que os cientistas tem para cada área.

3.2 *SUCUPIRA: A system for Information extraction of the Lattes Platform to identify academic social networks*

Já Alves, Yanasse e Soma (2011) apresentam o SUCUPIRA, sistema de extração de informações da Plataforma Lattes onde principal o objetivo foi identificar relacionamentos acadêmicos entre os pesquisadores. Atualmente, apenas um tipo de relacionamento é identificado. Entretanto, com as informações que já são extraídas do CV Lattes dos pesquisadores foi possível identificar outros tipos de relacionamentos como, por exemplo, relações de orientado/orientador, participação em bancas etc. Além disso, pode-se adicionar informação a própria representação da rede social, por exemplo, incluindo peso nas arestas representando o número de vezes que um determinado relacionamento acontece. Com isso, foi possível visualmente observar os quão intensos tais relacionamentos ocorrem.

Apesar de apresentar uma excelente ferramenta utilizando como fonte de dados de colaboração científica o *lattes*, ainda deixa a desejar no aspecto identificar através de métricas de ARS autores ou publicações relevantes. Além disso a ferramenta não possibilita a geração da rede a partir de uma área de conhecimento específica e não foi desenvolvido uma forma de se identificar comunidades dentro da rede. Apesar disso permite a comparação entre os autores escolhidos.

3.3 Minerando publicações científicas para análise da colaboração em comunidades de pesquisa

O Revoredo et al. (2012), apresentou uma abordagem para análise de uma comunidade científica a partir de suas publicações, onde sua finalidade foi traçar o perfil desta comunidade quando gera-se os indicadores obtidos a partir da mineração das publicações científicas. A

proposta foi avaliada considerando a comunidade nacional de Sistemas de Informação, através das publicações nas edições de 2008 a 2011 do Simpósio Brasileiro de Sistemas de Informação (SBSI). Esses artigos devem seguir o modelo de artigos da Sociedade Brasileira de Computação e ter formato Adobe *Portable Document Format* (PDF). Foi utilizada uma base de treinamento considerando três edições dos anais do SBSI (2008-2010) e para validação do classificador aprendido os anais da edição de 2011. Dessa forma, foram considerados 57 artigos para treinamento, e 27 para validação.

Neste estudo a forma como os dados são minerados utilizando como entrada de dados unicamente arquivos PDF deixa o estudo e a própria ferramenta utilizada focada em um pequeno grupo de dados e vai de contramão a busca de dados de publicações em bibliotecas digitais que possuem um grande acervo uma maior possibilidade de encontrar dados mais precisos.

3.4 Um Estudo sobre os Impactos dos Relacionamentos Sociais na Avaliação do Mérito Científico

O Guedes e Duarte (2013) aplicaram técnicas de análise de redes sociais a redes formadas pelas colaborações científicas entre pesquisadores. Tais relações podem assumir diversas formas, como orientação em programas de pós-graduação, coautorias em artigos científicos, participações em bancas ou comissões de avaliação, etc. Com isso identificaram os relacionamentos entre pesquisadores na área de Ciência da Computação, visando verificar possíveis efeitos desses relacionamentos na concessão das chamadas bolsas de produtividade em pesquisa do CNPq e na avaliação dos programas de pós-graduação na área de computação pela CAPES.

3.5 Análise da rede de coautoria do simpósio brasileiro de bancos de dados

O artigo de Procópio, Laender e Moro (2011) apresentou uma análise da rede de coautoria do Simpósio Brasileiro de Banco de Dados (SBBD) em 2010 quando o mesmo completou

25 anos Para isso foram coletados dados bibliográficos das edições realizadas até aquele momento para extração de uma série de estatísticas, tais como a média de artigos por autor, média de artigos por edição, média de coautores por artigo, etc. Além disso, foi realizada uma análise da rede de coautoria do SBBD, examinando tanto suas características estruturais quanto sua evolução temporal.

3.6 Characterizing and Predicting Community Members from Evolutionary and Heterogeneous Networks

O artigo de Zhao et al. (2008) estudou o relacionamento entre autores e comunidades usando técnicas de mineração. Eles propuseram caracterizar e prever membros de uma comunidade a partir da evolução de dados heterogêneos da web. Primeiramente criaram um *framework* para analisar a evolução de redes heterogêneas. Então a rede acadêmica, que foi extraída de um milhão de artigos científicos, foi usada como um exemplo para ilustrar o *framework*. Finalmente, duas aplicações exemplo da rede acadêmica foi apresentada. Resultados experimentais com uma grande rede heterogênea mostrou que o *framework* proposto pode produzir bons resultados em termos de recomendação de membros de uma comunidade. Também conhecimento e ideias podem ser ganhos analisando o padrão de evolução de uma comunidade

3.7 Considerações

A partir da revisão da literatura anterior, todo trabalho que se propõe a desenvolver algum tipo de pesquisa relacionada a relevância de publicações ou autores deverá ter ciência das seguintes premissas:

1. As bibliotecas digitais somente tem metadados de citações indexadas por elas, e nem sempre as citações estão formatadas de maneira correta.
2. Em relação a criação de redes coautoria, estas bibliotecas digitais em sua maioria não mantêm uma padronização de nomes, fazendo que se torne difícil um processo automatizado para geração de tal.

3. Boa parte das ferramentas disponíveis para obtenção e análise de dados científicos ainda não realizam cálculo de métricas ou estatísticas sobre determinada área de conhecimento.
4. Em sua maioria os estudos sobre produção científica focam na análise da coautoria das publicações.

De acordo com os pontos elencados anteriormente, identificamos alguns pontos críticos em estudos desta natureza. Desta forma, a partir da análise dos principais estudos relacionados diretamente ao nosso tema e, visando avançar o estado da arte nesse tipo de pesquisa, nosso trabalho irá apresentar os seguintes diferenciais.

1. Enquanto a maioria dos trabalhos se focam em fazer uma ferramenta que obtenha os dados que se deseja, nosso trabalho se propõe a gerar como artefato uma ferramenta que possibilite a análise de qualquer área de conhecimento.
2. Nossa metodologia permitirá que a partir de um termo de busca terá como saída dados quantitativos e qualitativos referentes aos resultados obtidos.
3. Trabalharemos com conceitos de teoria dos grafos e da teoria das redes complexas para inferirmos sobre a relevância de autores, publicações e meios de publicação.
4. Nosso estudo permitirá a classificação de sub-áreas da rede a partir de palavras-chaves das publicações pertencentes a elas.

Capítulo 4

Procedimentos Metodológicos

O presente trabalho tem interesse em obter respostas para a seguinte questão geral de pesquisa: “Pode o uso de técnicas de análise de redes sociais auxiliarem a mineração de conteúdo relevante de publicações científicas?”. Com objetivo de endereçar a citada questão de pesquisa foi proposta uma metodologia para análise de conteúdo relevante associado a publicações científicas. Tal metodologia é baseada em redes sociais e está centrada nas seguintes atividades: (i) formalização do processo para criação da rede de citações, (ii) provimento de suporte automatizado para criação da rede de citações e (iii) definição de um *framework* de avaliação da metodologia para análise de conteúdo relevante. Nas seções a seguir descreveremos em detalhes cada uma dessas atividades.

4.1 Processo para criação e análise da rede de citações

O objetivo principal dessa atividade é a formalização de um processo para identificação e mapeamento dos artigos e comunidades relevantes de uma área de conhecimento. Deste modo, a realização dessa atividade envolve uma série de passos detalhados na Figura 4.1.

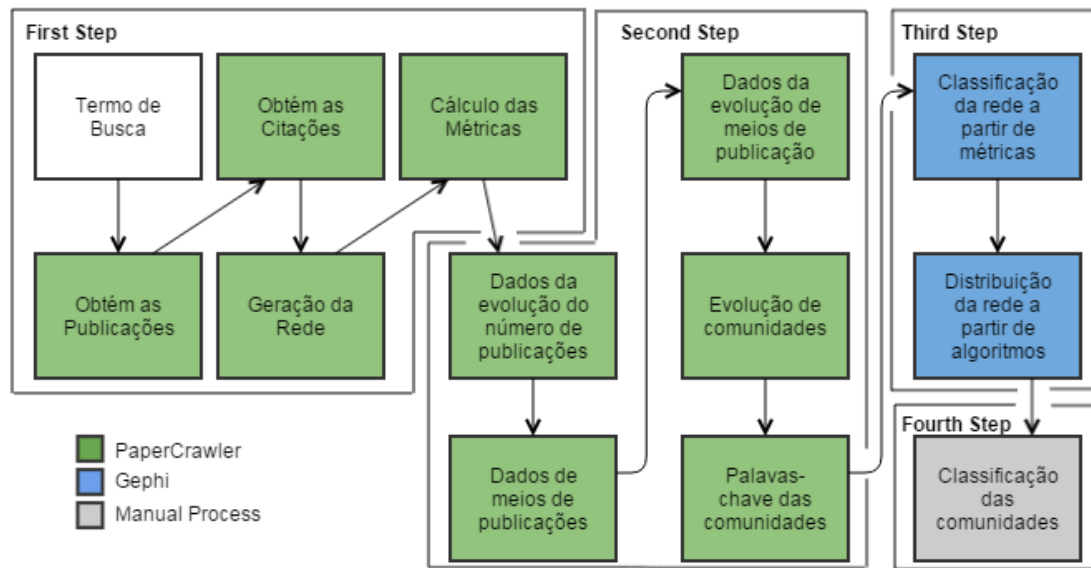


Figura 4.1: Processo Metodológico

Passo 1: Modelagem da rede - O primeiro passo da metodologia é obter as publicações e suas citações a partir de uma *query* (esta podendo representar uma área de conhecimento, como por exemplo, “Análise de Redes Sociais” ou “Sistemas Embarcados”), com isso pode-se gerar a rede de publicações e realizar cálculo das métricas de ARS, como grau, grau de entrada, grau de saída, centralidade de intermediação, modularidade e *Page Rank*, utilizadas para diversos fins neste estudo entre eles a identificação da relevância e a classificação das comunidades.

Passo 2: Análise da rede - O segundo passo é analisar os dados obtidos referentes a área de conhecimento como os meios de publicações, a evolução das publicações ao longo dos anos, a evolução do número de publicações para os meios de publicações ao longo dos anos, a evolução das comunidades ao longo dos anos e por último obter as *keywords* de cada comunidade.

Passo 3: Classificação e distribuição da rede - O terceiro passo é obter uma visualização amigável da rede para então analisar a área de conhecimento como um todo.

Passo 4: Definição e classificação de comunidades - E finalmente o quarto passo é a partir das *keywords* identificadas em cada comunidade analisar de forma manual qual seria a

melhor classificação para as 5 maiores e mais relevantes comunidades.

4.2 Criação do suporte automatizado

O objetivo dessa atividade é definição da arquitetura e implementação do suporte automatizado (i.e. ferramenta *Paper Crawler*) para criação e análise da rede de citações. Para implementação da ferramenta foi necessário um “*Tool Chain*” ou “Cadeia de Ferramentas” (em tradução livre), descritas na seção 4.3. A partir deste conceito as ferramentas podem ser utilizadas em cadeia, de modo que a saída de cada uma das ferramentas torna-se a entrada para a outra, embora o termo seja utilizado amplamente para se referir a qualquer conjunto de ferramentas interligadas visando o desenvolvimento de uma ferramenta de propósito único. Desta maneira, a Tabela 4.1 descreve as principais tecnologias utilizadas para implementação da ferramenta.

Tabela 4.1: Tecnologias utilizadas na construção da arquitetura

Descrição	Tecnologia	Ferramenta
Linguagem de Programação	Java	jdk1.7.0_22
Camada de Persistência	GEXF	<i>Gephi toolkit</i>
Camada de Integração	<i>API HTTP</i> Apache	httpclient 4.3
Camada de Visualização	Terminal	-

Percebemos que, de acordo com a divisão proposta na tabela acima, a ferramenta apresenta uma arquitetura dividida em três camadas (*layers*): A camada de persistência é a parte da aplicação responsável por realizar o salvamento dos dados necessários em arquivos, utilizando os recursos da *API gephi toolkit*, para assim serem futuramente analisados. Na camada de integração definimos as conexões externas a ferramenta, onde basicamente foram utilizadas as funcionalidades da *API Http* do apache no intuito de se realizar conexões *HTTP* as bibliotecas acadêmicas, obtendo assim os *htmls* a serem analisados. Na camada de visualização figuram as mensagens exibidas ao usuário sobre todo o andamento do processo.

Funcionamento do *Paper Crawler*

Basicamente podemos descrever o funcionamento da aplicação com base no fluxo de dados descrito na Figura 4.2. Inicialmente o usuário escolhe um termo a ser pesquisado, a

partir deste termo é realizado o processo (1) onde em sua concepção a ferramenta *Paper Crawler*¹ realiza requisições *HTTP* para cada biblioteca digital acadêmica escolhida pelo usuário, como *IEEEExplore*², *ACM*³, *Springer Link*⁴, entre outros, desde que os módulos para cada biblioteca estejam inseridos na ferramenta. Como esta foi feita de maneira modularizada permite a adição de novas bibliotecas com menor esforço, atualmente para esta pesquisa só foi possível fazer o módulo da biblioteca *IEEEExplore*. Este processo é feito de forma paralela em um *pool* de *threads* sendo possível até 50 requisições simultâneas. O resultado destas requisições são páginas *HTML* com as listagens dos resultados, nelas a ferramenta realiza leituras para extrair os metadados dos artigos retornados, como, por exemplo, o título, autores, local de publicação, ano da publicação, *link* de *download*, *link* para os detalhes, entre outros.

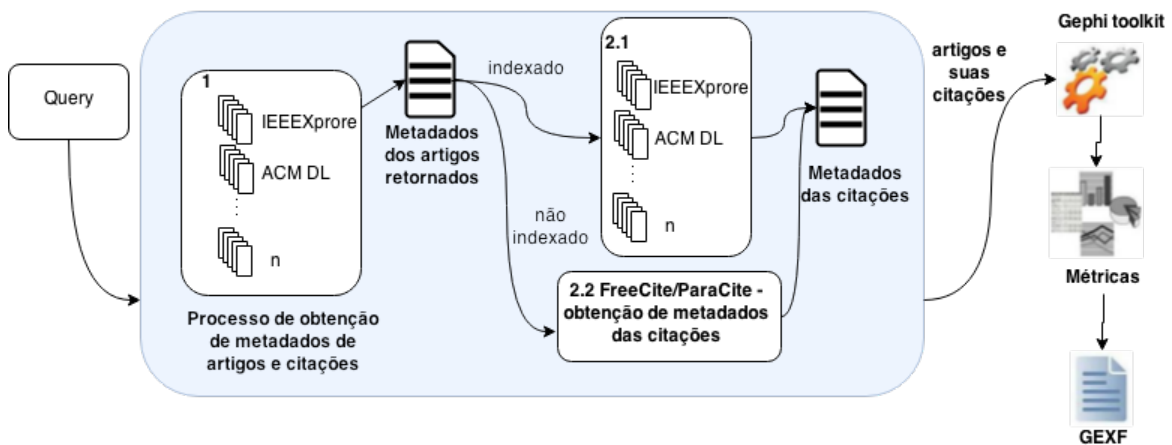


Figura 4.2: Paper Crawler

Após a obtenção dos artigos o usuário deve escolher que método de identificação de citações ele deseja utilizar. No contexto do presente trabalho temos dois métodos de identificação: **método indexado** e **método não indexado**. Com isso, a ferramenta acessa o *HTML* de detalhes de cada artigo e a partir dela identifica alguns metadados restantes como as *keywords* e autores da publicação, só então a ferramenta executa a identificação das citações conforme a escolha do usuário. O **método indexado** considera somente as citações que são indexadas pela biblioteca digital acadêmica e que fazem parte da lista de artigos obtidos

¹<https://code.google.com/p/appsnaauthorrank/>

²<http://ieeexplore.ieee.org>

³<http://dl.acm.org>

⁴<http://link.springer.com>

inicialmente, este processo também é feito de forma paralela em um *pool* de **threads** sendo possível também até 50 requisições simultâneas. Em contrapartida, o **método não indexado** considera todas as referências utilizando analisadores de citações online, neste caso o *Freecite*⁵ e o *Paracite*⁶, pois a maioria das bibliotecas digitais acadêmicas não indexam estas referências. A entrada para estes analisadores é o texto da citação obtidos nos metadados do artigo. Através de requisições *HTTP* aos analisadores obtém-se as páginas *XML* como resultado e a ferramenta *Paper Crawler* faz uma leitura destes *XMLs* e obtém como resultado os metadados das citações.

Com todos os artigos e referências identificados, inicia-se a geração da rede de citações e cálculo das métricas utilizando a biblioteca *gephi toolkit* então como resultado temos um arquivo de extensão *GEXF* com a rede nele. Com o arquivo da rede gerado a *Paper Crawler* faz um conjunto de análises quantitativas com o intuito de identificar a evolução dos artigos ao longo dos anos, as *keywords* mais relevantes para cada comunidade, a evolução de publicações ao longo dos anos para cada comunidade, a identificação dos meios de publicação mais relevantes e a evolução de publicações ao longo dos anos para os meios de publicação, a partir disso um arquivo de LOG é salvo para futura análise do usuário

4.3 Ferramentas e tecnologias

Nas seções a seguir iremos expor as principais tecnologias utilizadas para a construção da ferramenta *Paper Crawler*. Adicionalmente, iremos expor as principais razões da escolha de tais tecnologias. Para uma consulta mais detalhada a respeito do conteúdo a seguir, consulte as referências adicionais do trabalho.

4.3.1 Java

Em 1991, a *Sun Microsystems* iniciou um projeto chamado *Green*. O resultado foi uma linguagem de programação baseada em C/C++, denominada de Java. A *Sun* expôs esta linguagem oficialmente em 1995, chamando a atenção do mundo dos negócios por causa da capacidade de desenvolver sistemas em plataforma *web*. Atualmente, Java é uma das lingua-

⁵<http://http://freecite.library.brown.edu/welcome>

⁶<http://paracite.eprints.org/>

gens mais utilizadas para o desenvolvimento de aplicativos de grande porte, para o aprimoramento de servidores *Web*, para o desenvolvimento de aplicações móveis e etc (DEITEL, 2005).

4.3.2 Gephi

A *Gephi* é uma ferramenta para pessoas que desejam explorar e entender grafos. Como o *photoshop* mas para dados, o usuário interage com a representação, manipula as estruturas, formas e cores para revelar propriedades escondidas. O objetivo é dar suporte a analistas de dados a fazerem hipóteses, intuitivamente descobrir padrões, isolar singularidades na estrutura ou falhas. É uma ferramenta complementar para estatísticas tradicionais, como o pensamento visual com interfaces interativas é agora reconhecido para facilitar o raciocínio (BASTIAN; HEYMANN; JACOMY, 2009).

4.3.3 Gephi toolkit

O *Gephi Toolkit* é um projeto modularizado (*Graph, Layout, Filters, IO...*) desenvolvido em Java. O *toolkit* é só um único JAR que qualquer um pode reutilizar em novas aplicações Java e obter tarefas que podem ser feitas no *Gephi* automaticamente, a partir de uma linha de comando por exemplo. A habilidade de usar as características do *Gephi* assim em outras aplicações Java aumenta as possibilidades e promete ser muito útil.

A *Gephi* é projetada de maneira modular dividido em diferentes módulos. Todas as características são envoltas em módulos separados, por exemplo um modulo de estrutura de grafos, um modulo para os algoritmos de layout e assim vai. Além disso, módulos de negócio são separados das interfaces de usuário. Isto permite manter somente módulos de negócio e rever a UI sem problemas. Este é o propósito do *toolkit*, que mantém somente os módulos pertencentes ao core do *Gephi*.

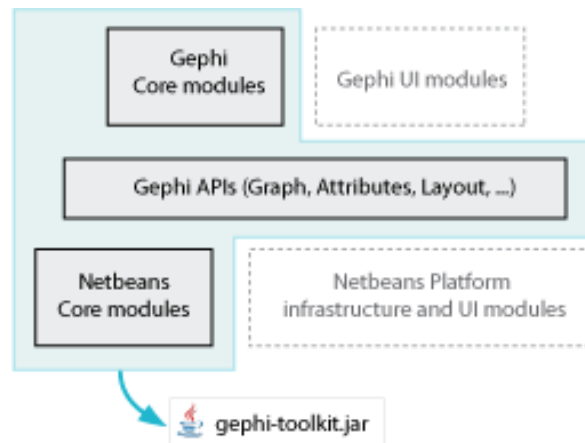


Figura 4.3: Gephi toolkit

A Figura 4.3 mostra a arquitetura do *Gephi* e que parte está incluída na *toolkit*.

Capítulo 5

Estudos de caso

No presente capítulo iremos expor os principais resultados obtidos no contexto desta pesquisa. Assim, este capítulo apresenta os resultados associados a três estudos de caso: Resultados do Estudo da área de análise de redes sociais (Seção 5.1), resultados da análise das edições do *Webmedia* (Seção 5.2) e resultados do Estudo da área de Sistemas Embarcados (Seção 5.3).

5.1 Análise de redes sociais

Nessa seção apresentaremos os resultados obtidos pela ferramenta *Paper Crawler* para a cadeia de busca “*Social Network Analysis*” and “*online*” no contexto do repositório de publicações digitais *IEEEExplore*, realizado no início do ano de 2014. Com objetivo de melhor organizar a apresentação dos resultados associados a esse estudo de caso, dividimos o conteúdo da seção conforme a seguir: Na seção 5.1.1 serão apresentadas diversas análises em cima dos dados obtidos pelo *Paper Crawler*, na seção 5.1.2 serão expostos os resultados obtidos referentes a análise de relevância das publicações. Em ambas as seções os resultados estão divididos em dois casos relacionados aos métodos de captura de referências descritos no Capítulo 3. Finalmente, na seção 5.1.3 descrevemos algumas conclusões alcançadas neste estudo

É importante mencionar que no momento da execução desse estudo de caso, a ferramenta *Paper Crawler* estava em fase de desenvolvimento e alguns dos dados explanados no Capítulo 4 ainda não tinham sido idealizados. Exemplificando, as funcionalidades associadas ao

(i) cálculo das métricas, (ii) classificação das comunidades e (iii) obtenção de dados quantitativos, descritos na fase 2 da metodologia (Capítulo 4), da área de conhecimento pesquisada ainda não estavam disponíveis na versão da ferramenta utilizada no estudo de caso. A versão da ferramenta utilizada no estudo de caso basicamente gerava apenas a rede a partir dos artigos e referências obtidos. Todo o processo de análise, cálculo das métricas e visualização da rede eram realizados com suporte de uma ferramenta de terceiros -i.e. *Gephi*.

5.1.1 Análise quantitativa da área

A seguir iremos expor os resultados associados a dois experimentos no contexto do estudo de caso de redes sociais. No primeiro experimento iremos apresentar os resultados associados ao uso do **método não indexado**. Por fim, no segundo experimento iremos apresentar os resultados associados ao uso do **método indexado**.

Primeiro Experimento - Uso do método de identificação não indexado

Iniciamos esta análise a partir do segundo método de captura de referências, nele capturamos todas as referências isso faz com que obtenhamos diversas publicações de diversas áreas de conhecimento relacionadas a área de **Análise de Redes Sociais**. A evolução ao longo dos anos do número de publicações pode ser ilustrada na Figura 5.1. Podemos observar que depois do ano de 2009 houve uma queda no número de publicações.

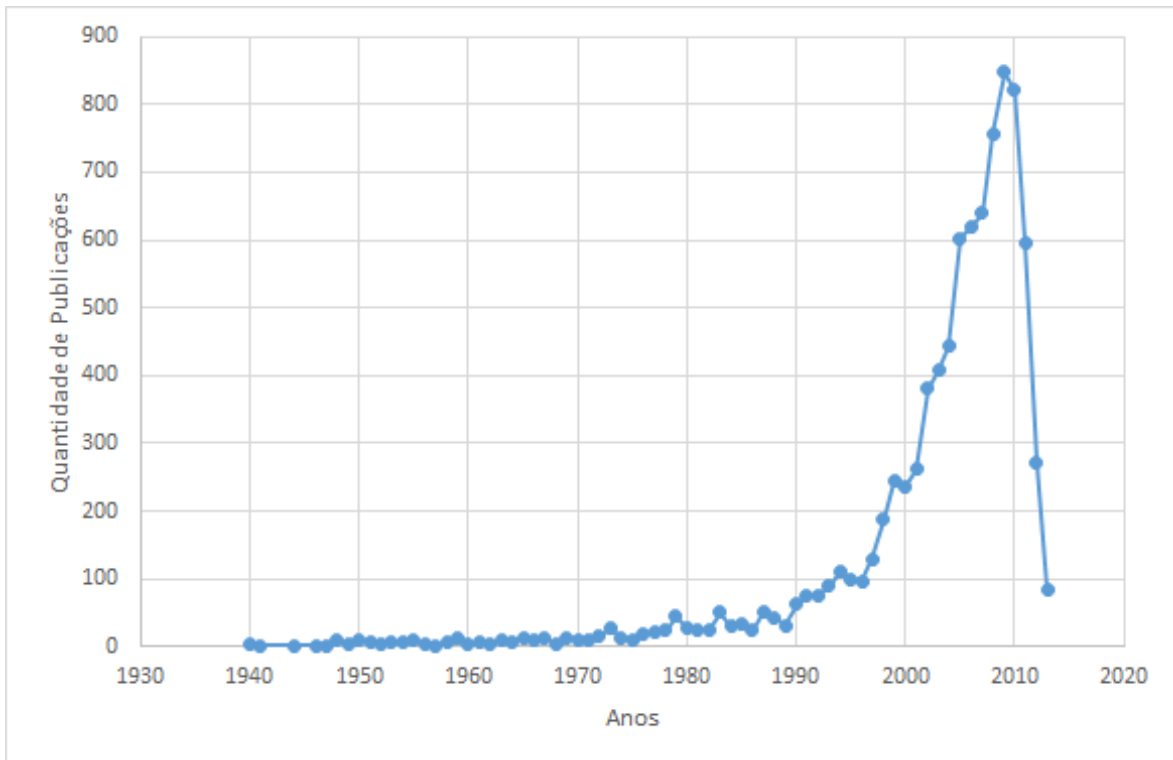


Figura 5.1: Número de Publicações ao longo dos anos

Dentre os dados obtidos identificamos que as publicações encontradas foram publicadas em cerca de 2559 locais de publicação e então utilizando como base a planilha de dados do *Gephi* calculamos manualmente o número de publicações para cada veículo. Vale salientar que não foi possível identificar o veículo de publicação em cerca de 40% nas referências, pois nenhum dos analisadores de citações utilizados conseguiu identificar isto no texto da citação. Podemos observar na Figura 5.2 os quinze principais meios de publicação, considerando como métrica o número de publicações de cada um deles.

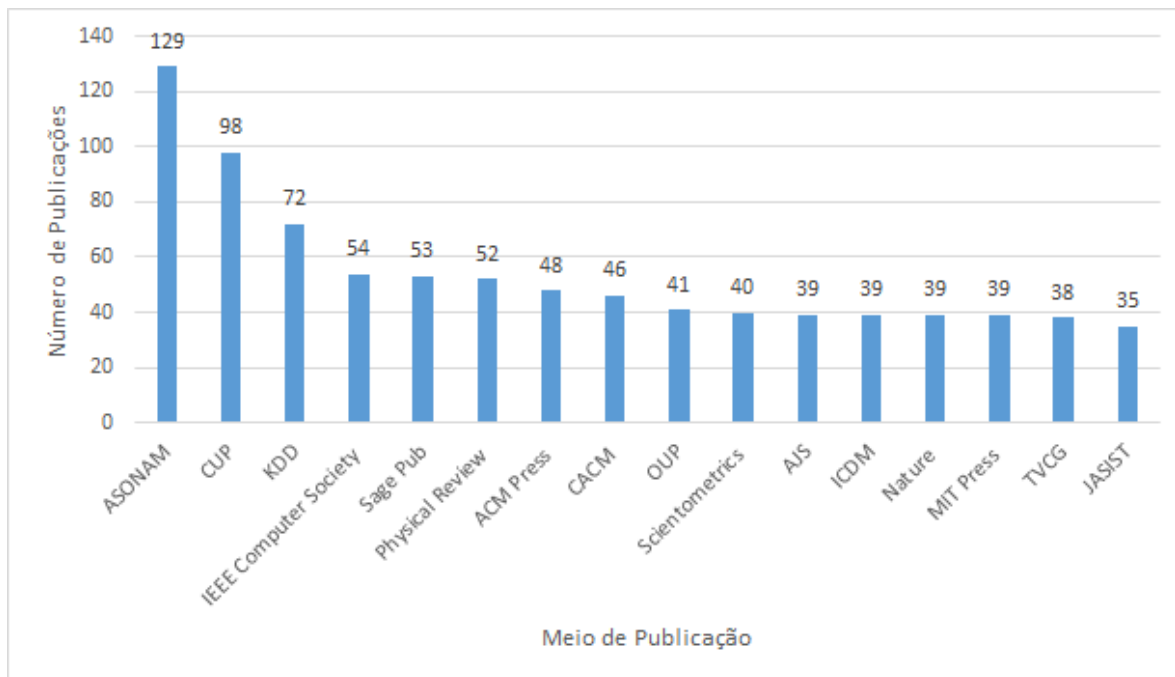


Figura 5.2: Meios de Publicação Mais Relevantes

Neste caso observamos que o evento que foi identificado com o maior número de publicações foi o *Asonam - International Conference on Advances in Social Networks Analysis and Mining*, com 129 associadas. Esse resultado pode ser considerado previsível, pois esse evento é considerado pela comunidade científica como um dos mais relevantes para área de Análise de Redes Sociais.

Segundo experimento - Uso do método de identificação indexado

No segundo experimento foi possível analisar o crescimento de publicações somente da área de **Análise de Redes Sociais** ao longo dos anos. Para isso foi utilizado o primeiro método de captura de citações (i.e. **método indexado**), que considera somente os artigos indexados pelo *IEEEExplore* retornados na busca inicial. Assim, foram obtidos os resultados associados a curva de crescimento ilustrada na Figura 5.3.

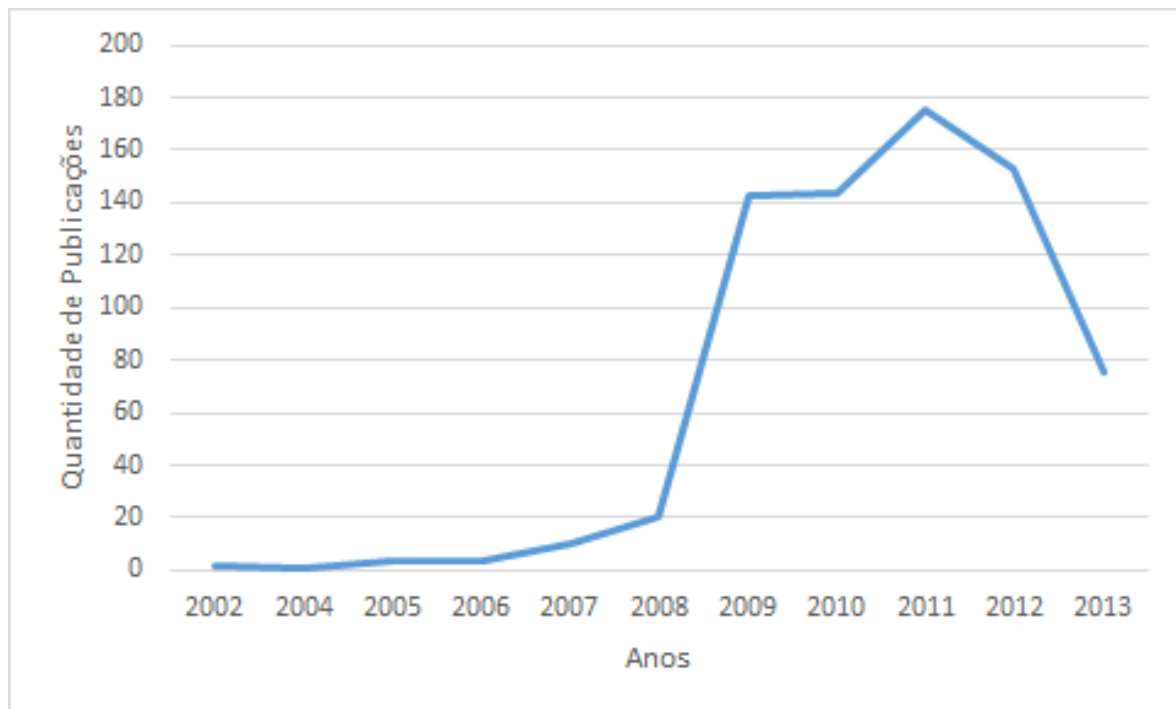


Figura 5.3: Número de Publicações ao longo dos anos para a área de ARS

Podemos observar que, considerando os dados da Figura 5.3, o crescimento é muito mais sinuoso nos anos 2000. Essa taxa de crescimento pode estar associada com a popularização das Redes Sociais *Online* neste período. O que acontece no caso ilustrado na Figura 5.1 é que as publicações mais antigas de temas sobre ARS ou sobre outras áreas que foram citadas modificaram a curva de crescimento de publicações para o início dos anos 90.

Vale salientar que os artigos que foram indexados pelo *IEEE Xplore* e apareceram no resultado da busca foram publicados em cerca de 267 locais de publicação. Nesse caso como todos os metadados estavam no *IEEE Xplore* não houve problemas em relação a encontrar os dados do meio de publicação. Na Figura 5.4 estão ilustrados os quinze principais meios de publicação da área de Análise de Redes Sociais.

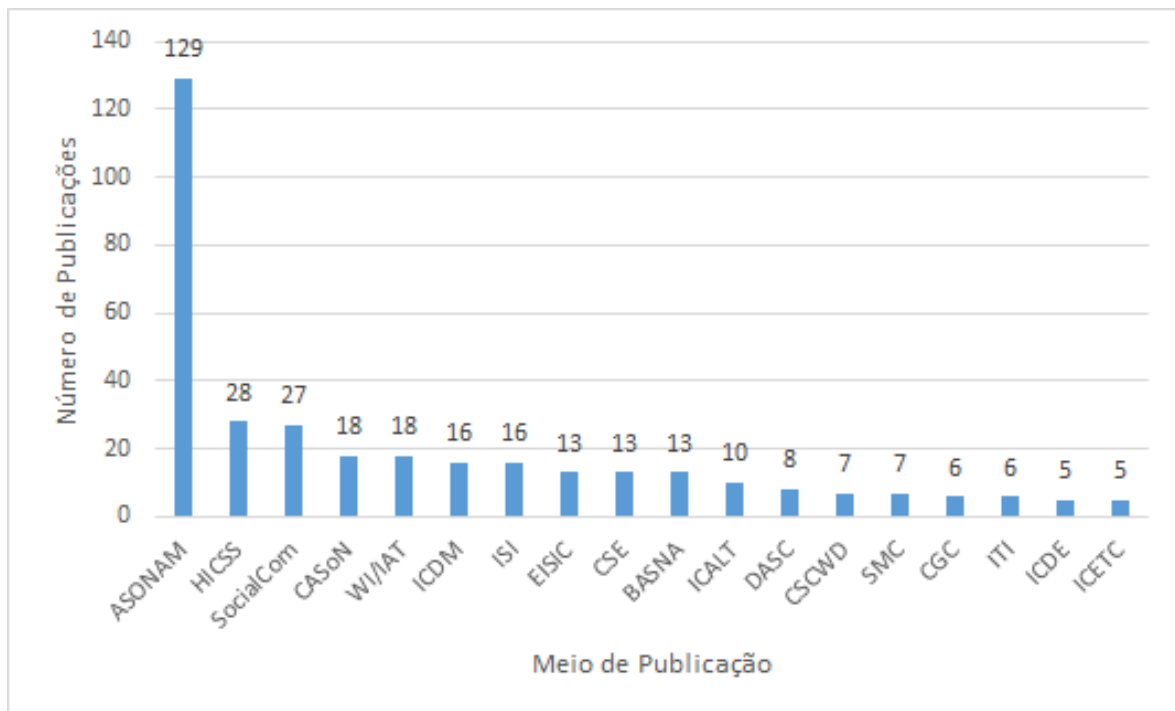


Figura 5.4: Meios de publicação mais relevantes

Novamente observamos que o evento que foi identificado com o maior número de publicações foi o *Asonam - International Conference on Advances in Social Networks Analysis and Mining*, com 129 associadas. Apesar da quantidade de publicações em cada veículo de publicação ser uma métrica quantitativa que pode indicar relevância, sua precisão as vezes pode ser contestada, pois eventos com alta visibilidade que recebem grande quantidade de submissões e possuem grande número de publicações tendem a ter uma diminuição na qualidade de suas produções e além disso suas produções pode ser pouco citadas (WEALE; BAILEY; LEAR, 2004).

Por fim, analisando os dois experimentos podemos concluir que o uso do *Paper Crawler* pode ser útil para identificar a evolução do número de publicações de uma determinada área ao longo do tempo. Esses dados podem ser utilizados para analisar a relevância de uma área de pesquisa. Adicionalmente, a *Paper Crawler* fornece indícios dos principais meios de publicação de determinada área de pesquisa. Esse dado é particularmente útil para pesquisadores que desejam identificar quais são os principais meios de publicação de determinada área de pesquisa.

5.1.2 Análise da Rede de Citações

A seguir iremos expor os resultados associados a dois experimentos no contexto do estudo de caso de redes sociais. No primeiro experimento iremos apresentar os resultados associados ao uso do **método não indexado**. Por fim, no segundo experimento iremos apresentar os resultados associados ao uso do **método indexado**.

Primeiro Experimento - Uso do método não indexado

A partir da cadeia de busca a ferramenta *Paper Crawler* consultou o *IEEE Xplore* e obteve 735 artigos como amostra inicial de artigos. Em seguida, extraíndo as referências de cada artigo da amostra inicial, foi gerada uma rede com cerca de 10800 nós representando os artigos. A rede incluía basicamente os artigos 735 artigos da amostra inicial e suas referências. Adicionalmente, a rede possuía cerca de 11700 arestas representando a associação entre os artigos. A partir desta rede utilizamos a ferramenta *Gephi* para análise e cálculo das métricas com o intuito de identificar as publicações mais relevantes. Para isso utilizamos duas métricas, o *PageRank*, a partir do qual podemos identificar nós que são mais citados por artigos também muito citados, e o Grau de Entrada, que neste caso representa o número de citações recebidas por um artigo. Podemos observar na Tabela 5.1 as 5 publicações com maior número de citações para ambas as métricas.

(a) Publicações com maior grau de entrada associado

Publicação	Grau
(WASSERMAN; FAUST, 1994)	69
(SCOTT, 2000)	23
(GIRVAN; NEWMAN, 2002)	18
(NEWMAN, 2003)	17
(NEWMAN; GIRVAN, 2004)	17

(b) Publicações com mais *Page Rank*

Publicação	<i>Page Rank</i>
(SCOTT, 2000)	1,84E+12
(GIRVAN; NEWMAN, 2002)	1,64E+12
(NEWMAN; GIRVAN, 2004)	1,58E+12
(WASSERMAN; FAUST, 1994)	4,03E+11
(NEWMAN, 2003)	1,58E+11

Tabela 5.1: Publicações mais relevantes

Como podemos observar nos resultados da Tabela 5.1, as mesmas publicações estão entre as cinco mais relevantes. Mesmo não havendo semelhança na ordem de relevância entre as duas métricas, isso provavelmente significa que há uma relação entre as duas métricas. O interessante é que se analisarmos a tabela do ponto de vista do número de citações a publicação Wasserman e Faust (1994) é considerada mais relevante, mas do ponto de vista do *Page Rank* essa mesma publicação foi considerada a quarta mais relevante. Significando que nos resultados apresentados na Tabela 5.1(b), o livro Wasserman e Faust (1994), em relação a Scott (2000), no nicho de dados obtido possui menos publicações que o citou e que também foram muito citadas. Logo, o *Page Rank* pode ser considerado uma métrica mais apropriada para análise de relevância segundo Cui, Wang e Zhai (2010) e Maslov e Redner (2008), já que esta além de analisar o próprio número de citações do nó, ele verifica também o número de citações de quem citou este nó.

Verificamos também como está distribuído o número de citações por publicações, observamos que houve uma grande quantidade de artigos com uma única citação (em torno de 9000), o restante foi bem distribuído entre valores abaixo de 697. Na Figura 5.5 podemos observar esta distribuição.

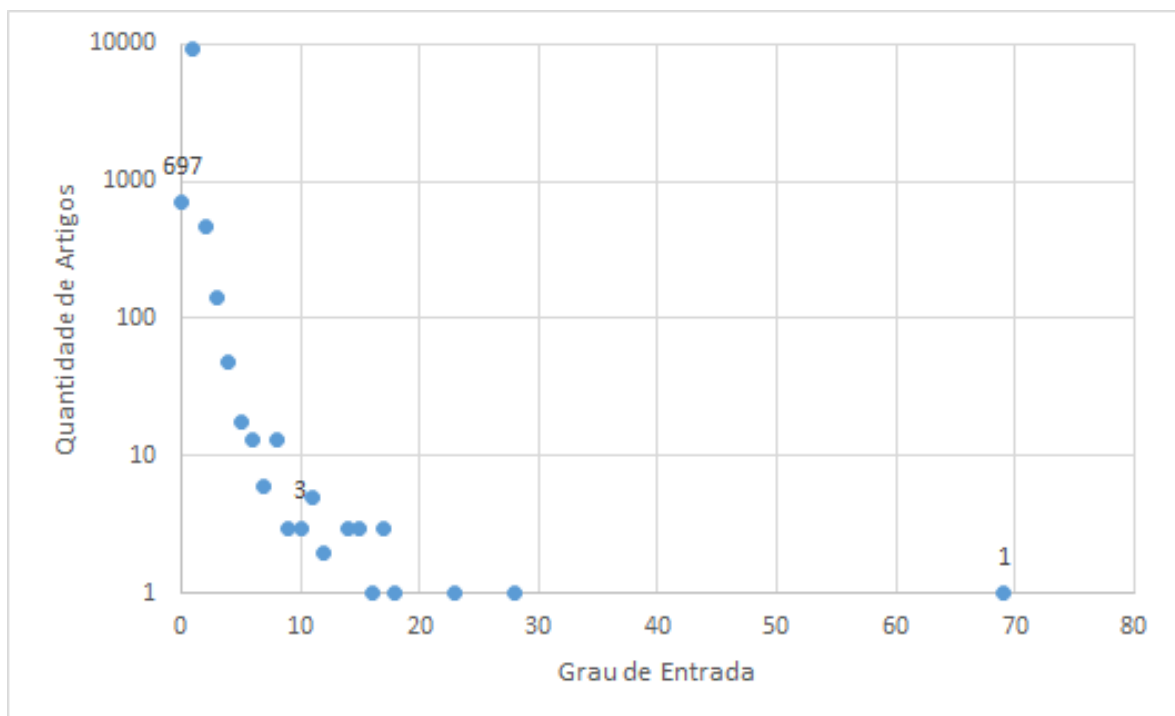


Figura 5.5: Distribuição de publicações por citação

Segundo experimento - Uso do método indexado

Analisando os artigos indexados pelo *IEEE Xplore* obtivemos uma rede formada por 735 nós e 31 arestas. Esta rede se mostrou pouco conectada, significando que dentre os resultados há poucas citações. Para tentar identificar quais deles são mais relevantes na rede, primeiramente consideramos a métrica grau de entrada, já que com ela poderíamos obter a quantidade de citações de cada publicação entre si. O maior número de citações foi 2, logo o resultado se tornou pouco satisfatório. A lista dos artigos considerados mais relevantes nesta rede segundo o grau de entrada pode ser observada na Tabela 5.2.

Tabela 5.2: Quantidade de Citações das publicações mais relevantes

Publicação	Número de Citações
(ERLIN; YUSOF; RAHMAN, 2009)	2
(NOLKER; ZHOU, 2005)	2
(FISHER, 2005)	2
(SEMENOV; VEIJALAINEN; BOUKHANOVSKY, 2011)	2
(WANG; LI, 2007)	1
(GLOOR et al., 2009)	1
(XUE; SHI; YANG, 2010)	1

Agora considerando a métrica *Page Rank*, podemos identificar os artigos que receberam mais citações de artigos também muito citados na rede. Na Tabela 5.3 estão descritos o *Page Rank* dos artigos mais relevantes na rede.

Tabela 5.3: *Page Rank* das publicações mais relevantes

Publicação	<i>PageRank</i>
(NGONMANG; TCHUENTE; VIENNET, 2012)	3,86E-03
(ERLIN; YUSOF; RAHMAN, 2009)	3,55E-03
(NOLKER; ZHOU, 2005)	3,55E-03
(FISHER, 2005)	3,55E-03
(SEMENOV; VEIJALAINEN; BOUKHANOVSKY, 2011)	2,99E-03
(WANG; LI, 2007)	2,43E-03
(XUE; SHI; YANG, 2010)	2,43E-03

É interessante observar que na Tabela 5.2 há dois artigos que também estão presentes na Tabela 5.3. Isso pode indicar mais um aspecto de relevância, já que eles foram os de maior valor nas duas métricas utilizadas. Também vale salientar que como a rede é pouco conectada e o *page rank*, como já havia frisado, identifica os artigos que receberam mais citações

de artigos também muito citados na rede, o resultado obtido por esta métrica pode ser considerado pouco satisfatório. Quanto mais conexões, ou seja, citações entre as publicações em questão melhor será a precisão do resultado obtido.

5.1.3 Considerações

Podemos afirmar que este estudo conseguiu apontar, que é possível identificar o nível de relevância de publicações de determinadas linhas de pesquisa através da utilização de métricas de análise de redes sociais.

Durante a fase de projeto e criação da ferramenta *Paper Crawler* foram encontrados alguns problemas, como a dificuldade de criar ou encontrar um analisador de citações com bom desempenho. Os analisadores escolhidos, o *Freecite* e o *Paracite*, apesar de serem de grande ajuda, ainda apresentam grandes problemas a serem resolvidos na resolução das citações. Ainda há muitas inconsistências nos dados retornados e, além disso, o tempo de resposta destes torna a ferramenta criada para este estudo muito lenta. Para a geração da rede citada nos resultados foi necessário um tempo médio de 2 horas e 30 minutos só para obter os dados em formato XML, retornado por estes analisadores. Além disso, muitas publicações possuem referências fora do padrão, o que dificulta ainda mais o trabalho dos analisadores.

Ao usar os analisadores a taxa de erro foi reduzida, pois foi realizada uma combinação dos recursos do *Freecite* e do *Paracite*, já que identificamos que no *Freecite* havia o problema de algumas informações virem erradas, como no caso do local de publicação e do ano para algumas referências. Com isso, ele era menos preciso, mas possuía uma grande quantidade de metadados que poderiam ser utilizados. O *Paracite* foi mais preciso, mas só trazia como informação o título, ano e local de publicação, então só foi utilizado para obter estas duas informações adicionais. Mesmo assim, ainda não possível identificar cerca de 40% dos meios de publicação.

5.2 WebMedia

O Simpósio Brasileiro de Sistemas Multimídia e Web, *WebMedia*, chegou em 2014 à sua vigésima edição consolidado como o principal evento nacional em sua área de atuação. Com o objetivo de caracterizar a comunidade científica atuante na área, analisamos o corpus de

453 artigos científicos publicados nas dezoito edições anteriores do evento realizadas entre os anos de 1995 e 2012.

Este estudo se desenvolveu focado na caracterização da comunidade científica atuante no evento, isto se deu através da criação de uma rede de coautoria, em que os nós representam pesquisadores, e que sempre que conectados, representam a divisão de autoria de um ou mais artigos publicados em alguma das edições do evento.

Foram utilizadas técnicas de Análise de Redes Sociais para permitir a avaliação não só de componentes isolados como também, e principalmente, dos relacionamentos entre estes componentes, permitindo-nos apontar pesquisadores mais relevantes/atuentes nas edições anteriores do evento de acordo com uma série de diferentes métricas de análise.

Este experimento foi um caso a parte em relação aos outros dois estudos de caso, primeiramente o foco deste estudo foi analisar as redes de coautoria, segundo os dados que estavam disponíveis em <http://paperz.zn.inf.br/>, então foi necessário o desenvolvimento de outro analisador para esse site. Além disso, lá não tinha informações sobre citações realizadas pelos artigos, então não poderíamos realizar uma análise de uma rede de citações, E por último este foi o único caso em que precisou-se gerar redes de coautoria para cada edição do evento. Explorado isso, exploramos os autores dos artigos publicados nas edições do *WebMedia* entre 1995 e 2012. A ferramenta *Paper Crawler* realizou extração dos dados dos artigos, sendo eles título, ano, autores e instituições de ensino. A partir destes dados a ferramenta gerou uma descrição das redes de coautoria para cada ano do evento, considerando também as instituições de atuação de cada autor.

Estas redes foram modeladas como grafos não-dirigidos onde os vértices representam os autores e as arestas indicam as relações de coautoria entre eles. O tamanho do nó é proporcional ao número de relações do autor e a cor dos nós identifica a que instituições o autor pertence. As redes foram geradas da seguinte forma:

- i O conjunto de autores de um artigo é representado através de uma clique, onde todos os autores, representados por nós, estão conectados entre si
- ii Existe uma aresta ligando nós de cliques distintas se os autores representados por estes dois nós publicaram algum artigo em conjunto.

5.2.1 Análise dos Anais do WebMedia

A partir das redes de coautoria geradas utilizamos uma ferramenta o *Gephi* para exibir e analisar um grafo com a evolução das redes de coautoria do *WebMedia*. Na figura 5.6 podemos observar a evolução das interações entre os autores e coautores de 3 em 3 anos desde 1995 até 2012.

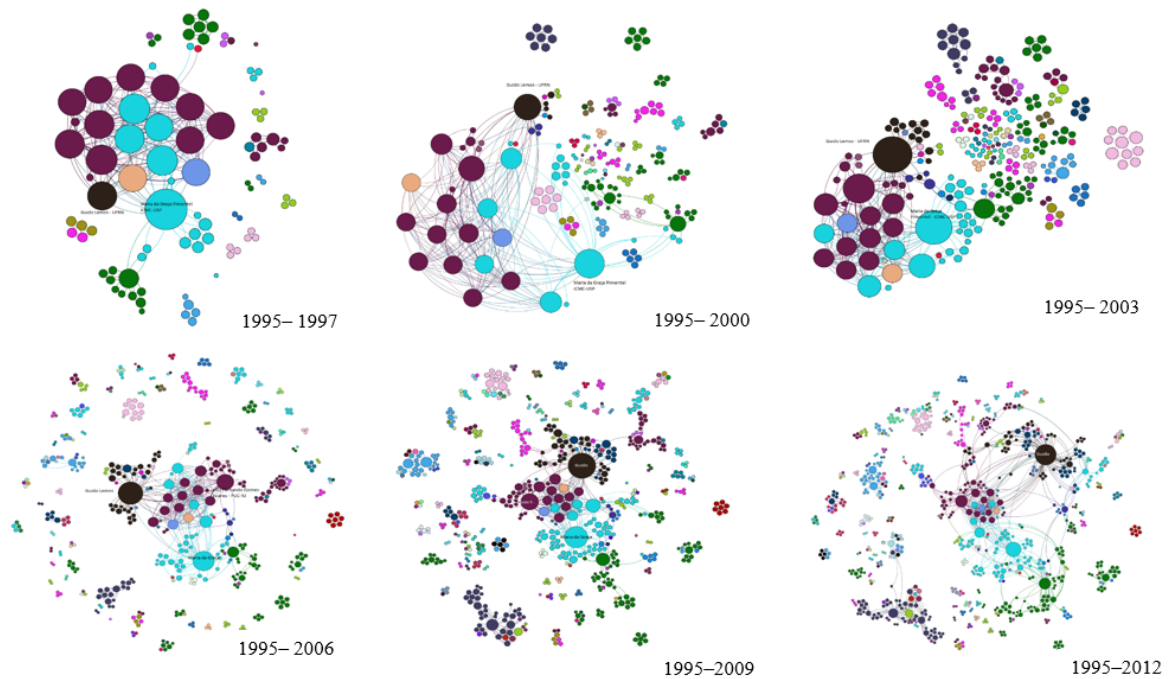


Figura 5.6: Evolução da rede de coautorias do *WebMedia*.

Inicialmente, entre 1995 e 1997, a rede possuía cerca de 96 (autores) nós e 273 arestas. Considerando todo o período avaliado, de 1995 a 2012, a rede evoluiu para 820 nós e 1869 arestas, um crescimento de 754%. Em todas as redes ilustradas na Figura 5.6 pode-se observar a formação de um **Componente Gigante**, ou seja, de uma sub-rede conectando uma grande parte dos nós da rede original, o que indica que a comunidade de autores do *WebMedia* é bastante coesa. O componente gigante da comunidade de autores do *WebMedia* é composto por 354 autores, ou 43% do total de autores presentes no evento. Os demais autores estão agrupados em outros 86 componentes conectados, ou sub-redes, que podem representar os diferentes grupos de pesquisa representados no evento.

O grafo na Figura 5.7 exibe o componente gigante com nós proporcionais ao grau de

cada autor, ou seja, ao número de conexões de coautoria que este autor estabeleceu ao longo dos anos através dos artigos publicados no *WebMedia*.

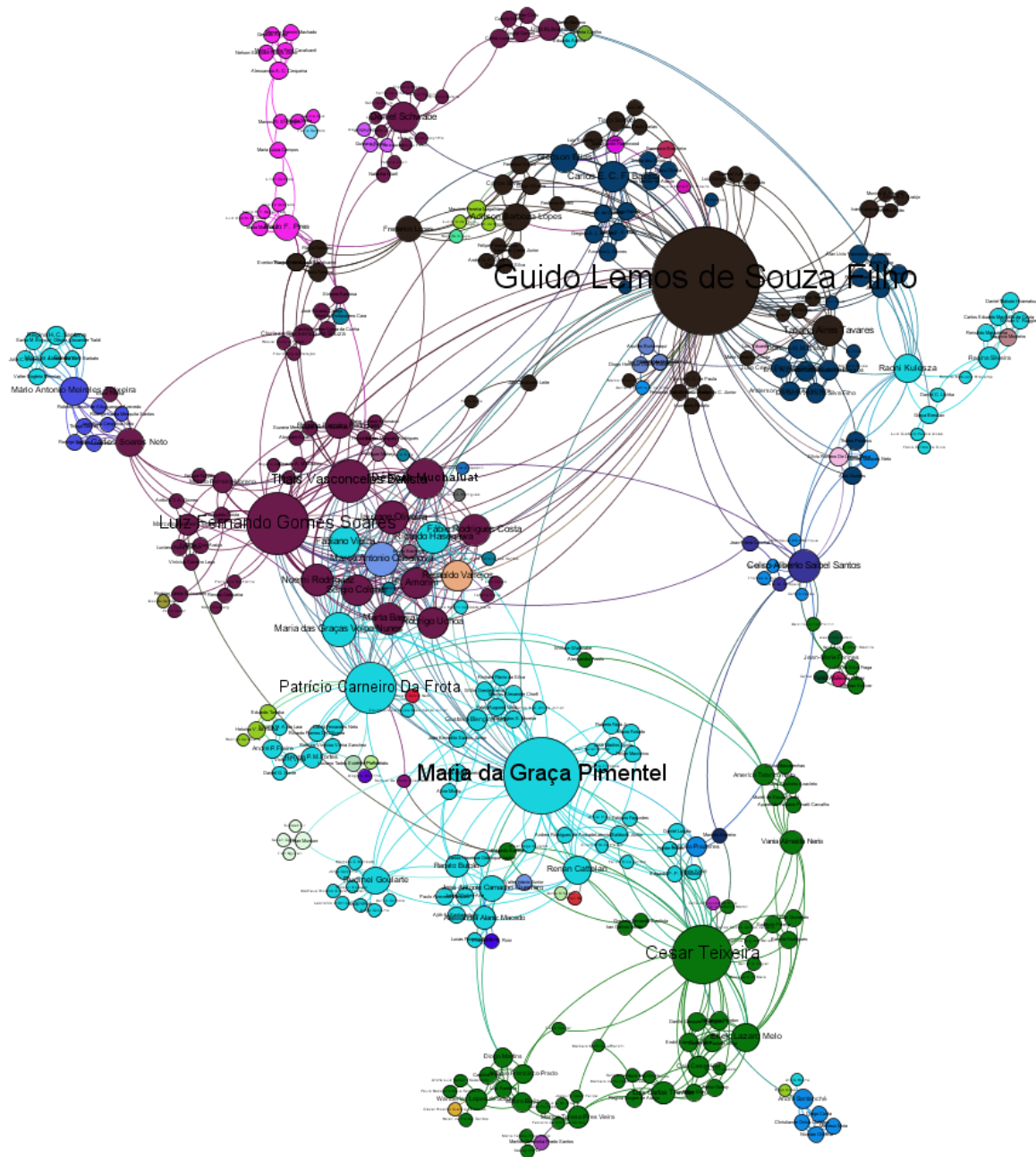


Figura 5.7: Componente Gigante da rede de 1995 a 2012.

Baseado em duas métricas de centralidade, a centralidade de grau e centralidade de intermediação, listamos na tabela 5.4 os autores considerados mais centrais no estabelecimento e manutenção destas redes de coautorias. O grau de um nó representa a quantidade de conexões, no nosso caso, de relações de coautoria, estabelecidas ao longo dos anos. Sendo assim,

os nomes na coluna da esquerda da tabela 5.4 indicam os autores com o maior número de relações de coautoria no artigos científicos publicados no *WebMedia* entre 1995 e 2012. Este número indica que tais autores tiveram papel fundamental na criação do componente gigante observado na figura 5.7 e dos demais componentes conectados presentes na rede de coautorias porém não deixa claro qual sua importância na manutenção dessa rede uma vez que autores podem estabelecer conexões com outros autores nós na rede.

A centralidade de intermediação (*betweenness*) apresenta uma medida da centralidade de um nó em uma rede. A centralidade de intermediação de um nó é proporcional a quantidade de vezes que este nó aparece nos caminhos mínimos entre quaisquer dois outros nós na rede. Ou seja, ela sintetiza o poder do nó de aproximar autores e uma rede de coautoria. Notem que a lista de autores mais centrais de acordo com esta métrica apresenta divergência em relação à lista de autores de maior grau. Isto acontece porque não é apenas o grau que determina a centralidade de um nó e sim as centralidades dos nós com os quais este nó está conectado. Um nó que se conecta com nós localizados na periferia da rede tende a ter uma centralidade de intermediação maior do que um nó que só se conecta a nós mais centrais uma vez que intermediará a conexão dos nós periféricos com o núcleo da rede.

Tabela 5.4: Estatísticas referentes a rede de coautoria

Grau	Centralidade de Intermediação
Guido Lemos	Guido Lemos
Maria da Graça Pimentel	Maria da Graça Pimentel
Luiz Fernando G. Soares	Luiz Fernando G. Soares
Cesar Teixeira	Cesar Teixeira
Patrício Carneiro da Frota	Celso Alberto S. Santos
Jussara Almeida	Patrício Carneiro da Frota
Thais Vasconcelos Batista	Raoni Kulesza
Marcos André Gonçalves	Thais Vasconcelos Batista
Débora Muchaluat	Carlos Eduardo C. F. Batista
José Valdeni de Lima	Paulo Pires

A Figura 5.8(a) apresenta a distribuição do número de coautores por artigo. Observa-se que 34% dos artigos têm dois coautores e cerca de 17% têm três coautores. Verificou-se a existência de um artigo com 17 coautores. Já em relação ao número de artigos por autor, podemos ver na figura 5.8(b) que a grande maioria dos autores (cerca de 77%) publicou apenas uma única vez no *WebMedia*, aproximadamente 11% publicaram duas vezes e 5% publicaram cinco ou mais vezes.

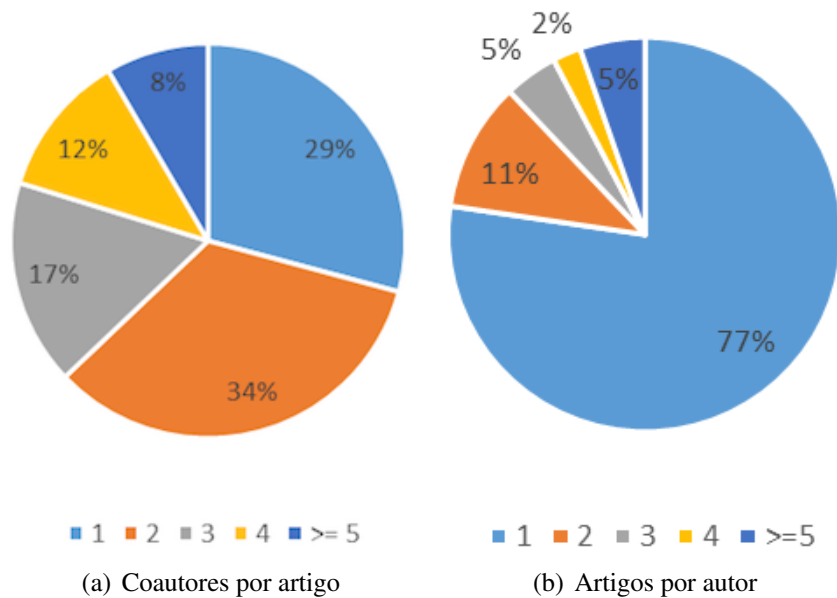
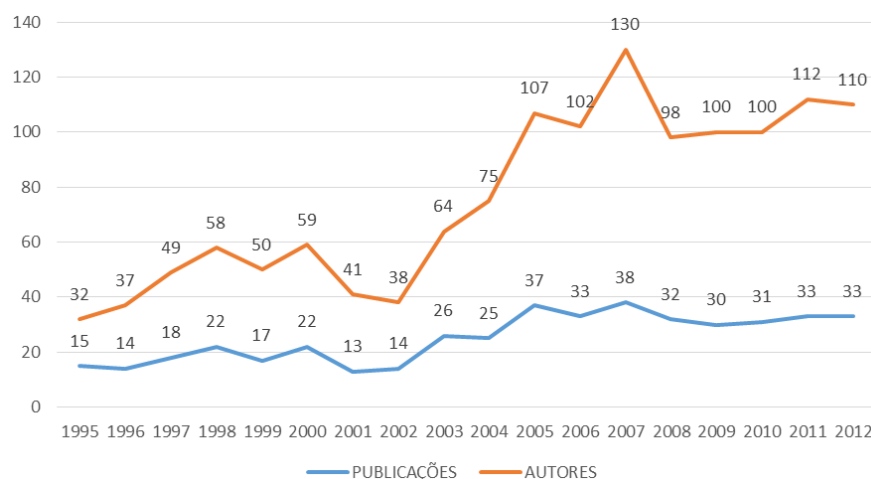


Figura 5.8: Relação entre autores e artigos

Observando a quantidade de publicações e autores por ano do evento geramos uma curva de crescimento de publicações e de autores do *WebMedia*, ilustrada na figura 5.9. É possível constatar que o número de autores cresceu de forma mais rápida que o número de artigos, o que mostra que o *WebMedia* tem atraído novos autores ao longo de sua existência.

Figura 5.9: Crescimento de publicações e autores do *WebMedia*.

5.2.2 Considerações

Este estudo processou os metadados dos 453 artigos publicados nas 18 edições do *WebMedia* realizadas entre os anos de 1995 e 2012 com o intuito de analisar a evolução da estrutura da

comunidade científica responsável pelos artigos publicados no evento.

O uso de técnicas de análise de redes sociais juntamente com exploração visual manual permitiram e facilitaram a combinação de elementos para entendimento e avaliação das estruturas de rede que emergem das relações de coautoria na produção dos artigos científicos. A análise de coautoria tornou visível a comunidade científica atuante no *WebMedia*, onde esta constitui um núcleo coeso que vem evoluindo desde as primeiras edições do evento.

Os dados mostrados indicam que há um visível crescimento na quantidade de novos autores com publicações no *WebMedia* porém a quantidade de publicações e submissões se estabilizou por volta de 2005.

Um fator que chama atenção é a baixa inserção e impacto internacional do evento. Apenas 28% dos artigos publicados nestas 18 edições anteriores foram escritos em inglês e apenas 19 dentre os 453 artigos aceitos para publicação contaram com a contribuição de autores estrangeiros.

5.3 Sistemas Embarcados

Esta seção é dedicada à apresentação dos resultados obtidos pela ferramenta *Paper Crawler* para a cadeia de busca “*Embedded Systems*” para a base de publicações *IEEE Xplore*, realizado no final do ano de 2014.

Na seção 5.3.1 realizamos diversas análises em cima dos dados obtidos pelo *Paper Crawler*, na seção 5.3.2 explanamos os resultados obtidos referentes a análise de relevância das publicações, utilizando o primeiro método de captura de referências descrito no Capítulo 4, na seção 5.3.3 descrevemos os resultados obtidos durante a classificação de comunidades da rede. E na seção 5.3.4 descrevemos algumas considerações alcançadas neste estudo.

5.3.1 Análise quantitativa da área

A partir dos dados retornados pela ferramenta realizamos a análise da evolução do número de publicações na área de **Sistemas Embarcados** ao longo dos anos, como pode ser ilustrado na Figura 5.10. Podemos observar que depois do ano de 2012 houve um declínio no número de publicações, possivelmente por dois fatores o primeiro que em alguns meios de publicações

há uma demora no envio de suas publicações para o *IEEE Xplore* e o segundo que realmente houve uma diminuição no desenvolvimento de pesquisas desta área.

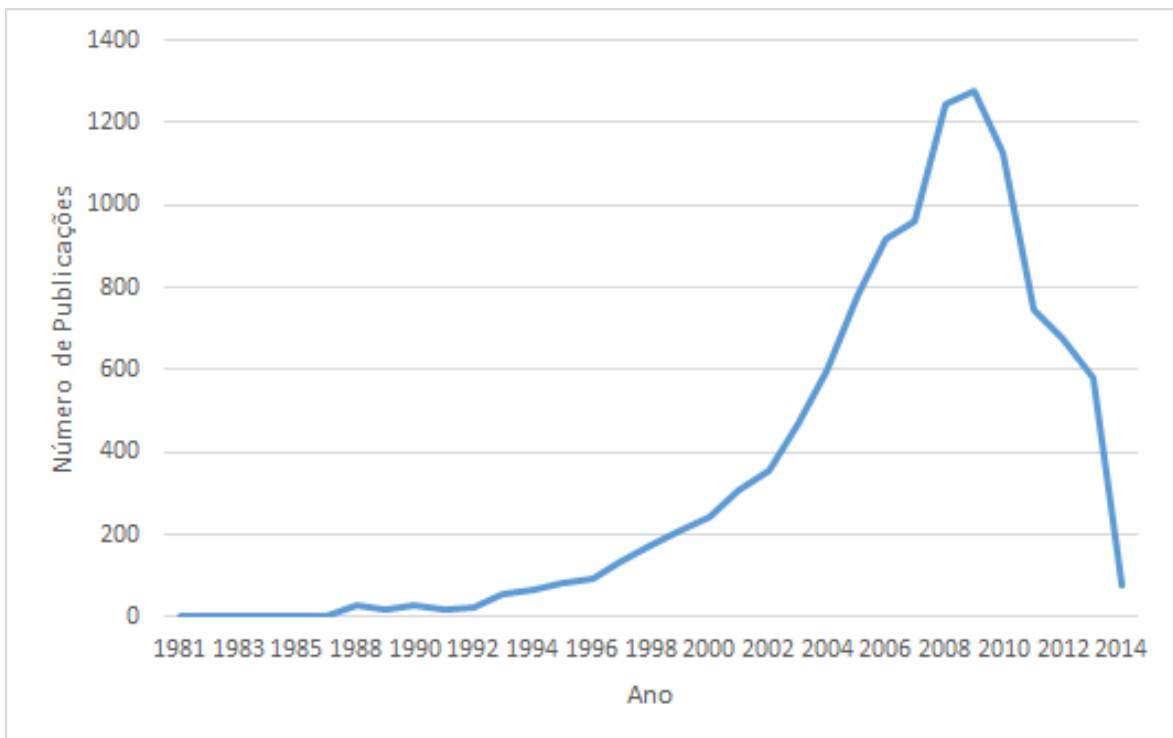


Figura 5.10: Evolução ao longo dos anos

As publicações obtidas foram publicadas em cerca de 11975 locais de publicação. Podemos observar na Figura 5.11 os quinze principais meios de publicação para as publicações consideradas no experimento.

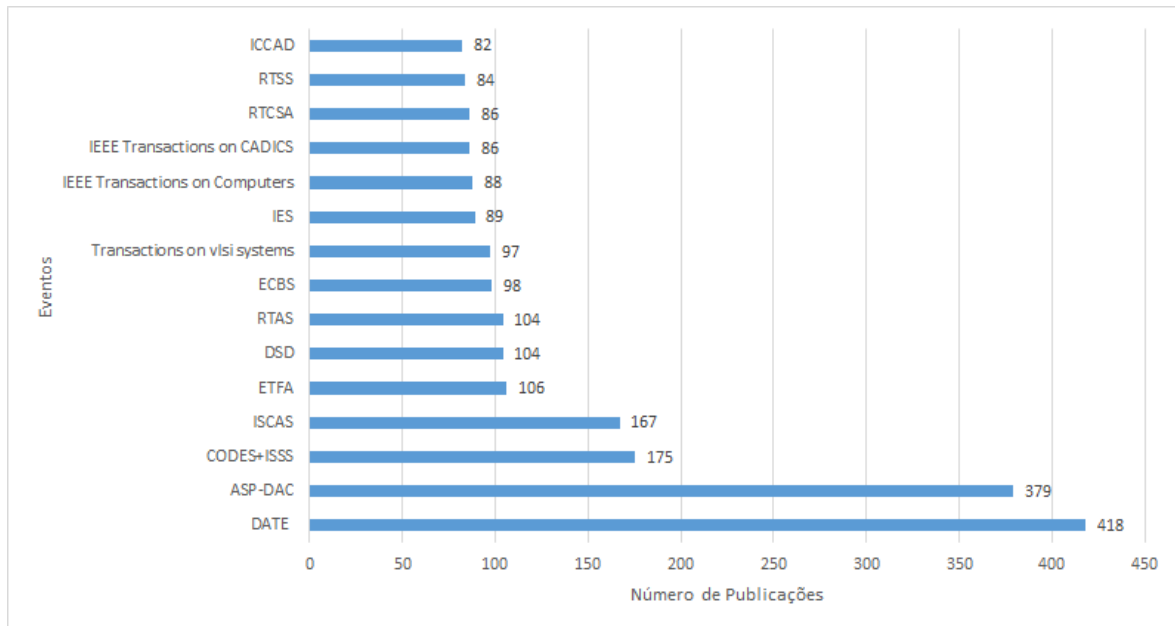


Figura 5.11: Meios de Publicação

O veículo de publicação que neste experimento possuiu o maior número de publicações foi o *DATE - Design Automation and Test in Europe*. Sendo este evento um dos mais conceituados na área de sistemas embarcados.

A partir dos meios de publicação identificamos a maneira como estes evoluíram ao longo dos anos, na Figura 5.12 está ilustrado os 5 locais de publicação mais relevantes e sua evolução ao longo dos anos

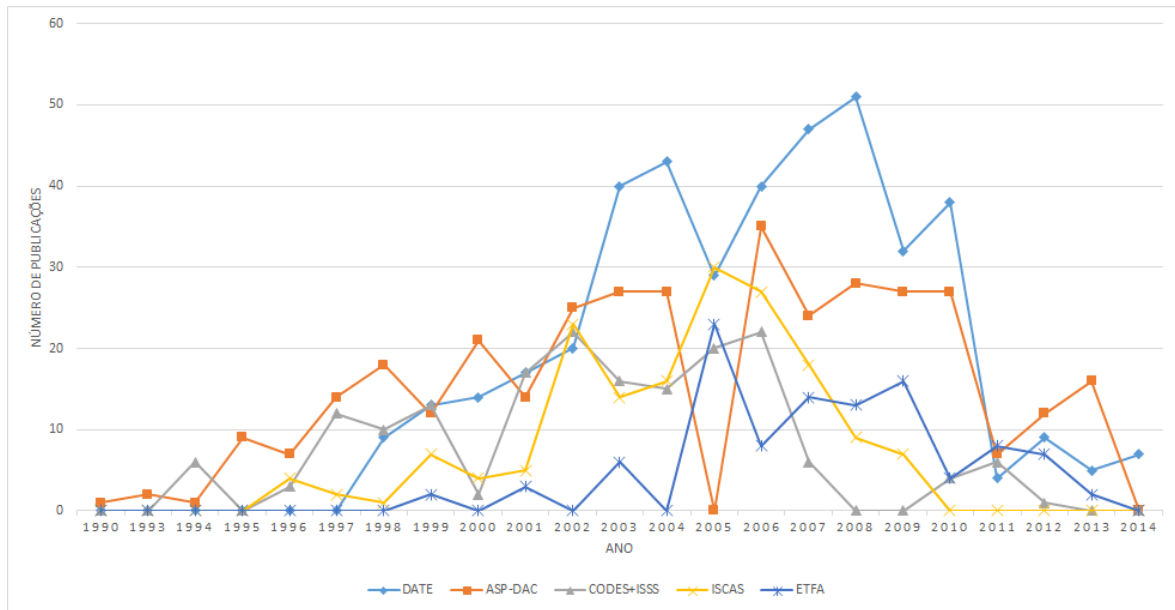
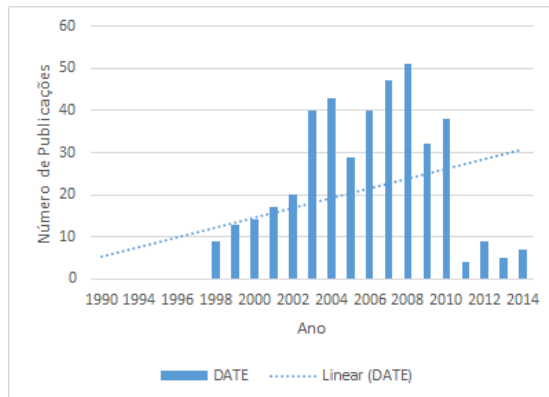


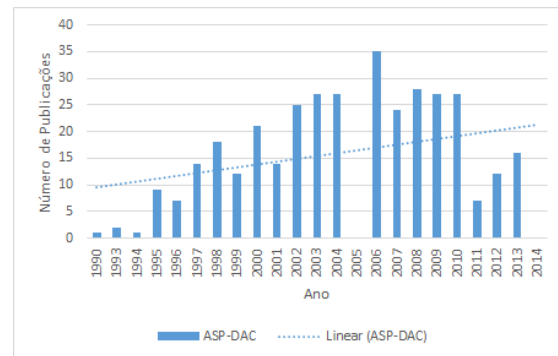
Figura 5.12: Evolução de Publicações dos Meios de Publicação

Podemos observar que o evento DATE teve um alto crescimento no número de publicações a partir de 2008, superando 50 publicações no ano 2008.

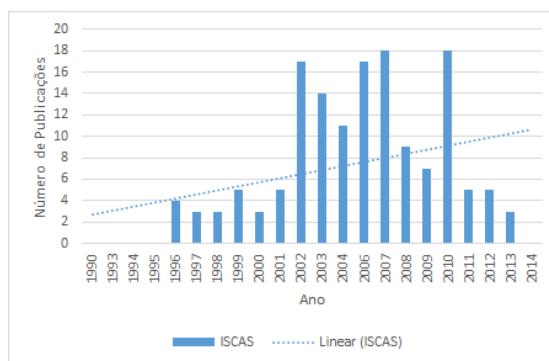
Como dito no início da seção há um decréscimo do número de publicações a partir de 2010, possivelmente por que em alguns meios de publicações demoram no envio de suas publicações para o *IEEE Xplore*. Então como analisar se estes eventos estão tendo uma tendência de crescimento? Para identificarmos isso, geramos a linha de tendência de publicações para cada evento, esta linha nos mostra a tendência de crescimento ou não do número de publicações para os meios de publicação analisados. Os cinco mais relevantes estão ilustrados na Figura 5.13.



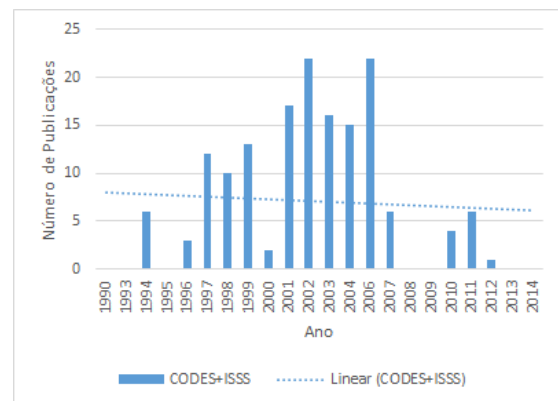
(a) Tendência do Date



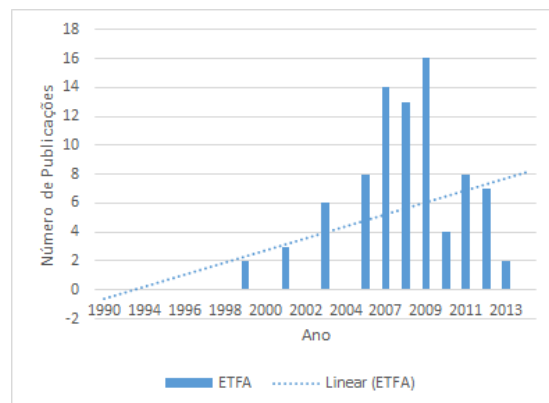
(b) Tendência do Asp-Dac



(c) Tendência do Iscas



(d) Tendência do Codes+Isss



(e) Tendência do Etfa

Figura 5.13: Tendência dos meios de publicação mais relevantes

Observamos que dos 5 eventos ilustrados na Figura 5.13 o único em que a linha de tendência está em decréscimo é o meio de publicação *Codes+Isss - International Conference on Hardware/Software Codesign and System Synthesis* ilustrado na Figura 5.13(d).

5.3.2 Análise da Rede de Citações

A partir da cadeia de busca a ferramenta obteve 11303 publicações. Em seguida, o sistema extraiu as referências de cada artigo recuperado desde que estas também estivessem listadas nas publicações obtidas inicialmente, gerando assim uma rede de citações com 11303 nós e 6589 arestas. Na Figura 5.14 podemos observar uma das visualizações da rede a criadas partir do *Gephi*.

Para obter a visualização da rede inicialmente particionamos a rede pela métrica modularidade, deixando assim visível as comunidades presentes na rede, em seguida era necessário destacar os nós mais importantes classificando-os pelo tamanho através da métrica desejada (neste caso o grau de entrada, *page rank* e centralidade de intermediação) e finalmente distribuir a rede para que esta se tornasse visualmente mais fácil de analisar, para isso utilizamos dois algoritmos de distribuição o *yifan hu* proporcional (HU, 2005) e *noverlap*.

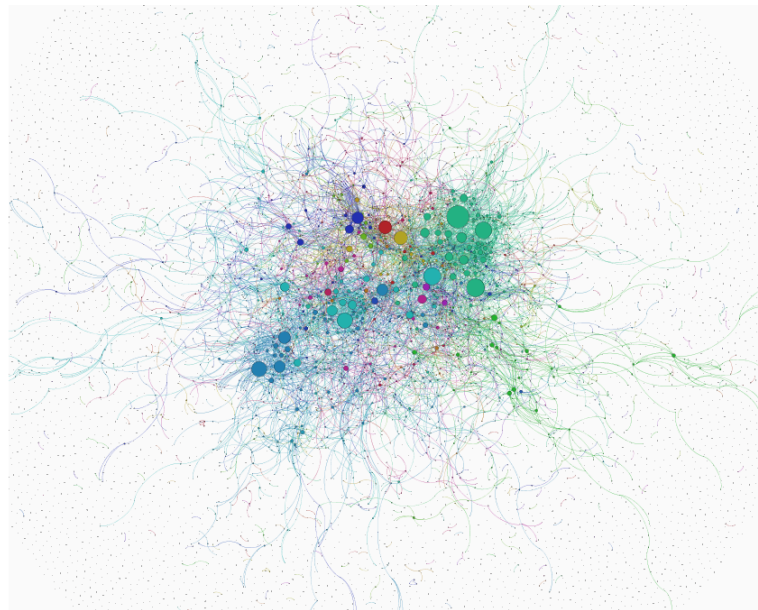


Figura 5.14: Rede classificada por Grau de entrada

Então a partir dos dados obtidos calculamos a distribuição de publicações, onde identificamos as publicações que possuíam zero, um e dois ou mais citações, o resultado do cálculo está ilustrado na Figura 5.15.

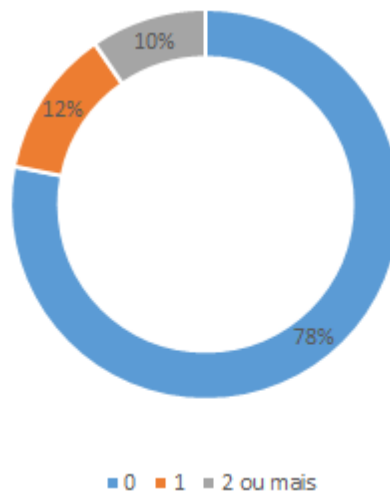


Figura 5.15: Distribuição de publicações

Realizando uma análise a Figura 5.15 podemos identificar que 10% das publicações foram citadas duas ou mais vezes, 12% das publicações foram citadas uma vez e por fim que 78% das publicações não foram citadas. Significando que em sua maioria as publicações citadas não estão indexadas pelo repositório escolhido.

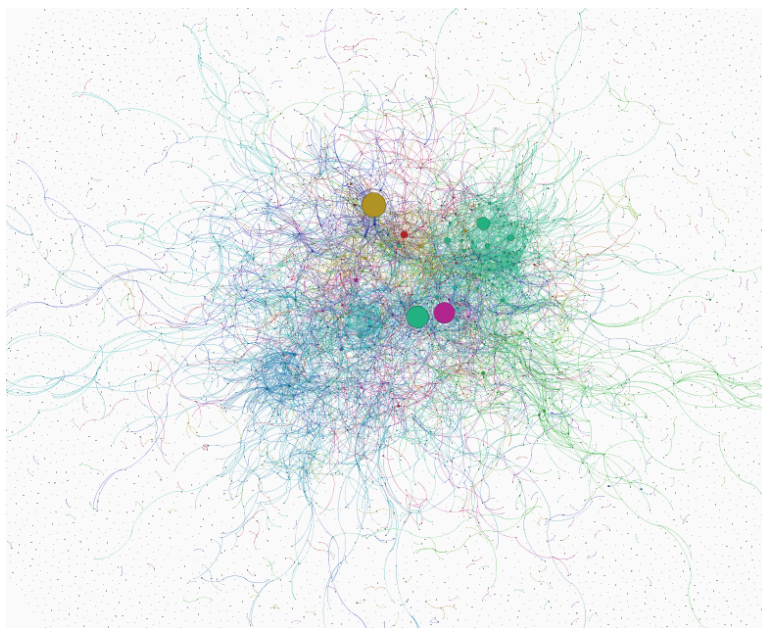


Figura 5.16: Rede classificada por Page Rank

Observando a Figura 5.16 podemos identificar os nós com mais *Page Rank*, estes são bem centrais a rede então caso removermos partes importantes da rede deixam de se comunicar.

Na Tabela 5.5 estão descritas as 10 publicações com o maior número de citações e as 10

publicações com o maior *Page Rank*.

(a) Publicações com mais citações		(b) Publicações com mais <i>Page Rank</i>	
Publicação	Número de Citações	Publicação	<i>Page Rank</i>
(GUPTA; MICHELI, 1993)	71	(ZAVE, 1982)	1,71E-02
(DICK; RHODES; WOLF, 1998)	58	(LI; MALIK; WOLFE, 1995)	1,56E-02
(TIWARI; MALIK; WOLFE, 1994)	55	(LI, 1995)	1,47E-02
(WOLF, 1994)	53	(GUPTA; MICHELI, 1993)	8,67E-03
(WOLFE; CHANIN, 1992)	49	(HAREL et al., 1988)	4,83E-03
(BANAKAR et al., 2002)	48	(WOLFE; CHANIN, 1992)	3,94E-03
(HAREL et al., 1988)	42	(CHIODO et al., 1994)	3,75E-03
(EDWARDS et al., 1997)	42	(YEN; WOLF, 1995)	3,20E-03
(PANDA; DUTT; NICOLAU, 1997b)	39	(WOLFE; CHANIN, 1992)	2,75E-03
(KEUTZER et al., 2000a)	37	(TIWARI; MALIK; WOLFE, 1994)	2,52E-03

Tabela 5.5: Publicações mais relevantes

Analisando a Tabela 5.5 podemos verificar que a publicação mais citada foi a Gupta e Micheli (1993) com 71 citações. Se analisarmos a tabela 5.5(b), porque esta mesma citação não ficou com o maior *Page Rank*? O *Page Rank* identifica os nós que são mais citados por artigos também muito citados, significando que a publicação em questão não atingiu este critério quando comparada com os resultados da publicação Zave (1982), logo o *Page Rank* pode ser considerado uma métrica mais apropriada para análise de relevância segundo Cui, Wang e Zhai (2010) e Maslov e Redner (2008), então as publicações Zave (1982), Li, Malik e Wolfe (1995), Li (1995) são mais relevantes que a publicação Gupta e Micheli (1993).

5.3.3 Classificação das Comunidades

A partir do cálculo da métrica modularidade definimos as comunidades da rede, foram identificadas 7257 comunidades, das quais consideramos mais relevantes neste estudo as que possuíam o maior número de nós, sendo as 5 maiores descritas na Tabela 5.6 e ilustradas na Figura 4.1.

Tabela 5.6: 5 maiores comunidades			
Comunidade	Qde nós	Arestas	Densidade
A	382	844	0,006
B	269	499	0,007
C	205	276	0,007
D	199	280	0,007
E	180	260	0,008

Filtrando a rede da Figura 5.14 para que possamos visualizar somente as maiores co-

munidades descritas na Tabela 5.6 temos como resultado a ilustração da Figura 5.17. Nela podemos visualizar os nós com maiores citações pertencentes a cada comunidade, como também a interação entre elas.



Figura 5.17: 5 maiores comunidades

A ferramenta também calculou média das métricas grau de entrada, centralidade de intermediação e *page rank* para as comunidades, estando descrito na tabela 5.7 os valores destas métricas para as 5 comunidades mais relevantes.

Tabela 5.7: Média das Métricas para as 5 comunidades mais relevantes

Comunidade	Média		
	Grau de Entrada	Centralidade de Intermediação	<i>Page Rank</i>
A	2,75E+00	1,74E+02	2,53E-04
B	2,07E+00	1,73E+01	1,26E-04
C	1,64E+00	8,23E+01	1,16E-04
D	1,57E+00	2,98E+01	1,08E-04
E	1,69E+00	1,72E+01	1,34E-04

Com a identificação das comunidades realizamos a análise da evolução do número de publicações por comunidade ao longo dos anos, entre as 5 mais relevantes o crescimento foi bem parecido entre elas, mas a comunidade B teve um crescimento de maior destaque nos anos de 1993 a 2000.

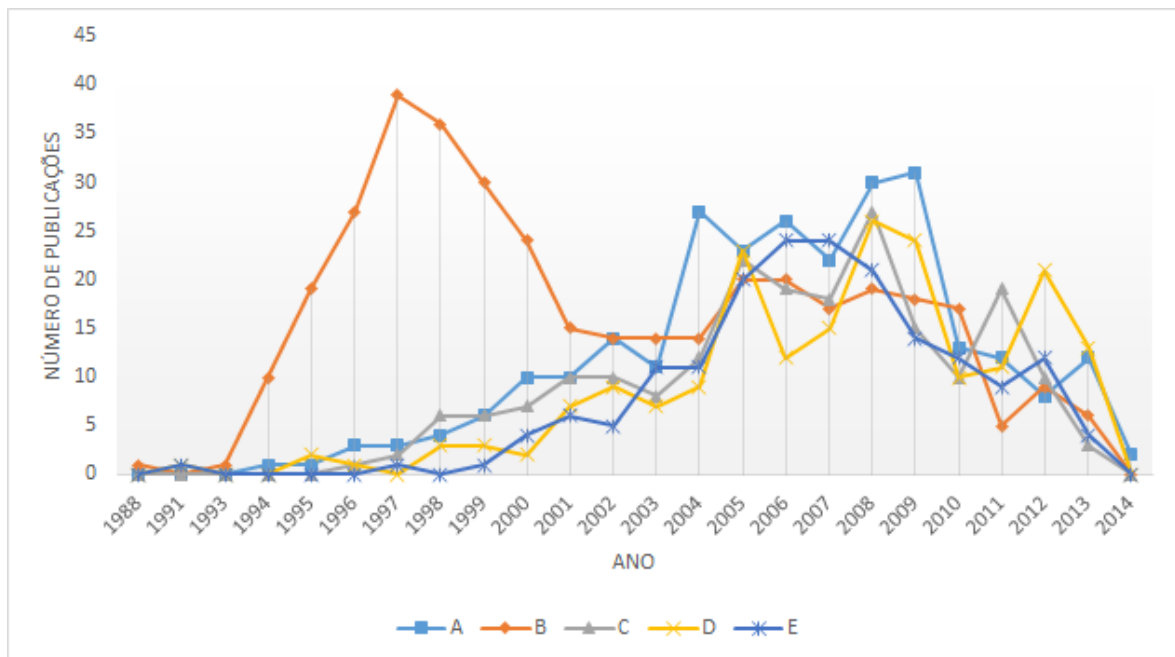


Figura 5.18: Evolução de Publicações das 5 maiores comunidades

Para verificar a tendência de crescimento de cada comunidade geramos a linha de tendência ilustrada na Figura 5.19, onde verificamos que a comunidade com mais tendência de ter um decréscimo no número de publicações nos próximos anos foi a comunidade B ilustrada na Figura 5.19(b).

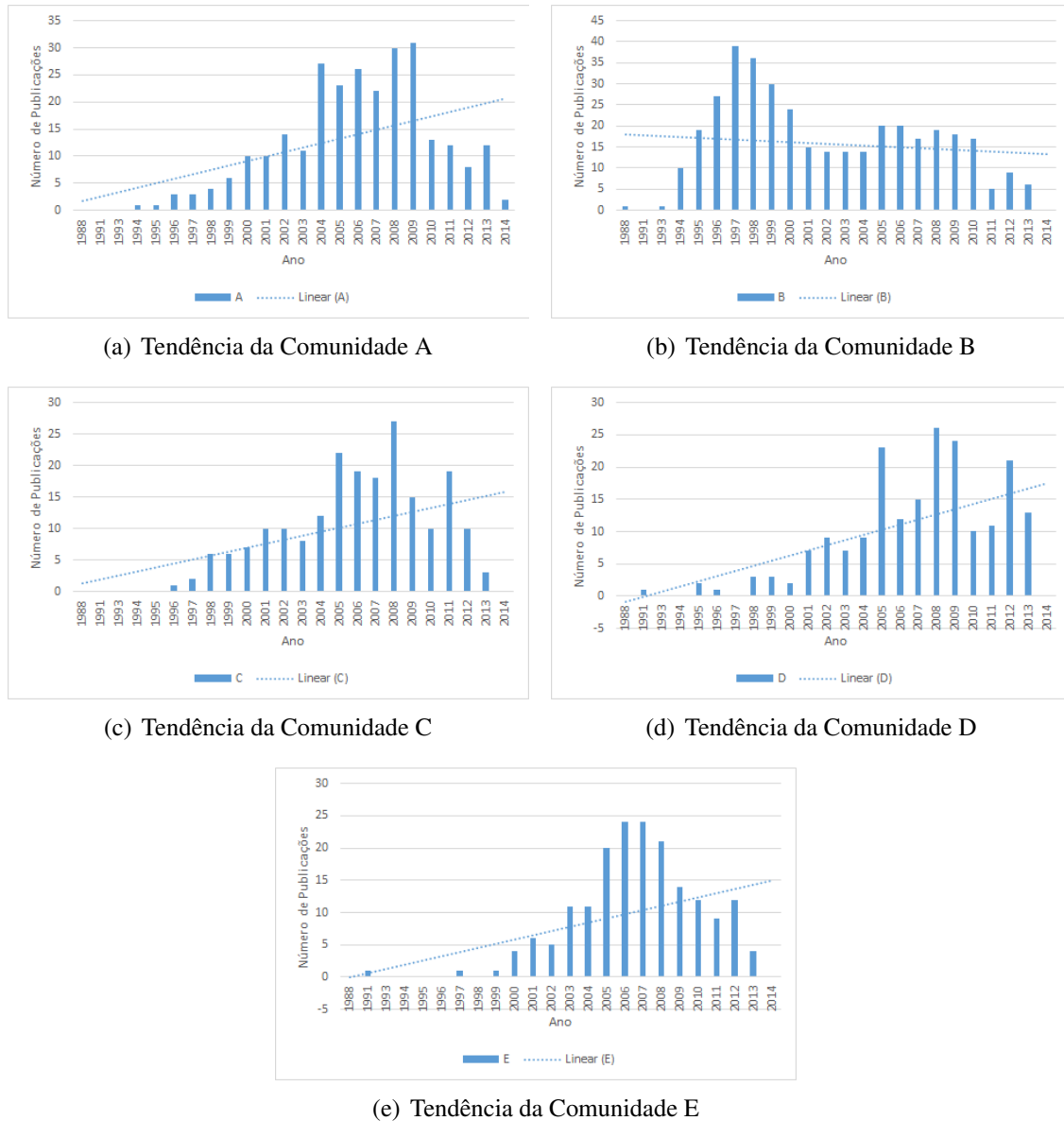


Figura 5.19: Tendência das comunidades mais relevantes

Para realizar a classificação das comunidades primeiramente identificamos as *keywords* das publicações presentes em cada comunidade mais o número de vezes que estas apareciam. Foram desconsideradas as palavras-chave *Embedded System*, *Embedded Software*, *Hardware* e *Application Software*, pois estas eram muito comuns em todas as comunidades. Com isso cada comunidade foi classificada conforme a *keyword* mais relevante possível, a classificação para as cinco comunidades mais relevantes está descrita nas Tabelas 5.8, 5.9, 5.10, 5.11 e 5.12.

Tabela 5.8: 5 keywords mais relevantes para comunidade A

<i>Keyword</i>	Número de referências
<i>Embedded system</i>	334
<i>Hardware</i>	229
<i>Embedded software</i>	92
<i>Costs</i>	88
<i>Application software</i>	84

Tabela 5.9: 5 keywords mais relevantes para comunidade B

<i>Keyword</i>	Número de referências
<i>Embedded system</i>	227
<i>Hardware</i>	82
<i>Energy consumption</i>	78
<i>Application software</i>	74
<i>Computer science</i>	51

Tabela 5.10: 5 keywords mais relevantes para comunidade C

<i>Keyword</i>	Número de referências
<i>Embedded system</i>	156
<i>Hardware</i>	98
<i>Application software</i>	56
<i>Computer architecture</i>	50
<i>Embedded software</i>	31

Tabela 5.11: 5 keywords mais relevantes para comunidade D

<i>Keyword</i>	Número de referências
<i>Embedded system</i>	138
<i>Hardware</i>	62
<i>Application software</i>	53
<i>Real time systems</i>	51
<i>Embedded software</i>	51

Tabela 5.12: 5 keywords mais relevantes para comunidade E

<i>Keyword</i>	Número de referências
<i>Embedded system</i>	143
<i>Energy consumption</i>	84
<i>Processor scheduling</i>	64
<i>Voltage control</i>	51

Também identificamos as publicações mais relevantes presentes em cada comunidade,

considerando a métrica *Page Rank*. Nas Tabelas 5.13, 5.14, 5.15, 5.16 e 5.17 estão descritas as publicações mais relevantes de cada comunidade.

Tabela 5.13: 5 publicações mais relevantes da comunidade A - *Costs*

Publicação	<i>Page Rank</i>
(LI; MALIK; WOLFE, 1995)	1,56E-02
(GUPTA; MICHELI, 1993)	8,67E-03
(WOLF, 1994)	3,94E-03
(CHIODO et al., 1994)	3,75E-03
(YEN; WOLF, 1995)	3,20E-03

Tabela 5.14: 5 publicações mais relevantes da comunidade B - *Energy Consumption*

Publicação	<i>Page Rank</i>
(PANDA; DUTT; NICOLAU, 1996)	2,35E-03
(PANDA; DUTT; NICOLAU, 1997b)	1,91E-03
(PANDA; DUTT; NICOLAU, 1997a)	1,16E-03
(BANAKAR et al., 2002)	1,11E-03
(STEINKE et al., 2002)	1,02E-03

Tabela 5.15: 5 publicações mais relevantes da comunidade C - *Computer architecture*

Publicação	<i>Page Rank</i>
(KEUTZER et al., 2000b)	1,39E-03
(SANGIOVANNI-VINCENTELLI; MARTIN, 2001)	7,60E-04
(SGROI; LAVAGNO; SANGIOVANNI-VINCENTELLI, 2000)	7,33E-04
(PIMENTEL; ERBAS; POLSTRA, 2006)	6,97E-04
(ITO; CARRO; JACOBI, 2001)	6,19E-04

Tabela 5.16: 5 publicações mais relevantes da comunidade D - *Real time systems*

Publicacao	<i>Page Rank</i>
(KARSAI et al., 2003)	8,80E-04
(EKER et al., 2003)	7,47E-04
(HERRERA et al., 2003)	5,22E-04
(MARTIN; LAVAGNO; LOUIS-GUERIN, 2001)	4,60E-04
(KUKKALA et al., 2005)	4,16E-04

Tabela 5.17: 5 publicações mais relevantes da comunidade E - *Processor scheduling*

Publicação	Page Rank
(KRISHNA; LEE, 2000)	2,26E-03
(LOCKE; VOGEL; MESLER, 1991)	1,73E-03
(SHIN; CHOI; SAKURAI, 2000)	51,45E-03
(LORENTE; LIPARI; BINI, 2006)	8,82E-04
(ZHU; MUELLER, 2004)	7,09E-04

5.3.4 Considerações

Este estudo processou os metadados dos 11303 artigos publicados em diversos meios de publicação realizados entre os anos de 1981 e 2014 com o intuito de analisar as publicações mais relevantes durante este período e as comunidades que foram mais relevantes, na área de sistemas embarcados.

É importante ressaltar que em todas as fases deste estudo a metodologia como também a ferramenta foram de crucial importância para o seu desenvolvimento, a partir dos dados extraídos a ferramenta já nos dava um conjunto de estatísticas sobre a análise da área em questão.

A partir dos dados analisados é possível afirmar que a área de sistema embarcados está com um crescimento menor nos últimos anos, mas apesar disso os meios de publicações relevantes identificados em sua maioria tem uma tendência de crescimento.

A classificação das comunidades relevantes da rede nos mostrou as sub-áreas que no nicho estudado possuem o maior numero de publicações, onde podemos afirmar que estas possuem relevância na área de Sistemas Embarcados. O método utilizado para classificar as comunidades através das palavras-chave pertencentes a cada publicação destas comunidades facilitou o processo. Verificamos que em sua maior parte as publicações pertencentes a uma comunidade realmente abordam algum assunto de proximidade a classificação dada. Isso futuramente permitirá descobrir a partir dos metadados de uma publicação fora da rede em que comunidade este se encaixa.

5.4 Publicações

Até o presente momento nosso trabalho resultou em duas publicações. A primeira publicação apresentou genericamente a proposta inicial de trabalho, revisão bibliográfica sobre redes

sociais, as metodologias existentes para análise e os resultados preliminares obtidos através dos dados área de Análise de Redes Sociais. O artigo foi denominado “Uma Ferramenta para Extração Semi-Automática e Análise de Relevância de Artigos Científicos”, apresentado durante o III *Brazilian Workshop on Social Network Analysis and Mining* (III BRASNAM) componente do Congresso da Sociedade Brasileira de Computação (CSBC) realizado no mês de julho do ano de 2014 na cidade de Brasília-DF.

Na segunda publicação relacionada ao nosso projeto de pesquisa conseguimos aprofundar os resultados obtidos no trabalho publicado descrito anteriormente e ainda, conseguimos levantar novas hipóteses a cerca da problemática em questão. O artigo foi denominado “WebMedia XX Quem Somos e o Que Temos Feito nas Últimas Duas Décadas” e foi aceito e apresentado durante as atividades do Simpósio de Multimídia e Web (WebMedia 2014) realizado no mês de novembro do ano de 2014 João Pessoa-PB.

Finalmente, encerrando o ciclo de artefatos gerados com este tema de estudo, pretendemos compilar os resultados obtidos da área de Sistemas Embarcados em um artigo completo que deverá ser submetido para alguma revista ligada a área.

Capítulo 6

Considerações Finais

Neste capítulo serão discutidas as principais conclusões obtidas a partir dos estudos de caso apresentados no Capítulo 5. Desta forma o presente capítulo encontra-se dividido em três seções: Em Ameaças à validade serão apontados os principais fatores que dificultaram o desenvolvimento da presente pesquisa; Em Discussão ampliaremos o debate acerca das conclusões obtidas e expostas; Finalmente em Trabalhos Futuros estão descritas algumas sugestões de pesquisas.

6.1 Ameaças à validade

Cabe o comentário de que embora fosse realizado um cuidadoso planejamento, a presente pesquisa pode ter sido afetada por diferentes tipos de fatores. Adicionalmente, tais fatores resultados associados a essa pesquisa. Os fatores estão relacionados a (i) validade externa: Tamanho e diversidade da amostra, escolha dos repositórios de artigos, desambiguação de nomes, analisadores de citações online (ii) validade interna: Escolha de métricas. Cada um dos fatores bem como as ações para mitigar seus impactos sob os resultados da presente pesquisa, são descritos abaixo:

1. **Tamanho e diversidade da amostra** - Para o estudo de caso da Seção 5.1 sobre ARS no primeiro experimento obtivemos uma quantidade razoável de publicações para serem analisadas, mas verificando o segundo experimento obtivemos uma quantidade pequena de publicações e ainda estas pouco se citavam, fazendo com que neste caso algumas dificuldades na identificação de publicações relevantes fossem encontradas,

para solucionar isso no estudo de caso da Seção 5.3 decidimos analisar publicações de uma área que tivesse uma maior quantidade de publicações, por isso escolhemos a área de sistemas embarcados. Com isso, obtivemos um nicho bem maior de publicações e uma rede bem mais conectada, indicando que estas publicações se citavam tornando a análise bem mais interessante.

2. **Escolha dos repositórios de artigos** - A escolha do repositório *IEEE Xplorer* se deu por que este era o mais simples de realização o processo de paginação e assim obter o resultado, mas apesar disso essa decisão fez que limitássemos os dados a esta base. Mesmo assim sua utilização deu uma visão geral do resultados obtidos nos estudos de caso. Para solucionar isso deixamos a ferramenta modularizada, onde caso houvesse a necessidade de incluir novos repositórios fosse possível. Outro problema em relação aos repositórios é que estes mudam com alguma frequência a estrutura de seus sites invalidando os analisadores de html utilizados.
3. **Desambiguação de nomes** - No início queríamos realizar a análise dos autores mais relevantes também, utilizando redes de coautoria, mas como o *IEEE Xplore* não padronizava os metadados dos autores tínhamos o problema da desambiguação de nomes e portanto não foi possível realizar isso para os estudos de caso seção 5.1 e seção 5.3. No caso do estudo de caso sobre os anais do *webmedia* (seção 5.2), isto foi possível por que os metadados referentes aos autores estavam bem padronizados facilitando a análise.
4. **Analisadores de Citações Online** Durante o percorrer do estudo de caso sobre análise de redes sociais (Seção 5.1) identificamos que a viabilidade de se obter os metadados de todas as citações de uma publicação no *IEEE Xplore* é muito baixa, pois mesmo com a utilização de analisadores de citação online a taxa de acerto não era muito boa, principalmente por problemas na padronização dos textos de referências presentes no HTML do *IEEE Xplore*, fora a demora na obtenção dos dados. Por isso no estudo posterior sobre Sistemas Embarcados (Seção 5.3) decidimos obter somente citações que são indexadas pelo *IEEE Xplore* e que estão presentes na lista de publicação obtidas a partir da cadeia de busca, evitando assim obtenção de estudos sobre outras áreas de conhecimento como ocorreu no estudo de caso sobre *ARS*.

5. **Escolha de métricas** - Existem inúmeras métricas (i.e *page rank*, grau de entrada, centralidade de intermediação, índice h, *qualis*, *cite rank*, entre outras) quando se trata da identificação de relevância de publicações, autores e veículos de publicação. Neste estudo procuramos utilizar métricas de ARS conhecidas e de simples entendimento que nos desse um resultado passível de validação, mesmo assim o foco do estudo não é a métrica em si mas validar através delas que é possível encontrar artigos que fossem relevantes a área de conhecimento analisada através de uma metodologia.

6.2 Discussão

Com base nos dados apurados, pode-se afirmar que a partir da construção e análise de uma rede composta de publicações e suas citações, existe a possibilidade de fazer uso de métricas de ARS para obter informações importantes a respeito da produção científica de determinada área e identificar suas publicações mais relevantes. Este estudo se mostrou satisfatório, uma vez que conseguiu identificar publicações relevantes e classificar comunidades de uma área de conhecimento, tudo isso utilizando uma metodologia específica para análise de relevância de publicações, possibilitando assim dar suporte a pesquisadores que estejam iniciando sua pesquisa e precisam de algum *background* para ela.

A partir da utilização da ferramenta *Paper Crawler* tivemos todo suporte necessário para análise dos dados obtidos, suas funcionalidades em conjunto com a metodologia definida permitiram a capacidade de automatizar todo o processo analisado nos estudos de caso descritos no Capítulo 5. Apesar da existência de outras métricas para clusterização de rede, a utilização da métrica modularidade nos foi satisfatória, sua capacidade de identificar grupos na rede fez com que pudéssemos visualizar a interação entre estas comunidades.

Inicialmente na identificação das publicações mais relevantes utilizamos a métrica grau de entrada que permitia identificar o número de citações que uma publicação recebeu, mas apesar de ser um indicativo de relevância era mais interessante utilizar métricas de centralidade como o *Page Rank* que nos permite identificar as publicações que eram citadas por publicações que possuíam muitas citações.

Nossa questão de pesquisa foi “Pode o uso de técnicas de análise de redes sociais auxiliarem a mineração de conteúdo relevante associado a publicações científicas?”, através do

decorrer deste estudo tentamos responde-la, alcançando as hipóteses traçadas no Capítulo 4. Podemos afirmar que estas a partir dos resultados explanados nos estudos de caso foram verificadas.

A **hipótese H1** afirmava que através de uma metodologia baseada em redes sociais poderíamos identificar meios de publicação que fossem relevantes para a área pesquisada, como foi observado nos estudos de caso 5.1 e 5.3 conseguimos identificar os meios de publicação que possuíam o maior número de publicações.

A **hipótese H2** afirmava que através de uma metodologia baseada em redes sociais poderíamos identificar publicação científicas que fossem relevantes para a área pesquisada, os resultados que comprovam isso foram observados nos estudos de caso 5.1 e 5.3, lá conseguimos identificar publicações científicas relevantes através do número de citações e também do *page rank*. Já a **hipótese H3** dizia que utilizar métricas de centralidade são bons indicadores na identificação de publicações relevantes, isso foi mostrado a partir de estudos correlacionados e da própria natureza da métrica *page rank*.

A **hipótese H4** afirmava que através de métricas centralidade seria possível analisar a evolução de uma comunidade científica ao longo do tempo, como foi observado no estudo de caso 5.2 analisamos o crescimento da comunidade científica do *Webmedia*.

E por fim, a **hipótese H5** afirmava que era possível através de uma metodologia baseada em redes sociais realizar a classificação de comunidade de uma rede de citações. No estudo de caso 5.3 foi possível observar que a utilização das palavras-chave contidas nas publicações poderiam ser utilizadas para nomear as comunidades da rede e assim encontrar sub-áreas dentro da área de conhecimento analisada.

6.3 Trabalhos Futuros

Como Trabalhos Futuros vamos tentar resolver a desambiguação de nomes para os autores permitindo assim uma análise da evolução da comunidade científica. Desenvolver novos módulos para as outras bibliotecas digitais, fazendo com que obtenhamos o maior número de resultados possíveis. Também pretendemos buscar outras métricas que possibilitem a identificação de publicações relevantes, comparado seu desempenho as métricas utilizadas nesta pesquisa. E por último modificar a ferramenta *Paper Crawler* para que fique apropriada

para um ambiente *WEB*.

Bibliografia

ABREU-RODRIGUES, J. A qualidade da publicação científica. *Psicologia: Teoria e Pesquisa*, scielo, v. 25, p. iii – iv, 03 2009. ISSN 0102-3772. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-37722009000100001&nrm=iso>.

ALBA, R. D. A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, v. 3, p. 3–113, 1973. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.89.8697>>.

ALVES, A. D.; YANASSE, H. H.; SOMA, N. Y. Sucupira: A system for information extraction of the lattes platform to identify academic social networks. In: *Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on*. [S.l.: s.n.], 2011. p. 1–6.

BANAKAR, R. et al. Scratchpad memory: a design alternative for cache on-chip memory in embedded systems. In: *Hardware/Software Codesign, 2002. CODES 2002. Proceedings of the Tenth International Symposium on*. [S.l.: s.n.], 2002. p. 73–78.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009. Disponível em: <<http://www.aiai.org/ocs/index.php/ICWSM/09/paper/view/154>>.

BERKOWITZ, S. D. *An introduction to structural analysis: The network approach to social research*. [S.l.]: Butterworth-Heinemann, 1982.

BOLLOBÁS, B. The evolution of random graphs—the giant component. In: *Random Graphs*. Second. Cambridge University Press, 2001. p. 130–159. ISBN 9780511814068. Cambridge Books Online. Disponível em: <<http://dx.doi.org/10.1017/CBO9780511814068.008>>.

BRANDES, U. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, v. 25, p. 163–177, 2001.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, v. 30, n. 1-7, p. 107–117, 1998.

CAMARGO, L. S. d. A.; VIDOTTI, S. A. B. G. Elementos de personificação em repositórios institucionais. *Conferência Ibero-Americana de publicações eletrônicas no contexto da comunicação científica*, 2006.

CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. de. Identificando a reputação de pesquisadores usando um modelo de perfil adaptativo. *XXXVIII SEMISH*, 2011.

CHIODO, M. et al. Hardware-software codesign of embedded systems. *Micro, IEEE*, v. 14, n. 4, p. 26–36, Aug 1994.

CRILLY, T. *50 Mathematical Ideas You Really Need to Know*. Book Sales, 2007. ISBN 9781847241474. Disponível em: <<http://books.google.com.br/books?id=XfYSKAAACAAJ>>.

CUI, J.; WANG, F.; ZHAI, J. Citation networks as a multi-layer graph: Link prediction and importance ranking. 2010.

DEITEL, H. M. *Java: como programar*. Pearson Prentice Hall, 2005. ISBN 9788576050193. Disponível em: <<http://books.google.com.br/books?id=U5AyAgAACAAJ>>.

DICK, R. P.; RHODES, D. L.; WOLF, W. Tgff: task graphs for free. In: *Hardware/Software Codesign, 1998. (CODES/CASHE '98) Proceedings of the Sixth International Workshop on*. [S.l.: s.n.], 1998. p. 97–101.

EDWARDS, S. et al. Design of embedded systems: formal models, validation, and synthesis. *Proceedings of the IEEE*, v. 85, n. 3, p. 366–390, Mar 1997.

EKER, J. et al. Taming heterogeneity - the ptolemy approach. *Proceedings of the IEEE*, v. 91, n. 1, p. 127–144, Jan 2003.

ERLIN, B.; YUSOF, N.; RAHMAN, A. A. Analyzing online asynchronous discussion using content and social network analysis. In: *ISDA*. IEEE Computer Society, 2009. p. 872–877. ISBN 978-0-7695-3872-3. Disponível em: <<http://dblp.uni-trier.de/db/conf/isda/isda2009.html#ErlinYR09>>.

FISHER, D. Using egocentric networks to understand communication. *IEEE Internet Computing*, v. 9, n. 5, p. 20–28, 2005. Disponível em: <<http://dblp.uni-trier.de/db/journals/internet/internet9.html#Fisher05>>.

FREEMAN, L. C. A set of measures of centrality based on betweenness. *Sociometry*, v. 40, p. 35–41, 1977.

FREITAS, M. C.; PEREIRA, H. B. B. Contribuição da análise de redes sociais para o estudo sobre os fluxos de informações e conhecimento. 6º Encontro Nacional de Ciência da Informação (CINFORM), 2005.

GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Science*, v. 99, n. 12, p. 7821–7826, 2002.

GLOOR, P. A. et al. Web science 2. 0: Identifying trends through semantic social network analysis. In: *CSE (4)*. IEEE Computer Society, 2009. p. 215–222. Disponível em: <<http://dblp.uni-trier.de/db/conf/cse/cse2009-4.html#GloorKNFS09>>.

GUEDES, A. C. S. M.; DUARTE, A. N. Um estudo sobre os impactos dos relacionamentos sociais na avaliação do mérito científico. *19th Brazilian Symposium on Multimedia and the Web*, 2013.

GUPTA, R. K.; MICHELI, G. D. Hardware-software cosynthesis for digital systems. *Design Test of Computers, IEEE*, v. 10, n. 3, p. 29–41, Sept 1993.

HAREL, D. et al. Statemate: a working environment for the development of complex reactive systems. In: *Computer Systems and Software Engineering, 1988. Proceedings. , Third Israel Conference on*. [S.l.: s.n.], 1988. p. 1–3.

HERRERA, F. et al. Systematic embedded software generation from systemc. In: *Design, Automation and Test in Europe Conference and Exhibition, 2003*. [S.l.: s.n.], 2003. p. 142–147.

HU, Y. F. Efficient and high quality force-directed graph drawing. *The Mathematica Journal*, v. 10, p. 37–71, 2005.

ITO, S. A.; CARRO, L.; JACOBI, R. P. Making java work for microcontroller applications. *Design Test of Computers, IEEE*, v. 18, n. 5, p. 100–110, Sep 2001.

KARSAI, G. et al. Model-integrated development of embedded software. *Proceedings of the IEEE*, v. 91, n. 1, p. 145–164, Jan 2003.

KEUTZER, K. et al. System-level design: orthogonalization of concerns and platform-based design. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, v. 19, n. 12, p. 1523–1543, Dec 2000.

KEUTZER, K. et al. System-level design: orthogonalization of concerns and platform-based design. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, v. 19, n. 12, p. 1523–1543, Dec 2000.

KRISHNA, C. M.; LEE, Y. H. Voltage-clock-scaling adaptive scheduling techniques for low power in hard real-time systems. In: *Real-Time Technology and Applications Symposium, 2000. RTAS 2000. Proceedings. Sixth IEEE*. [S.l.: s.n.], 2000. p. 156–165.

KUKKALA, P. et al. Uml 2. 0 profile for embedded system design. In: *Design, Automation and Test in Europe, 2005. Proceedings*. [S.l.: s.n.], 2005. p. 710–715 Vol. 2.

LI, S. M. Y.-T. S. Performance analysis of embedded software using implicit path enumeration. In: *Design Automation, 1995. DAC '95. 32nd Conference on*. [S.l.: s.n.], 1995. p. 456–461.

LI, Y. T. S.; MALIK, S.; WOLFE, A. Performance estimation of embedded software with instruction cache modeling. In: *Computer-Aided Design, 1995. ICCAD-95. Digest of Technical Papers. , 1995 IEEE/ACM International Conference on*. [S.l.: s.n.], 1995. p. 380–387.

LOCKE, C. D.; VOGEL, D. R.; MESLER, T. J. Building a predictable avionics platform in ada: a case study. In: *Real-Time Systems Symposium, 1991. Proceedings. , Twelfth*. [S.l.: s.n.], 1991. p. 181–189.

LORENTE, J. L.; LIPARI, G.; BINI, E. A hierarchical scheduling model for component-based real-time systems. In: *Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International*. [S.l.: s.n.], 2006. p. 8 pp. –.

LUCE, R. D.; PERRY, A. D. A method of matrix analysis of group structure. *Prychometrika*, out. 1949.

MARTELETO, R. M. Cultura da modernidade: discursos e práticas informacionais. *Revista da Escola de Biblioteconomia da UFMG*, v. 23, n. 2, p. 115–137, 1994.

MARTELETO, R. M. Análise de redes sociais - aplicação dos estudos de transferência da informação. *Ciência da Informação*, scielo, v. 30, p. 71 – 81, 04 2001. ISSN 0100-1965. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652001000100009&nrm=iso>.

MARTIN, G.; LAVAGNO, L.; LOUIS-GUERIN, J. Embedded uml: a merger of real-time uml and co-design. In: *Hardware/Software Codesign, 2001. CODES 2001. Proceedings of the Ninth International Symposium on*. [S.l.: s.n.], 2001. p. 23–28.

MASLOV, S.; REDNER, S. Promise and pitfalls of extending google's pagerank algorithm to citation networks. *The Journal of Neuroscience*, v. 28, n. 44, p. 11103–11105, 2008.

NEWMAN, M. E. J. The structure and function of complex networks. *SIAM review*, JSTOR, v. 45, n. 2, p. 167–256, 2003. ISSN 0036-1445.

NEWMAN, M. E. J. Coauthorship networks and patterns of scientific collaboration. In: *Proceedings of the National Academy of Sciences*. [S.l.: s.n.], 2004. p. 5200–5205.

NEWMAN, M. E. J. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, v. 103, n. 23, p. 8577–8582, jun. 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=retrieve&db=pubmed&list_uids=16723398&dopt=AbstractPlus>.

NEWMAN, M. E. J. Mathematics of networks. *The New Palgrave Encyclopedia of Economics*, 2007.

NEWMAN, M. E. J. *Networks: An Introduction*. New York: Oxford University Press, 2010.

NEWMAN, M. E. J.; GIRVAN, M. Finding and evaluating community structure in networks. 2004.

NGONMANG, B.; TCHUENTE, M.; VIENNET, E. Local community identification in social networks. *Parallel Processing Letters*, v. 22, n. 1, 2012. Disponível em: <<http://dblp.uni-trier.de/db/journals/ppl/ppl22.html#NgonmangTV12>>.

NOLKER, R. D.; ZHOU, L. Social computing and weighting to identify member roles in online communities. In: SKOWRON, A. et al. (Ed.). *Web Intelligence*. IEEE Computer Society, 2005. p. 87–93. ISBN 0-7695-2415-X. Disponível em: <<http://dblp.uni-trier.de/db/conf/webi/webi2005.html#NolkerZ05>>.

PANDA, P.; DUTT, N. D.; NICOLAU, A. Memory organization for improved data cache performance in embedded processors. In: *System Synthesis, 1996. Proceedings. , 9th International Symposium on*. [S.l.: s.n.], 1996. p. 90–95.

PANDA, P. R.; DUTT, N. D.; NICOLAU, A. Architectural exploration and optimization of local memory in embedded systems. In: *System Synthesis, 1997. Proceedings. , Tenth International Symposium on*. [S.l.: s.n.], 1997. p. 90–97.

PANDA, P. R.; DUTT, N. D.; NICOLAU, A. Efficient utilization of scratch-pad memory in embedded processor applications. In: *European Design and Test Conference, 1997. ED TC 97. Proceedings*. [S.l.: s.n.], 1997. p. 7–11.

PATSOPOULOS, N. A.; ANALATOS, A. A.; IOANNIDIS, J. A. Relative citation impact of various study designs in the health sciences. *JAMA*, v. 293, n. 19, p. 2362–2366, 2005. Disponível em: <<http://dx.doi.org/10.1001/jama.293.19.2362>>.

PIMENTEL, A. D.; ERBAS, C.; POLSTRA, S. A systematic approach to exploring embedded system architectures at multiple abstraction levels. *Computers, IEEE Transactions on*, v. 55, n. 2, p. 99–112, Feb 2006.

PROCÓPIO, P.; LAENDER, A. H.; MORO, M. M. Análise da rede de coautoria do simpósio brasileiro de bancos de dados. *Anais do XXVI Simpósio Brasileiro de Bancos de Dados*, p. 131–138, 2011.

RAMOS, M. V. M.; NETO, J. J.; VEGA, Í. S. *Linguagens Formais: Teoria, Modelagem e Implementação*. 1. ed. Porto Alegre: Bookman, 2009. 656p p. ISBN 978-85-7780-453-5.

REVOREDO, K. et al. Minerando publicações científicas para análise da colaboração em comunidades de pesquisa. *Brazilian Workshop on Social Networks Analysis and Mining*, 2012.

SANGIOVANNI-VINCENTELLI, A.; MARTIN, G. Platform-based design and software design methodology for embedded systems. *Design Test of Computers, IEEE*, v. 18, n. 6, p. 23–33, Nov 2001.

SCOTT, J. *Social Network Analysis: A Handbook*. Second. Sage Publications, 2000. ISBN 0761963391. Disponível em: <http://www.amazon.com/Social-Network-Analysis-Professor-Scott/dp/0761963383/ref=sr_1_1?ie=UTF8&s=books&qid=1256622319&sr=1-1>.

SCOTT, J.; CARRINGTON, P. J. *The SAGE Handbook of Social Network Analysis*. SAGE Publications, 2011. ISBN 9781446250112. Disponível em: <<https://books.google.com.br/books?id=mWlsKkIuFNgC>>.

SEMENTOV, A.; VEIJALAINEN, J.; BOUKHANOVSKY, A. A generic architecture for a social network monitoring and analysis system. In: BAROLLI, L.; XHAFI, F.; TAKIZAWA, M. (Ed.). *NBiS*. IEEE Computer Society, 2011. p. 178–185. ISBN 978-1-4577-0789-6. Disponível em: <<http://dblp.uni-trier.de/db/conf/nbis/nbis2011.html#SemenovVB11>>.

SGROI, M.; LAVAGNO, L.; SANGIOVANNI-VINCENTELLI, A. Formal models for embedded system design. *Design Test of Computers, IEEE*, v. 17, n. 2, p. 14–27, Apr 2000.

SHIN, Y.; CHOI, K.; SAKURAI, T. Power optimization of real-time embedded systems on variable speed processors. In: *Computer Aided Design, 2000. ICCAD-2000. IEEE/ACM International Conference on*. [S.l.: s.n.], 2000. p. 365–368.

STEINKE, S. et al. Assigning program and data objects to scratchpad for energy reduction. In: *Design, Automation and Test in Europe Conference and Exhibition, 2002. Proceedings*. [S.l.: s.n.], 2002. p. 409–415.

SÁ, C. C. D.; ROCHA, J. *Treze Viagens Pelo Mundo da Matemática*. U. PORTO, 2010. ISBN 9789898265340. Disponível em: <<https://books.google.com.br/books?id=25kri1SrOAUC>>.

TIWARI, V.; MALIK, S.; WOLFE, A. Power analysis of embedded software: a first step towards software power minimization. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, v. 2, n. 4, p. 437–445, Dec 1994.

WANG, Y.; LI, X. Social network analysis of interaction in online learning communities. In: SPECTOR, J. M. et al. (Ed.). *ICALT*. IEEE Computer Society, 2007. p. 699–700. ISBN 978-0-7695-2916-5. Disponível em: <<http://dblp.uni-trier.de/db/conf/icalt/icalt2007.html#WangL07>>.

WASSERMAN, S.; FAUST, K. *Social network analysis: Methods and applications*. Cambridge university press, 1994. v. 8. Disponível em: <http://scholar.google.com/scholar.bib?q=info:gET6m8icitMJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&as_vis=1&ct=citation&cd=0>.

WEALE, A.; BAILEY, M.; LEAR, P. The level of non-citation of articles within a journal as a measure of quality: a comparison to the impact factor. *BMC Medical Research Methodology*, v. 4, n. 1, p. 14, 2004. ISSN 1471-2288. Disponível em: <<http://www.biomedcentral.com/1471-2288/4/14>>.

WOLF, W. H. Hardware-software co-design of embedded systems [and prolog]. *Proceedings of the IEEE*, v. 82, n. 7, p. 967–989, Jul 1994.

WOLFE, A.; CHANIN, A. Executing compressed programs on an embedded risc architecture. In: *Microarchitecture, 1992. MICRO 25. , Proceedings of the 25th Annual International Symposium on*. [S.l.: s.n.], 1992. p. 81–91.

XUE, W.; SHI, J.; YANG, B. X-rime: Cloud-based large scale social network analysis. In: *IEEE SCC*. IEEE Computer Society, 2010. p. 506–513. ISBN 978-0-7695-4126-6. Disponível em: <<http://dblp.uni-trier.de/db/conf/IEEEscc/scc2010.html#XueSY10>>.

YEN, T.-Y.; WOLF, W. Communication synthesis for distributed embedded systems. In: *Computer-Aided Design, 1995. ICCAD-95. Digest of Technical Papers. , 1995 IEEE/ACM International Conference on*. [S.l.: s.n.], 1995. p. 288–294.

ZAVE, P. An operational approach to requirements specification for embedded systems. *Software Engineering, IEEE Transactions on*, SE-8, n. 3, p. 250–269, May 1982.

ZHAO, Q. et al. Characterizing and predicting community members from evolutionary and heterogeneous networks. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. [S.l.]: ACM, 2008. p. 309–318.

ZHU, Y.; MUELLER, F. Feedback edf scheduling exploiting dynamic voltage scaling. In: *Real-Time and Embedded Technology and Applications Symposium, 2004. Proceedings. RTAS 2004. 10th IEEE*. [S.l.: s.n.], 2004. p. 84–93.