# UM ESTUDO SOBRE A EVOLUÇÃO TEMPORAL DE COMUNIDADES CIENTÍFICAS

BRUNO LEITE ALVES

# UM ESTUDO SOBRE A EVOLUÇÃO TEMPORAL DE COMUNIDADES CIENTÍFICAS

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

Orientador: Alberto Henrique Frade Laender
Coorientador: Fabrício Benevenuto de Souza

Belo Horizonte
Agosto de 2013

Alves, Bruno Leite

A474e Um Estudo sobre a Evolução Temporal de Comunidades Científicas / Bruno Leite Alves. — Belo Horizonte, 2013
xxii, 77 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais - Departamento de Ciência da Computação
Orientador: Alberto Henrique Frade Laender
Coorientador: Fabrício Benevenuto de Souza

1. Computação - Teses. 2. Redes de Relações Sociais - Teses. 3. Instituições e Sociedades Científicas – Teses. 4. Teoria dos Grafos - Teses. 5. Redes Complexas - Teses. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*62(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

# FOLHA DE APROVAÇÃO

Um Estudo Sobre a Evolução Temporal de Comunidades Científicas

**BRUNO LEITE ALVES**

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ALBERTO HENRIQUE FRADE LAENDER - Orientador
Departamento de Ciência da Computação - UFMG

PROF. FABRÍCIO BENEVENUTO DE SOUZA - Coorientador
Departamento de Ciência da Computação - UFMG

PROFA. ANA PAULA COUTO DA SILVA
Departamento de Ciência da Computação - UFMG

PROFª. JONICE DE OLIVEIRA SAMPAIO
Departamento de Ciência da Computação - UFRJ

Belo Horizonte, 23 de agosto de 2013.

*Para Maria Célia, Juliano, Thiago e Ana Paula.*

# Agradecimentos

Meus mais sinceros agradecimentos para todos que me encorajaram e contribuíram de alguma forma para o desenvolvimento deste trabalho. Em especial, gostaria de agradecer:

- A Deus, pois sem Ele nada disso seria possível;
- Ao meu orientador Alberto H. F. Laender e coorientador Fabrício Benevenuto pelos ensinamentos, disposição, amizade, e principalmente por acreditarem que eu seria capaz;
- Ao professor Anderson A. Ferreira e a professora Mirella M. Moro por me ajudarem no início desta caminhada;
- Aos meus colegas do LBD. Em particular, Daniel Hasan, Diego Oliveira, Harlley Lima, Michele Brandão, Rodrygo Santos, Sérgio Canuto e Thiago Silva;
- A minha mãe Maria Célia por todo apoio, admiração e todo o amor incondicional que me fizeram sempre seguir em frente;
- Ao meu pai João Bosco, que mesmo com suas as dificuldades sempre torceu por mim;
- A toda minha família, em especial, meus irmãos Juliano e Thiago, minha querida avó e madrinha Geni Leite (*in memoriam*) e meu tio e padrinho José Flávio Leite (*in memoriam*);
- A minha namorada Ana Paula por sua companhia, afeição, respeito, amizade e todo o suporte;
- Aos meus amigos, Felipe Cáfaro, Olavo Shibata, Bruno Vaz, Adriano Tavares, Leonardo Kenji, Denise Notini e todos aqueles que torceram sempre por mim;
- A CAPES, CNPq, Fapemig e InWeb por financiar parcialmente este trabalho.

*"De tudo, ficaram três coisas: a certeza de que ele estava sempre começando, a certeza de que era preciso continuar e a certeza de que seria interrompido antes de terminar. Fazer da interrupção um caminho novo. Fazer da queda um passo de dança, do medo uma escada, do sono uma ponte, da procura um encontro."*

(Fernando Sabino)

# Resumo

Diversos esforços têm sido realizados para compreensão e modelagem de aspectos dinâmicos de comunidades científicas. Apesar do grande interesse, pouco se sabe sobre o papel que diferentes membros têm na formação da estrutura topológica da rede dessas comunidades. Nesta dissertação, investigamos o papel que os membros do núcleo de comunidades científicas têm na formação e evolução de sua rede de colaboração. Para isso, definimos o núcleo de uma comunidade com base em uma métrica, denominada *CoScore*, derivada do índice h que captura tanto a prolificidade quanto o envolvimento dos pesquisadores na comunidade. Nossos resultados subsidiam uma série de observações importantes relacionadas à formação e aos padrões de evolução das comunidades. Particularmente, mostramos que os membros do núcleo das comunidades atuam como pontes que conectam pequenos grupos de pesquisa. Além disso, esses membros são responsáveis pelo aumento do grau médio de toda a rede que representa a comunidade e pela redução de sua assortatividade. Mais importante, notamos que variações no conjunto de membros que compõem o núcleo das comunidades tendem a ser fortemente correlacionadas com variações dessas métricas. Mostramos ainda que nossas observações são importantes para caracterizar o papel dos principais membros na formação e na estrutura das comunidades.

# Abstract

Several efforts have been done to understand and model dynamic aspects of scientific communities. Despite the great interest, little is known about the role that different members play in the formation of the underlying network structure of such communities. In this dissertation, we investigate the roles that members of the core of scientific communities play in the formation and evolution of the collaboration network structure. To do that, we define a community core based on a metric, named *CoScore*, which is an h-index derived metric that captures both, the prolificness and the involvement of researchers in a community. Our results provide a number of key observations related to community formation and evolving patterns. Particularly, we show that members of the community core work as bridges that connect smaller clustered research groups. Furthermore, these members are responsible for an increase in the average degree of the whole community underlying the network and a decrease on the overall network assortativeness. More important, we note that variations on the set of members that form the community core tend to be strongly correlated with variations on these metrics. We also show that our observations are important to characterize the role of key members on the formation and structure of these communities.

# Lista de Figuras

# Lista de Tabelas

# Sumário

# Capítulo 1

# Introdução

## 1.1 Motivação

Desde os seus primórdios, a sociedade tem se organizado em comunidades, ou seja, grupos de indivíduos com interesses em comum[1]. Particularmente, a proliferação de novas tecnologias de comunicação baseadas na Internet tem facilitado a rápida formação e crescimento de comunidades *online* [Kleinberg, 2008]. Comunidades possuem uma grande quantidade de características e servem a vários propósitos. Elas podem ser desde pequenos grupos hermeticamente comprometidos com temas específicos, como comunidades científicas de determinadas áreas, até mesmo grupos de milhões de usuários ligados por um interesse em comum, tais como comunidades relacionadas a esporte ou fãs de uma celebridade.

Geralmente, indivíduos que são socialmente conectados em comunidades tendem a compartilhar interesses comuns e outras similaridades. Embora existam muitos fatores que possam determinar a formação de uma comunidade e o seu crescimento, existem duas forças que explicam a formação de comunidades: influência e homofilia. Por um lado, influência postula que indivíduos mudam para se tornarem mais similares a seus amigos na comunidade. Por outro lado, homofilia postula que indivíduos criam conexões sociais dentro de uma comunidade justamente porque já são similares. Esforços recentes têm mostrado evidências quantitativas que ambas as forças [Cha et al., 2010; Backstrom et al., 2006], teorias [Rogers, 1962; Watts & Dodds, 2007] e modelos existentes [Kempe et al., 2003, 2005] dependem da identificação de um grupo influente de indivíduos com o poder de afetar não somente a estrutura topológica de uma comunidade, mas também interferir na difusão e no fluxo de informação da comunidade.

[1] http://www.merriam-webster.com/dictionary/community

Nesta dissertação, apresentamos uma perspectiva diferente e um estudo complementar desse problema. Aqui, nos concentramos em estudar os papéis que indivíduos influentes em comunidades científicas desempenham na evolução das propriedades de tais comunidades. Intuitivamente, quando pesquisadores importantes e com grande influência em suas áreas decidem se juntar ou deixar uma comunidade científica, levam com eles recursos, experiência e até mesmo estudantes, e possivelmente influenciam outros membros a fazerem o mesmo. Para esse estudo, usamos dados da DBLP[2] para construir comunidades científicas, representadas pelas principais conferências dos SIGs[3] (*Special Interest Groups*) da ACM[4] (*Association for Computing Machinery*). Então, propomos uma estratégia para definir o núcleo de uma dada comunidade científica, juntamente com seus líderes em um dado período de tempo. Finalmente, investigamos como os aspectos do núcleo impactam a estrutura topológica da comunidade.

O estudo do núcleo de comunidades científicas pode ser visto de duas perspectivas diferentes. A primeira é a sociológica, vindo da necessidade de compreender como partes da sociedade evoluem, bem como responder a perguntas de longa data relacionadas com à interação entre os diferentes tipos de participante. Em contrapartida, sob uma perspectiva tecnológica, compreender estes aspectos é crítico para muitas aplicações relacionadas a predição de *links* [Getoor & Diehl, 2005]. Tal estudo, entretanto, tem sido difícil devido à caracterização de alguns conceitos, como conexões humanas e uma definição apropriada de liderança, sendo difícil até mesmo de se reproduzir em grande escala dentro de um laboratório de pesquisa.

## 1.2 Trabalhos Relacionados

Recentemente, esforços tem sido feitos com o intuito de analisar a estrutura das comunidades e a evolução de suas redes. Particularmente, Kumar et al. [2006] analisaram duas grandes redes através do tempo, Flickr[5] e Yahoo! 360, para encontrar uma segmentação dessas redes em indivíduos isolados, componentes intermediários, que também são comunidades isoladas, e o maior componente conectado. Nesse trabalho, os autores identificaram que alguns componentes possuem um centro constituído de nodos de maior grau com muitos nodos de menor grau conectados a eles e que foram caracterizados estrelas, isto é, componentes que são formados rapidamente, mas não são absorvidos pelo maior componente conectado. Assim, eles propuseram um modelo de

[2]http://dblp.uni-trier.de/
[3]http://www.acm.org/sigs
[4]http://www.acm.org
[5]http://www.flickr.com/

crescimento capaz de gerar redes com características similares às das redes estudadas. Este modelo se baseia em três tipos de nodo: passivos, recrutadores e agregadores. Usuários passivos se juntam à rede pela curiosidade ou pela insistência de amigos, mas nunca se comprometem efetivamente em atividades da rede. Recrutadores estão interessados em transformar comunidades *offline* em comunidades *online* e recrutar seus amigos a participarem da rede. Por fim, os agregadores são participantes que contribuem ativamente para o crescimento da rede e frequentemente se conectam a outros membros semelhantes.

Ducheneaut et al. [2007] extraíram e caracterizaram explicitamente, comunidades criadas a partir de cinco servidores do *World of Warcraft*, um jogo multijogador massivo. O estudo é centralizado em grupos de jogadores formados para realizar atividades em conjunto, denominados guildas. Tais grupos apresentam dificuldades consideráveis de administração e muitos não sobrevivem ao longo do tempo, sendo os líderes membros importantes para a longevidades das guildas. Os autores apresentaram algumas propriedades demográficas sobre esses grupos, tais como, distribuição do tamanho das guildas, interação de membros e não membros de guildas em eventos do jogo, e tamanho das guildas ao longo do tempo. Em seguida, mostraram o impacto das guildas na estrutura da rede utilizando métricas clássicas de redes complexas como centralidade e densidade. Eles sugeriram ainda um painel para monitorar o tamanho, número de subgrupos e densidade das guildas, além de uma visualização da topologia da rede.

Complementarmente ao estudo de Ducheneaut et al. [2007], Patil et al. [2012] analisaram e modelaram fatos que levam usuários a deixar ou entrar em guildas do jogo *World of Warcraft*. Nesse trabalho os autores propõem um modelo capaz de predizer se e quando um jogador irá deixar a comunidade e também qual o impacto dessa saída. Este modelo é baseado em várias características, tais como, níveis dos jogadores e das guildas, atividades do jogo e características sociais. Estas características sociais estão ligadas a um jogador podendo ser o seu número de amigos, o número de amigos que já deixaram a comunidade, e a percentagem de membros na comunidade que interagem com o jogador em questão, bem como a frequência. O modelo construído se mostrou efetivo em predizer quando os usuários deixarão suas comunidades e quando essas comunidades irão se desfazer na rede.

Viswanath et al. [2009] estudaram dois tipos de rede formados a partir de dados do Facebook[6]. A primeira rede foi construída utilizando os laços de amizade entre os usuários e a segunda utilizando as interações entre os usuários, e.g., postagens em mural e seus comentários. Eles constataram que muitos usuários aceitam pedidos de

[6]http://www.facebook.com

amizade por cortesia, de modo que a rede baseada em laços de amizades apresenta um crescimento surreal. A rede construída a partir das interações entre os usuários possui comportamento diferente. Nessa rede, os laços tendem a surgir e desaparecer rapidamente através do tempo, e a força dos laços apresenta, de forma geral, uma tendência decrescente das atividades que acompanham a idade dos *links* nas redes sociais. O trabalho também apresenta a métrica *resemblance*, utilizada para medir a sobreposição entre duas instâncias da rede, para mostrar que a rede possui um núcleo que persiste através do tempo. Finalmente, os autores apresentam uma análise utilizando métricas clássicas de redes complexas, mostrando que as propriedades estruturais das redes estudadas não apresentam grandes variações ao longo do tempo.

Em termos de modelos para redes sociais dinâmicas, Leskovec et al. [2005] investigaram nove redes, que incluem redes de citações de diferentes áreas da física, citações de patentes americanas e afiliações de pesquisadores, para mostrar que essas redes densificam ao longo do tempo, com o número de arestas crescendo de forma superlinear em relação ao número de nodos e que o caminho mínimo médio entre os nodos geralmente diminui através do tempo. Baseado nessas observações, eles desenvolveram um modelo de geração de redes que incorpora tais propriedades, chamado de *Forest Fire*. Esse modelo necessita apenas de dois parâmetros e é capaz de capturar padrões como a lei de potência da densificação e redução do diâmetro. Mais recentemente, Leskovec et al. [2008] apresentaram um estudo detalhado da evolução de redes por meio da análise de quatro grandes redes sociais *online*, são elas: Flickr, Delicious[7], Answers[8] e LinkedIn[9]. Eles investigaram uma grande variedade de estratégias de formação de redes para mostrar que a localização das arestas desempenha um papel crítico na evolução dessas redes. Baseados nessas observações, os autores desenvolveram um modelo de evolução em que novos nodos surgem na rede com uma taxa predeterminada. O modelo também considera que novas arestas possuem distâncias muito curtas, normalmente formando triângulos, sendo assim, o modelo segue este padrão para formação de novas arestas. Diferentemente dos esforços acima referidos, nosso trabalho foca em propriedades das comunidades e no papel que os líderes dessas comunidades desempenham na topologia da rede.

Alguns outros trabalhos também abordaram o estudo de comunidades científicas. Particularmente, Backstrom et al. [2006] estudaram comunidades na LiveJournal[10] e comunidades científicas extraídas da DBLP. Esses dois conjuntos de dados possuem

[7]http://delicious.com/
[8]http://www.answers.com/
[9]http://www.linkedin.com
[10]http://www.livejournal.com/

comunidades explícitas, onde conferências representam comunidades na DBLP. Os autores identificaram que as tendências que levam indivíduos a entrar em comunidades e comunidades a crescer rapidamente dependem sutilmente da topologia da rede. Por exemplo, a tendência de um indivíduo a entrar em uma comunidade é influenciada não só pelo número de amigos que ele possui dentro daquela comunidade, mas também de como esses amigos estão conectados entre si. Eles utilizaram técnicas de árvore de decisão para identificar as estruturas mais significativas que afetam estas comunidades, e desenvolveram uma nova metodologia capaz de mensurar a mudança de indivíduos entre comunidades e também verificar como isto está relacionado com mudanças de interesse dentro das comunidades.

Huang et al. [2008] usaram os dados da biblioteca digital CiteSeer[11] para construir uma rede de coautoria em Ciência da Computação abrangendo pesquisas realizadas entre 1980 e 2005. Eles realizaram estudos sobre tendências de evolução das colaborações sob duas perspectivas, rede completa e comunidades. Entre as suas principais observações, eles mostraram que a área de Ciência da Computação apresenta padrões de colaboração mais similares à Matemática do que à Biologia. Além disso, eles também quantificaram e compararam padrões de colaboração de seis comunidades dentro da área da Ciência da Computação: Inteligência Artificial, Aplicações, Arquitetura, Bancos de Dados, Sistemas e Teoria. Com base nessas informações, os autores propuseram um modelo de aprendizagem e predição de colaborações entre pares de autores. Diferente desses esforços, nesta dissertação focamos no estudo das propriedades do núcleo da comunidade, de modo que nossas análises são complementares a essas.

Quando se trata de identificar núcleos de comunidades, existem várias abordagens que extraem o núcleo tendo como base as propriedades topológicas da rede. Leskovec et al. [2010] compararam vários algoritmos de detecção de comunidades em vários tipos de rede. Neste trabalho, uma comunidade é definida como nodos que possuem mais ligações entre si do que com o restante da rede. Os autores apontam que, de forma geral, os algoritmos são otimizados e detectam as comunidades efetivamente. No entanto, existem classes de redes em que os algoritmos executam de forma subótima. Já Chakrabarti et al. [2006] propuseram um arcabouço baseado em agrupamento hierárquico aglomerativo e *k-means* para realizar o agrupamento de dados através do tempo. Eles consideraram dois critérios para avaliação: (i) o agrupamento ao longo do tempo não deve apresentar grandes diferenças em relação aos dados globais e (ii) o agrupamento não deve mudar drasticamente de uma iteração para outra. Eles construíram uma rede a partir de dados do Flickr e constaram que o arcabouço proposto atende

[11] http://citeseer.ist.psu.edu

aos dois critérios definidos. Hopcroft et al. [2004] também propuseram um arcabouço baseado em agrupamento hierárquico capaz de identificar comunidades na biblioteca digital CiteSeer. Além dessas comunidades, eles também apontaram um grupo de artigos que aparece várias vezes nos grupos identificados ao longo do tempo, dos quais eles denominaram como núcleo da rede. Tais abordagens não são aplicáveis em nosso contexto, já que estamos interessados no estudo das propriedades dos núcleos.

## 1.3 Contribuições

A seguir listamos as principais contribuições desta dissertação:

- Definição de uma métrica, chamada *CoScore*, que captura tanto a prolificidade quanto o envolvimento de um pesquisador em uma comunidade científica. Desta forma, nossa métrica é capaz de quantificar a importância de um determinado pesquisador em uma dada comunidade científica [Alves et al., 2013].
- Definição do conceito de núcleo de uma comunidade a partir da métrica proposta [Alves et al., 2013].
- Caracterização de mais de vinte comunidades científicas e uma discussão de como a métrica *CoScore* afeta as propriedades topológicas das redes ao longo do tempo [Alves et al., 2013].
- Visualização das comunidades estudadas, em que é possível observar os componentes da rede e como os membros dos núcleos se organizam nestes componentes.

## 1.4 Organização da Dissertação

O restante desta dissertação está organizado da seguinte forma. No Capítulo 2 apresentamos uma visão geral sobre redes complexas e suas características, bem como as principais definições e métricas utilizadas ao longo da dissertação. No Capítulo 3 descrevemos nossa abordagem e o conjunto de dados utilizado para construir as comunidades científicas, bem como nossa estratégia para computar o núcleo das comunidades. No Capítulo 4 investigamos os papéis que esses conjuntos de pesquisadores exercem dentro de suas comunidades. Finalmente, no Capítulo 5 apresentamos nossas conclusões e algumas direções para trabalhos futuros.

# Capítulo 2

# Redes Complexas

Neste capítulo apresentamos uma visão geral dos conceitos, perspectivas e aplicações de redes complexas, bem como as principais definições e métricas utilizadas no presente trabalho.

## 2.1 Introdução

O interesse pela forma e complexidade como entidades[1] estão conectadas na sociedade moderna tem ganhado força na última década. Em meio a este interesse existe a ideia das redes que Easley & Kleinberg [2010] definem como um padrão de interconexões entre um conjunto de entidades. Essas redes podem aparecer em vários contextos e podem ser vistas sob várias perspectivas.

As redes sociais, de acordo com Easley & Kleinberg [2010], são compostas por coleções de laços sociais entre amigos e apresentam grande crescimento de complexidade ao logo da história humana devido aos avanços tecnológicos que facilitaram o deslocamento geográfico, a comunicação global e a interação social.

Dunbar [1992] mostra que indivíduos tendem a se organizar em grupos e que tais indivíduos possuem uma capacidade cognitiva de manter relações sociais estáveis com um número limitado de indivíduos (uma variação entre 100 e 230). Mais tarde, Gonçalves et al. [2011] mostraram que essas características se mantêm em *redes sociais online* (OSNs - *Online Social Networks*), mostrando que a complexidade entre as ligações dos indivíduos tendem a manter-se mesmo com a evolução tecnológica.

Podemos identificar sistemas na sociedade atual que possuem estruturas semelhantes às de uma rede, i.e., possuem elementos interconectados. Essas estruturas se

[1] Por entidade entende-se qualquer coisa, concreta ou abstrata, que tenha importância em um dado domínio.

tornaram tão complexas de modo que uma pequena alteração em um dado ponto pode refletir por todo o sistema de forma catastrófica, como sistemas tecnológicos e econômicos, onde pequenas alterações podem causar falhas em cascatas ou crises financeiras.

Redes podem ser identificadas em vários outros contextos, como redes de fornecedores em operações de manufatura, *sites* com redes de usuários ou redes de *sites* que se interligam através de hiperlinks e empresas de mídia com suas redes de anunciantes. Em tais formulações, muitas vezes o foco acaba sendo mais na estrutura da própria rede do que na sua complexidade como um todo, onde alterações em elementos centrais podem causar reações inesperadas.

O estudo sobre redes e suas conexões, conhecido também como análise de redes complexas, é um tema interdisciplinar que permeia diversas áreas do conhecimento como Ciência da Computação, Matemática, Física, Biologia e Sociologia. De acordo com Boccaletti et al. [2006], uma rede complexa pode ser formalmente representada como um grafo, uma ferramenta matemática originada dos estudos de Euler, que inicialmente a utilizou para formalizar o problema das pontes de Königsberg, conhecido também como o problema das sete pontes [Euler, 1956].

A análise de redes complexas e a teoria dos grafos são utilizadas em diversas outras áreas, dentre elas a sociometria, que utiliza-se desses conceitos para realizar a modelagem dos atores e suas respectivas ligações, sendo os atores representados por nodos e suas ligações representadas por arestas. Com a rede social gerada, ou sociograma como é chamada na sociometria, é possível identificar: (i) os líderes aceitos, (ii) atores que por algum motivo estão marginalizados, (iii) grupos que se fecharam com algum interesse em comum, (iv) atores responsáveis por unir um ou mais grupos sem serem membros de tais grupos, também chamados de estrelas, (v) atores pontes, que são responsáveis por unirem um ou mais grupos dos quais fazem parte, e por fim, (vi) atores isolados, que não fazem parte de qualquer grupo [Moreno, 1978].

Com o surgimento e popularização das OSNs, tais estudos se intensificaram devido ao grande número de pessoas nessas redes e também pelo conteúdo gerado, podendo estes conteúdos serem desde simples mensagens de texto ou até mesmo conteúdo multimídia, como fotos e vídeos. De acordo com Benevenuto et al. [2012], OSNs são ambientes ricos para o estudo de várias áreas da Ciência da Computação, incluindo sistemas distribuídos, padrões de tráfego na Internet, mineração de dados, sistemas multimídia e interação humano-computador.

## 2.2 Definições e Características

Newman [2003] define uma rede como é um conjunto de itens conectados entre si. Em uma rede, um elemento de um dado sistema possui ligações com outros elementos do mesmo sistema, i.e., dado um grupo de elementos de um sistema qualquer, devemos determinar alguma regra que interligue tais elementos. A natureza do problema a ser modelado pode variar entre diversas áreas, podendo estes serem pessoas, neurônios, computadores, dentre outros. A semântica desta ligação varia de acordo com o problema, e.g., pessoas podem estar ligadas por um laço de amizade ou profissional, neurônios podem estar ligados através de sinapses e computadores podem estar interligados através de uma rede *ad hoc*.

Uma rede pode ser modelada como um grafo que possui um conjunto de nodos $N$ e um conjunto de arestas $E$ que conectam esses nodos, sendo o número de nodos da rede definido como $n = |N|$. Uma aresta, direcionada ou não, existe entre dois nodos, de acordo com o problema modelado. Em um grafo com arestas direcionadas (digrafo), cada aresta que conecta um nodo origem a um nodo destino possui uma direção. Em geral, as arestas das OSNs não possuem direção, pois o laço de amizade é bilateral. No Facebook, por exemplo, um usuário precisa enviar uma solicitação de amizade para uma determinada pessoa, sendo este pedido sujeito a aprovação. Após a aprovação, ambos se tornam amigos na rede sem qualquer distinção entre o laço de amizade. No entanto, existem redes onde o relacionamento possui direção, e.g., na rede de *microblogging* Twitter[2], os usuários seguem outros usuários, sendo que a recíproca não precisa necessariamente existir, i.e., uma entidade famosa no mundo real não precisa seguir seus fãs, mas provavelmente todos os seus fãs a seguirão.

OSNs possuem características que nos permitem modelá-las como redes, pois possuem elementos que se relacionam, consequentemente podemos mapear esta modelagem para um grafo. Na Tabela 2.1 listamos várias OSNs populares atualmente, conforme apresentado por Benevenuto et al. [2012].

Além das OSNs, temos também as redes de colaboração científica, que são formadas por pesquisadores que publicam trabalhos em fóruns científicos. Essas redes também podem ser modeladas como um grafo, onde cada nodo corresponde a um pesquisador na rede e uma aresta entre dois nodos indica que os pesquisadores publicaram pelo menos um trabalho em conjunto. Os dados mantidos pelas bibliotecas digitais DBLP e BDBComp[3], por exemplo, possibilitam este tipo de modelagem.

[2]http://www.twitter.com/
[3]http://www.lbd.dcc.ufmg.br/bdbcomp/

Tabela 2.1: *Redes sociais online* populares

| **Nome** | **Propósito** | **URL** |
|---|---|---|
| Google+ | Amizades | http://plus.google.com |
| Facebook | Amizades | http://www.facebook.com |
| MySpace | Amizades | http://www.myspace.com |
| Hi5 | Amizades | http://www.hi5.com |
| LinkedIn | Profissionais | http://www.linkedin.com |
| YouTube | Compartilhamento de vídeos | http://www.youtube.com |
| Flickr | Compartilhamento de fotos | http://www.flickr.com |
| LiveJournal | Blogs e diários | http://www.livejournal.com |
| Digg | Compartilhamento de *bookmarks* | http://digg.com |
| Twitter | Troca de mensagens curtas | http://twitter.com |
| LastFM | Compartilhamento de músicas | http://www.last.fm |

## 2.3 Métricas

Nesta seção apresentamos métricas que tipicamente são utilizadas em análises de redes complexas. As métricas são baseadas na topologia da rede e podem ser usadas para identificar suas características.

### 2.3.1 Grau dos Nodos

O grau de um nodo é o número de arestas incidentes àquele nodo, com os laços (uma aresta que possui o mesmo nodo como origem e destino) contando duas vezes. Esta é uma característica importante na estrutura da rede e segue a lei de potência em vários tipos de rede, como a Internet [Faloutsos et al., 1999], a Web [Barabási & Albert, 1999] e redes neurais [Braitenberg & Schüz, 1998]. Sendo assim, a probabilidade de um nodo ter grau $k$ é proporcional a $k^{-\alpha}$, sendo $\alpha$ uma constante obtida através de regressão linear. Este expoente é comumente utilizado para comparar diferentes redes e seus valores variam entre 1,0 e 3,5 [Ebel et al., 2002]. Em grafos direcionados, é comum analisar o grau dos nodos levando em consideração as arestas de entrada e de saída. Por exemplo, em uma rede de coautoria, um pesquisador é representado por um nodo, cujo grau indica o seu número de coautores.

### 2.3.2 Coeficiente de Agrupamento

O coeficiente de agrupamento (*clustering coefficient*) é um indicador de conectividade entre os nodos de um grafo. Este indicador informa o quão agrupados os vizinhos de um dado nodo se encontram na rede, i.e., o coeficiente de agrupamento indica o quão

interligados estão os coautores de um dado pesquisador na rede. Isto acontece porque os nodos tendem a criar grupos coesos caracterizados por um denso número de ligações. Sendo assim, a probabilidade de nodos se interconectarem na forma de grupos tende a ser maior do que ligações aleatórias na rede.

Figueiredo [2011] define o coeficiente de agrupamento de um nodo $i$, que pertence ao conjunto de nodos $i \in N$ de um dado grafo, como sendo a fração de arestas que os vizinhos de $i$ possuem entre si e o máximo de arestas possíveis que poderiam existir entre eles. Sendo $d_i$ o grau do nodo $i$, o número máximo de arestas entre seus vizinhos é $\binom{d_i}{2}$, ou seja, quando todos os pares de vizinhos de $i$ estão interconectados. Assim, sendo $E_i$ o número de arestas entre os vizinhos de $i$, temos que o coeficiente de agrupamento $c_i$ do nodo $i$ é dado por:

$$c_i = \frac{E_i}{\binom{d_i}{2}} = \frac{2E_i}{d_i(d_i - 1)}. \tag{2.1}$$

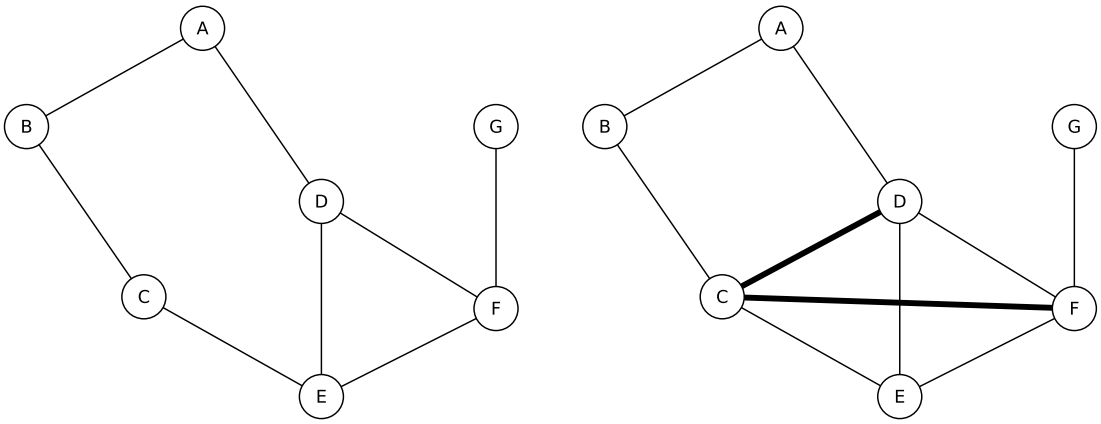
Note que o coeficiente de agrupamento definido na Equação 2.1 é aplicável somente a nodos com grau maior que um. De acordo com Figueiredo [2011], utilizando o coeficiente de agrupamento de cada nodo, podemos obter o coeficiente de agrupamento de toda a rede calculando a respectiva média aritmética, conforme:

$$\bar{c} = \frac{1}{n} \sum_{i \in N} c_i. \tag{2.2}$$

É importante ressaltar que a Equação 2.2 necessita que todos os nodos pertencentes a $N$ tenham seus coeficientes de agrupamento previamente calculados. Considerando a Figura 2.1a, onde cada nodo representa um pesquisador em uma rede de coautoria, o pesquisador $E$ possui um coeficiente de agrupamento 1/3, porque existe somente uma única aresta entre *D-F*, sendo que temos três pares de pesquisadores *D-F*, *D-C*, e *C-F*. Na Figura 2.1b podemos observar que o valor do coeficiente de agrupamento do pesquisador $E$ é incrementado para 1, porque existem três arestas entre *D-F*, *D-C*, e *C-F* entre os mesmos três pares.

### 2.3.3 Componentes

Segundo Easley & Kleinberg [2010], um grafo é conectado quando existem arestas interligando todos os seus nodos. Naturalmente, um grafo é desconectado quando nem todos os seus nodos estão interligados, de modo que eles se dividem em grupos de nodos interconectados entre si, porém desconectados do restante da rede, sendo que dois grupos não se sobrepõem. Na Figura 2.2, podemos considerar que o grafo consiste




(a) Antes de novas arestas se formarem (b) Depois de novas arestas se formarem

Figura 2.1: Exemplo de coeficiente de agrupamento

de três partes, ou seja, três componentes conectados: um composto pelos nodos $K$, $I$, $J$ e $H$, outro composto pelos nodos $L$ e $M$ e o terceiro composto pelos demais nodos. Por exemplo, em uma rede de coautoria, poderíamos dizer que cada componente é um grupo de pesquisa, sendo que os pesquisadores de um dado grupo não trabalham com pesquisadores de outro grupo.

Easley & Kleinberg [2010] definem um componente conectado (*connected component*), formalmente, como um subconjunto de nodos, tal que: (i) exista um caminho interligando todos os nodos daquele subconjunto e (ii) o subconjunto não é parte de um conjunto maior, onde cada nodo possa chegar a qualquer outro. Assim, podemos intuitivamente definir que um componente é: (i) internamente conectado e (ii) como um todo, é uma "parte" desconectada das demais partes do grafo. Por exemplo, o subconjunto $C$, $F$ e $E$ da Figura 2.2 não poderia ser chamado de componente conectado, pois violaria a condição (ii). Embora o subconjunto atenda a condição (i), ele faz parte de um subconjunto maior que engloba o nodos de $A$ a $G$.

Em um grafo direcionado, um componente é chamado de fortemente conectado quando existe pelo menos um caminho direcionado interligando todos os pares de nodos. Quando tal caminho existe, porém não direcionado, o componente é chamado de fracamente conectado. Benevenuto et al. [2012] exemplificam o modelo *bow tie*, definido por Broder et al. [2000], em que um grafo possui um componente central fortemente conectado, também chamado de *core*, que pode ser alcançado ou alcançar outros grupos de componentes.

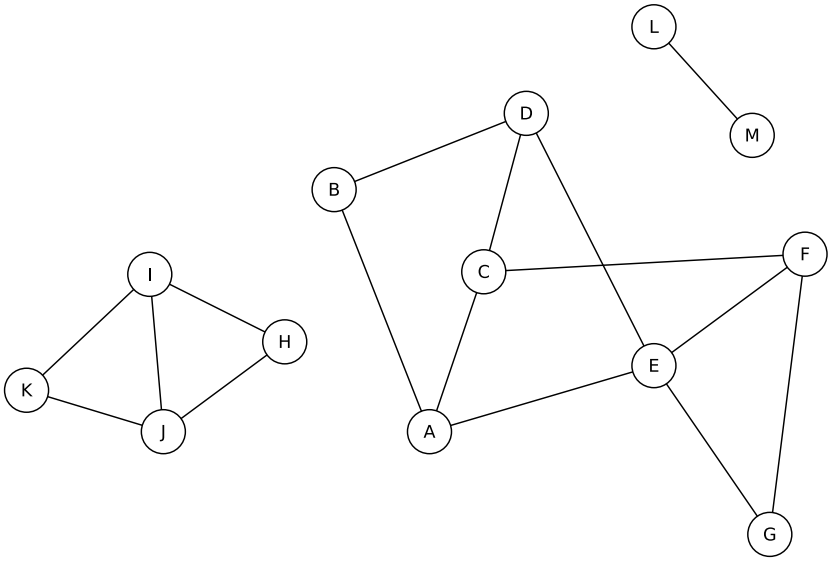



Figura 2.2: Um grafo com três componentes conectados

### 2.3.4 Caminho Mínimo Médio e Diâmetro

Os caminhos de uma rede são um aspecto importante dela. Um caminho pode ser definido como uma sequência de nodos sem repetição onde existe uma aresta entre cada par de nodos adjacentes na sequência. Por exemplo, na Figura 2.2 podemos dizer que existe um caminho entre $A$ e $F$, sendo este caminho formado pelas arestas $A$-$E$ e $E$-$F$. O comprimento de um caminho é dado pelo número de arestas que o define ou pelo número de nodos contidos no caminho menos um. Ainda na Figura 2.2, o caminho contendo os nodos $A$, $B$, $D$, $C$ e $F$ possui tamanho quatro.

Naturalmente existem muitos caminhos entre dois nodos quaisquer de um componente. Desta forma, o caminho mínimo entre dois nodos é definido como sendo o comprimento do menor caminho entre eles. Assim, considerando os nodos $A$ e $F$ na Figura 2.2, o caminho mínimo entre eles é dois, definida pelos caminhos $A$, $C$ e $F$ ou $A$, $E$ e $F$.

O caminho mínimo médio de um grafo é a média do número de arestas em todos os caminhos mínimos existentes entre todos os pares de nodos do grafo. Geralmente esta medida é calculada no maior componente fortemente conectado para grafos direcionados ou no maior componente fracamente conectado para grafos não direcionados, uma vez que o grafo pode não ser totalmente conectado. Figueiredo [2011] define o caminho mínimo médio como a média aritmética dos caminhos mínimos entre todos os pares de nodos da rede, sendo $l(i, j)$ o caminho mínimo entre os nodos $i, j \in N$, onde

$N$ é o conjuntos de nodos da rede. O caminho mínimo médio $\bar{l}$ é definido como:

$$\bar{l} = \frac{\sum_{i,j \in N} l(i,j)}{\binom{n}{2}}. \tag{2.3}$$

A Equação 2.3 considera todos os pares não-ordenados, que ao todo são $\binom{n}{2}$. Outra métrica baseada em caminho mínimo é o diâmetro que também é calculado no maior componente fortemente conectado ou fracamente conectado. O diâmetro é o tamanho do maior caminho mínimo existente em todo o grafo, que Figueiredo [2011] define como:

$$L = \max_{i,j \in N} l(i,j). \tag{2.4}$$

### 2.3.5 *Betweenness*

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

O valor da métrica *betweenness* $B(e)$ de uma aresta $e$ pode, de acordo com Benevenuto et al. [2012], ser formalmente definido como o número de caminhos mínimos entre todos os pares de nodos que passam por $e$. Desta forma temos:

$$B(e) = \sum_{u \in N, v \in N} \frac{\sigma_e(u,v)}{\sigma(u,v)} \tag{2.5}$$

onde $\sigma(u,v)$ representa o número de caminhos mínimos entre $u$ e $v$, e $\sigma_e(u,v)$ representa o número de caminhos mínimos que incluem $e$. Assim, se existem vários caminhos mínimos entre $u$ e $v$, cada caminho recebe um peso de modo que o somatório dos pesos seja um.

De forma análoga, o valor da métrica *betweenness* pode ser computado para um nodo da rede ao invés de uma aresta. Assim teríamos essa métrica representando a importância de um dado nodo para a rede, onde os vários caminhos mínimos que passam por ele representam, de forma quantitativa, sua importância na rede. Por exemplo, em uma rede de coautoria, a existência de nodos com um alto valor para a métrica *betweenness* pode indicar que os respectivos pesquisadores atuam como pontes interligando vários grupos de pesquisa na rede. Desta forma, adotamos esta métrica de centralidade no restante da dissertação.

### 2.3.6 Assortatividade

A assortividade (*assortativity* ou *assortative mixing*) é uma métrica clássica de redes complexas que identifica o comportamento de como os nodos tendem a se agrupar na rede, e.g., uma rede de coautoria apresenta propriedades assortativas quando pesquisadores com o mesmo número de conexões tendem a se conectar com outros pesquisadores com o mesmo número de conexões.

A assortatividade pode ser representada visualmente a partir de um gráfico em que cada grau $k$ encontrado em pelo menos um nodo da rede é representado pelo grau médio $k_{nn}$ dos vizinhos dos nodos de grau $k$. Em grafos direcionados, este gráfico pode ser construído separadamente para graus de entrada e graus de saída. Segundo Ahn et al. [2007], esta métrica também pode ser expressa numericamente utilizando o coeficiente de correlação de Pearson:

$$r = \frac{\langle k_i k_j \rangle - \langle k_i \rangle \langle k_j \rangle}{\sqrt{(\langle k_i^2 \rangle - \langle k_i \rangle^2)(\langle k_j^2 \rangle - \langle k_j \rangle^2)}} \tag{2.6}$$

onde $k_i$ e $k_j$ são os graus dos nodos que constituem uma aresta e a notação $\langle\rangle$ representa a média sobre todas as arestas da rede.

Se a rede possui assortatividade negativa, nodos que possuem grau elevado tendem a se conectar a nodos com menor grau, e vice versa. O coeficiente $r$ pode variar entre -1 e 1, onde $r > 0$ indica que a rede possui propriedades assortativas, ou seja, nodos com graus semelhantes tendem a estabelecer conexões na rede, já $r < 0$ indica que a rede possui propriedades disassortativas, existindo maior probabilidade de encontrar arestas entre nodos de graus diferentes. Por exemplo, uma rede de coautoria com assortatividade negativa pode indicar que os pesquisadores seniores estão se conectando a pesquisadores menos experientes, e.g., alunos de doutorado.

# Capítulo 3

# Comunidades

Neste capítulo discutimos sobre as comunidades científicas que utilizamos em nossas análises e, em seguida, apresentamos uma definição para identificação do que chamamos de núcleo das comunidades.

## 3.1 Comunidades Científicas

Dada uma rede social, uma comunidade pode ser compreendida como um grupo denso de nodos dessa rede que possuem mais arestas interligando-os entre si, do que arestas interligando-os ao restante da rede. Existem múltiplas definições e estratégias para identificar comunidades e elas variam de acordo com o contexto [Kleinberg, 2008; Leskovec et al., 2010]. No nosso contexto, uma comunidade científica pode ser definida em termos de uma grande e consolidada conferência científica capaz de reunir pesquisadores que trabalham em uma mesma área de pesquisa ao longo de vários anos.

A fim de construir um conjunto de comunidades científicas, coletamos dados da DBLP[1] [Ley, 2009], uma biblioteca digital que contém mais de 2,1 milhões de publicações de 1,2 milhões de autores e que provê informações bibliográficas dos principais anais de conferências e periódicos da área de Ciência da Computação. A DBLP disponibiliza todo o seu conjunto de dados no formato XML (*eXtensible Markup Language*), o que facilita a obtenção dos dados e a construção de comunidades científicas inteiras.

Cada publicação é acompanhada por seu título, lista de autores, ano de publicação e veículo de publicação, i.e., conferência ou periódico. Para o propósito do nosso trabalho, consideramos uma comunidade científica como um grafo em que os nodos representam pesquisadores e as arestas ligam os coautores de artigos de uma mesma

[1] http://dblp.uni-trier.de/

comunidade. A fim de definir tais comunidades, focamos nas publicações das principais conferências (*flagship*) dos SIGs da ACM. Assim, definimos uma comunidade científica como formada por pesquisadores interligados entre si por serem coautores de um algum artigo dessas conferências, fazendo com que elas atuem como comunidades onde coautorias são formadas. Ao criarmos tais comunidades, removemos conferências sem dados suficientes para uma análise temporal, bem como conferências cujo histórico completo não está registrado na DBLP.

No total, 22 comunidades científicas foram construídas. A Tabela 3.1 lista essas comunidades, incluindo o respectivo SIG, acrônimo, período considerado (algumas conferências tiveram seu período reduzido para evitar hiato nos dados), número total de autores, publicações e edições, bem como razões extraídas desses três últimos dados.

Tabela 3.1: Estatísticas da DBLP das conferências *flagship* dos SIGs da ACM

| SIG | Conferência | Período | Autores | Publicações | Edições | Aut/Edi | Pub/Edi | Aut/Pub |
|---|---|---|---|---|---|---|---|---|
| SIGACT | STOC | 1969-2012 | 2159 | 2685 | 44 | 49,07 | 61,02 | 0,80 |
| SIGAPP | SAC | 1993-2011 | 9146 | 4500 | 19 | 481,37 | 236,84 | 2,03 |
| SIGARCH | ISCA | 1976-2011 | 2461 | 1352 | 36 | 68,36 | 37,56 | 1,82 |
| SIGBED | HSCC | 1998-2012 | 846 | 617 | 15 | 56,40 | 41,13 | 1,37 |
| SIGCHI | CHI | 1994-2012 | 5095 | 2819 | 19 | 268,16 | 148,37 | 1,81 |
| SIGCOMM | SIGCOMM | 1988-2011 | 1593 | 796 | 24 | 66,38 | 33,17 | 2,00 |
| SIGCSE | SIGCSE | 1986-2012 | 3923 | 2801 | 27 | 145,30 | 103,74 | 1,40 |
| SIGDA | DAC | 1964-2011 | 8876 | 5693 | 48 | 184,92 | 118,60 | 1,56 |
| SIGDOC | SIGDOC | 1989-2010 | 1071 | 810 | 22 | 48,68 | 36,82 | 1,32 |
| SIGGRAPH | SIGGRAPH | 1985-2003 | 1920 | 1108 | 19 | 101,05 | 58,32 | 1,73 |
| SIGIR | SIGIR | 1978-2011 | 3624 | 2687 | 34 | 106,59 | 79,03 | 1,35 |
| SIGKDD | KDD | 1995-2011 | 3078 | 1699 | 17 | 181,06 | 99,94 | 1,81 |
| SIGMETRICS | SIGMETRICS | 1981-2011 | 2083 | 1174 | 31 | 67,19 | 37,87 | 1,77 |
| SIGMICRO | MICRO | 1987-2011 | 1557 | 855 | 25 | 62,28 | 34,20 | 1,82 |
| SIGMM | MM | 1993-2011 | 5400 | 2928 | 19 | 284,21 | 154,11 | 1,84 |
| SIGMOBILE | MOBICOM | 1995-2011 | 1151 | 480 | 17 | 67,71 | 28,24 | 2,40 |
| SIGMOD | SIGMOD | 1975-2012 | 4202 | 2669 | 38 | 110,58 | 70,24 | 1,57 |
| SIGOPS | PODC | 1982-2011 | 1685 | 1403 | 30 | 56,17 | 46,77 | 1,20 |
| SIGPLAN | POPL | 1975-2012 | 1527 | 1217 | 38 | 40,18 | 32,03 | 1,25 |
| SIGSAC | CCS | 1996-2011 | 1354 | 676 | 16 | 84,63 | 42,25 | 2,00 |
| SIGSOFT | ICSE | 1987-2011 | 3502 | 2248 | 25 | 140,08 | 89,92 | 1,56 |
| SIGWEB | CIKM | 1992-2011 | 4978 | 2623 | 20 | 248,90 | 131,15 | 1,90 |

## 3.2 Definição do Núcleo das Comunidades

Tentativas anteriores para identificar o núcleo de comunidades científicas são baseadas em abordagens algorítmicas que visam identificar conjuntos densos de nodos na rede [Seifi & Guillaume, 2012]. Entretanto, como planejamos investigar o papel do núcleo na estrutura da rede, qualquer abordagem que faz uso da estrutura da rede para identificar tais nodos poderia nos levar a um conjunto de pesquisadores enviesado. Em vez disso, focamos no desenvolvimento de uma métrica que quantificasse o envolvi-

mento de um pesquisador em uma comunidade científica durante um certo período de tempo. Intuitivamente, essa métrica deveria ser capaz de capturar (i) a prolificidade de um pesquisador em diferentes comunidades e (ii) a frequência do envolvimento daquele pesquisador com a comunidade em um certo período de tempo.

Em primeiro lugar, a fim de capturar a prolificidade de um pesquisador, usamos o índice h [Hirsch, 2005], uma métrica largamente adotada para esse propósito. Essa métrica consiste de um índice que tenta medir tanto a produtividade quanto o impacto dos trabalhos publicados de um dado pesquisador. Ela baseia-se no conjunto de artigos mais citados de um pesquisador e no número de citações desse pesquisador com pelo menos $h$ citações. Mais especificamente, um pesquisador tem um índice h $i$ se publicou $i$ artigos que receberam pelo menos $i$ citações. Assim, por exemplo, se um pesquisador possui 10 artigos com pelo menos 10 citações, seu índice h final é 10.

Em segundo lugar, como uma tentativa de capturar a importância de um pesquisador em uma comunidade específica em um certo período de tempo, multiplicamos o valor do seu índice h ao final desse período pelo número de publicações desse pesquisador nessa comunidade no mesmo período. Denominamos essa métrica de *Community Score* (*CoScore*) [Alves et al., 2013]. Mais formalmente, o *CoScore* de um pesquisador $p$ em uma comunidade $c$ durante um período de tempo $t$, $CoScore_{p,c,t}$, é dado pelo seu índice h ($h_{p,t}$) ao final do período $t$ multiplicado pelo seu número de publicações na comunidade $c$ durante $t$ ($\#\text{publicações}_{p,c,t}$), como expresso pela Equação 3.1.

$$CoScore_{p,c,t} = h_{p,t} \times \#\text{publicações}_{p,c,t} \tag{3.1}$$

Como podemos ver, a primeira parte da equação captura a importância de um pesquisador para a comunidade científica como um todo em um determinado período de tempo, independentemente de qualquer área de pesquisa específica, e a segunda parte pesa essa importância baseada na atividade do pesquisador em uma certa comunidade. A fim de computar o *CoScore* para os membros de uma comunidade, definimos o núcleo de uma comunidade em um certo período de tempo como sendo formado pelos pesquisadores com o melhor escore naquela comunidade em termos de seu *CoScore* em um dado período.

A seguir, na Subseção 3.2.1, detalhamos como estimamos o índice h dos pesquisadores ao longo do tempo. Então, na Subseção 3.2.2, discutimos como definimos dois importantes limiares: o tamanho do núcleo da comunidade e a janela de tempo usada em nossas análises.

### 3.2.1 Estimativa do Índice H dos Pesquisadores

Existem várias ferramentas que medem o índice h de pesquisadores, das quais o Google Citations[2] é a mais proeminente. No entanto, para ser incluído neste sistema, o pesquisador precisa se inscrever e criar explicitamente seu perfil. Em uma coleção preliminar com parte dos perfis de autores da DBLP, descobrimos que menos de 30% desses autores tinham um perfil no Google Citations. Assim, esta estratégia poderia reduzir nosso conjunto de dados e, potencialmente, introduzir algum viés na análise das comunidades.

Para evitar essa limitação, utilizamos os dados do projeto SHINE[3] (*Simple HINdex Estimator*) para estimar o índice h dos pesquisadores. O SHINE fornece um *website* que permite seus usuários verificar o índice h de quase duas mil conferências de Ciência da Computação. Para isso, seus desenvolvedores realizaram uma coleta no Google Scholar[4], buscando pelo título dos artigos publicados nessas conferências, o que lhes permitiu efetivamente estimar o índice h das conferências alvo com base nas citações computadas pelo Google Scholar. Embora o SHINE permita somente consultas ao índice h de conferências, seus desenvolvedores gentilmente nos permitiram acessar o seu conjunto de dados para estimarmos o índice h dos pesquisadores baseados nas conferências coletadas. Embora isso possa gerar um viés em nossas estimativas, segundo Laender et al. [2008], os pesquisadores da área de Ciência da Computação tendem a publicar cerca 2,5 artigos em conferências para cada artigo publicado em periódicos, contrabalançando uma possível discrepância.

No entanto, existem duas importantes limitações com esta estratégia. A primeira limitação é referente ao dados coletados, uma vez que o SHINE não coleta todas as conferências existentes da área de Ciência da Computação, o índice h dos pesquisadores pode ser subestimado quando calculado com esses dados. Para investigar esta questão, comparamos o índice h de um conjunto de pesquisadores que possuem perfil no Google Citations, com seu índice h estimado com base nos dados do SHINE. Para isso, selecionamos aleatoriamente 10 pesquisadores de cada conferência da Tabela 3.1 e extraímos o índice h de seus perfis no Google Scholar. Em comparação com o índice h que estimamos a partir dos dados do SHINE, os valores do Google Citations são, em média, 50% maiores. A Figura 3.1 mostra o gráfico de dispersão para os dois índices h medidos. Podemos observar que, embora o índice h calculado a partir dos dados do SHINE seja menor, as duas medidas são altamente correlacionadas. O coeficiente de correlação de Pearson é de 0,85, apresentando uma forte correlação positiva, o que

[2]http://scholar.google.com/citations
[3]http://shine.icomp.ufam.edu.br/
[4]http://scholar.google.com

indica que os pesquisadores podem ter seu índice h proporcionalmente estimados em ambos os sistemas.

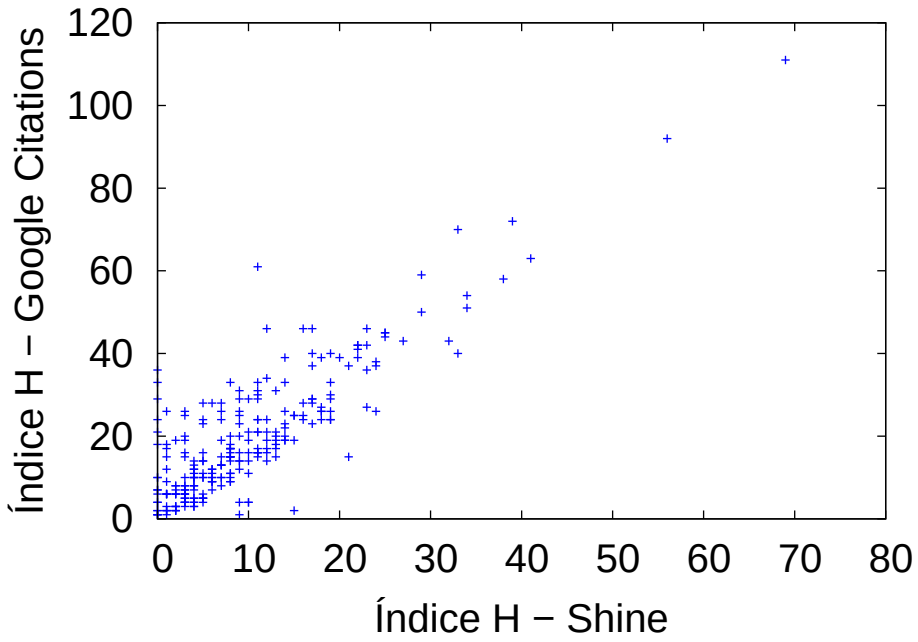
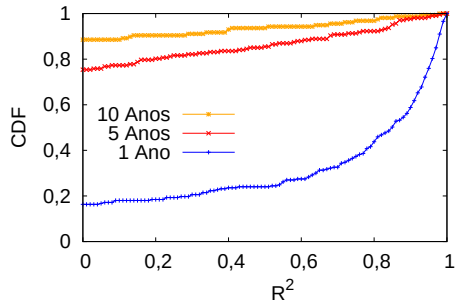



Figura 3.1: Correlação entre o índice h estimado e o Google Citations

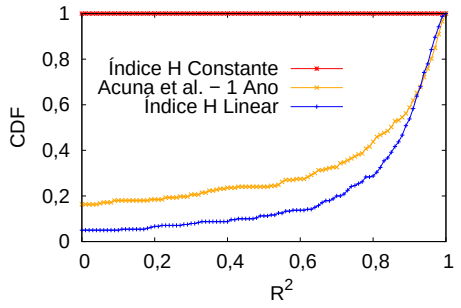
A segunda limitação está relacionada à evolução do índice h ao longo do tempo. O SHINE coleta somente o total atual de citações de cada artigo, sendo possível estimar apenas o valor final do índice h de um dado pesquisador. De forma análoga à abordagem anterior, selecionamos aleatoriamente 10 pesquisadores de cada conferência e extraímos do Google Citations o número de citações dos artigos de cada pesquisador ao longo do tempo, nos permitindo, assim, obter a curva de evolução do índice h desses pesquisadores.

Acuna et al. [2012] apresentam um método que inclui equações capazes de predizer o índice h de um pesquisador daqui a um, cinco ou dez anos. Desta forma, utilizamos cada equação para estimar o índice h dos pesquisadores e comparamos com os dados coletados do Google Citations utilizando regressão linear, $R^2$. De acordo com a CDF (*Cumulative Distribution Function*) na Figura 3.2a, podemos observar que a equação de um ano é a que mais se assemelha aos valores do Google Citations, tendo mais de 70% dos pesquisadores com $R^2$ superior a 60%. Com base nas equações definidas por Acuna et al. [2012], computamos o índice h dos pesquisadores utilizando três abordagens: (i) a primeira fixa o índice h atual do pesquisador ao longo do tempo, (ii) em seguida utilizamos a equação capaz de prever o índice h ano a ano dos pesquisadores definida por Acuna et al. [2012], e (iii) por fim, utilizamos uma evolução linear do

índice h do pesquisador considerando a sua primeira e última data de publicação. A Figura 3.2b mostra a CDF da $R^2$ entre os valores do Google Scholar e os valores estimados utilizando as três abordagens propostas, sendo possível observar que a abordagem utilizando a evolução linear é a que mais se aproxima dos valores do Google Scholar, tendo mais de 60% dos pesquisadores com $R^2$ superior a 80%.




(a) Valores gerados utilizando o método proposto por Acuna et al. [2012]




(b) Valores gerados utilizando as três estratégias

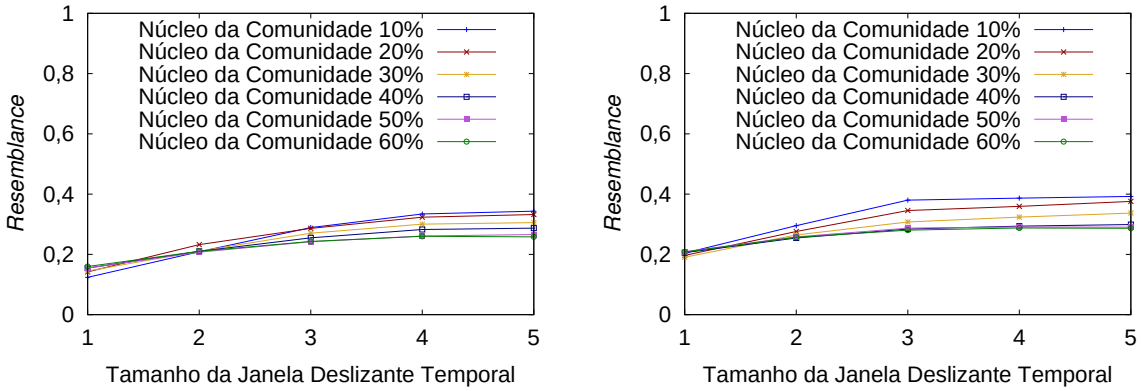
Figura 3.2: Evolução do índice h

### 3.2.2 Definição dos Limiares

Nossa estratégia para definir os dois limiares necessários para definir o núcleo das comunidades consiste em variar cada um deles e quantificar como eles impactam nas mudanças dos membros desse núcleo. Para medir essas mudanças, calculamos a métrica *resemblance*, conforme definida por Viswanath et al. [2009], que mede a fração dos membros do núcleo no tempo $t_0$ que permanecem no núcleo no tempo $t_1$. Para cada comunidade, variamos o tamanho da janela de 1 a 5 anos e o tamanho do núcleo de 10% a 60% do total dos respectivos pesquisadores.

Intuitivamente, uma alta variação da métrica *resemblance* indica uma escolha ruim dos limiares. Assim, procuramos por limiares cujas mudanças causassem pequenas alterações nos valores dessa métrica. A Figura 3.3 mostra os valores do *resemblance* em função do tamanho da janela, fornecendo diferentes curvas para o tamanho do núcleo da comunidade. Mostramos aqui apenas as curvas das comunidades SIGMOD e CHI, as curvas das demais comunidades podem ser encontradas no Apêndice A. Por inspeção visual definiríamos o tamanho do núcleo da comunidade como 10% devido à proximidade das curvas e o tamanho da janela como 2 ou 3, uma vez que a maior parte das comunidades mostra um valor do *resemblance* mais estável após esses valores. Para nos ajudar a decidir, calculamos o coeficiente angular das curvas com tamanho do

núcleo de 10% para cada comunidade e a média do coeficiente angular para elas. Com base nesses valores, definimos o tamanho da janela para nossos experimentos como sendo de 3 anos.



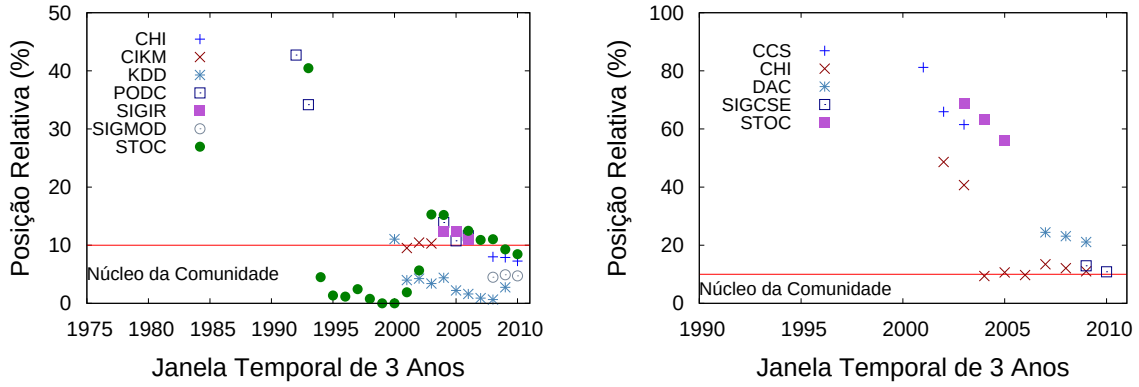

(a) SIGMOD (b) CHI

Figura 3.3: Média dos valores de *resemblance*

### 3.2.3 Validação

Com base no *CoScore*, esperamos que os membros do núcleo da comunidade sejam pesquisadores que contribuam ativamente com publicações em uma determinada comunidade. A validação desta suposição é, por natureza, subjetiva. Assim, fornecemos a seguir evidências que nossa abordagem captura corretamente essa característica esperada.

Primeiro, analisamos o *CoScore* de dois palestrantes convidados da conferência WWW 2013 realizada recentemente no Rio de Janeiro: Jon Kleinberg e Luis von Ahn. A Figura 3.4 mostra a posição em termos da percentagem (e.g., a posição 5% de uma dada comunidade), desses dois pesquisadores nas comunidades em que eles têm publicado. A linha inferior divide os membros do núcleo da comunidade dos demais membros. Podemos notar que Jon Kleinberg foi membro do núcleo da comunidade STOC, uma conferência teórica, por anos. Mais precisamente, ele foi parte do núcleo da STOC por doze anos, publicando sete artigos na STOC em um único período de três anos. Com o envolvimento de Kleinberg na KDD, ele se tornou menos ativo na STOC e saiu do núcleo dessa comunidade por algum tempo. Durante esse período, ele publicou muitos artigos na KDD, enquanto suas publicações na STOC foram reduzindo. Com relação ao pesquisador Luis von Ahn, podemos notar que ele é mais ativo na




(a) Jon Kleinberg

(b) Luis von Ahn

Figura 3.4: *CoScore* de dois palestrantes convidados da WWW 2013

comunidade CHI, na qual ele publicou seis artigos ao longo de sua vida acadêmica. Ele chegou ao núcleo da comunidade CHI em duas janelas de tempo, publicando quatro artigos na CHI em um único período.

Tabela 3.2: Pesquisadores das conferências CHI, ICSE, KDD e POPL que apareceram com mais frequência no núcleo da comunidade através dos anos.

| **CHI** | **ICSE** | **KDD** | **POPL** |
|---|---|---|---|
| **Scott E. Hudson** | **Victor R. Basili** | **Heikki Mannila** | Thomas W. Reps |
| **Hiroshi Ishii** | **Barry W. Boehm** | Hans-Peter Kriegel | Martín Abadi |
| **Steve Benford** | **Jeff Kramer** | **Jiawei Han** | John C. Mitchell |
| **George G. Robertson** | **Mary Shaw** | Martin Ester | Robert Harper |
| **Shumin Zhai** | Dewayne E. Perry | **Rakesh Agrawal** | Zohar Manna |
| **Brad A. Myers** | Don S. Batory | Bing Liu | Benjamin C. Pierce |
| **Robert E. Kraut** | Mary Jean Harrold | Ke Wang | Amir Pnueli* |
| **Elizabeth D. Mynatt** | **Lori A. Clarke** | **Padhraic Smyth** | **Barbara Liskov*** |
| **Ravin Balakrishnan** | Gruia-Catalin Roman | Philip S. Yu | Martin C. Rinard |
| **James A. Landay** | Premkumar T. Devanbu | Charu C. Aggarwal | Luca Cardelli |
| Ken Hinckley | Gail C. Murphy | **Vipin Kumar** | Thomas A. Henzinger |
| **Mary Czerwinski** | **Richard N. Taylor** | Wynne Hsu | **Ken Kennedy** |
| **Carl Gutwin** | **David Garlan** | Qiang Yang | **Matthias Felleisen** |
| **Gregory D. Abowd** | Michael D. Ernst | **Christos Faloutsos** | Edmund M. Clarke* |
| Michael J. Muller | James D. Herbsleb | William W. Cohen | Mitchell Wand |
| **Susan T. Dumais** | Lionel C. Briand | Pedro Domingos | David Walker |
| Loren G. Terveen | Gregg Rothermel | Eamonn J. Keogh | Simon L. Peyton Jones |
| **Steve Whittaker** | Kevin J. Sullivan | Alexander Tuzhilin | Shmuel Sagiv |
| W. Keith Edwards | **David Notkin** | Mohammed Javeed Zaki | Barbara G. Ryder |
| **John M. Carroll** | Douglas C. Schmidt | Mong-Li Lee | Alexander Aiken |

Em seguida, computamos a posição dos pesquisadores que aparecem com mais frequência no núcleo de suas comunidades científicas. Escolhemos as conferências CHI, ICSE, KDD, POPL, SIGCOMM, SIGIR e SIGMOD para mostrar os 20

Tabela 3.3: Pesquisadores das conferências SIGCOMM, SIGGRAPH, SIGIR e SIGMOD que apareceram com mais frequência no núcleo da comunidade através dos anos.

| SIGCOMM | SIGGRAPH | SIGIR | SIGMOD |
|---|---|---|---|
| **Scott Shenker** | **Donald P. Greenberg** | **W. Bruce Croft** | **Michael Stonebraker** |
| George Varghese | **Pat Hanrahan** | Clement T. Yu | **David J. DeWitt** |
| **Donald F. Towsley** | Demetri Terzopoulos | **Gerard Salton** | **Philip A. Bernstein** |
| Ion Stoica | **David Salesin** | Alistair Moffat | H. V. Jagadish |
| Hui Zhang | **Michael F. Cohen** | **Susan T. Dumais** | Christos Faloutsos |
| Deborah Estrin | **Richard Szeliski** | James Allan | **Rakesh Agrawal** |
| Hari Balakrishnan | John F. Hughes | Yiming Yang | **Michael J. Carey** |
| Robert Morris | N. Magnenat-Thalmann | Edward A. Fox | **H. Garcia-Molina** |
| Thomas E. Anderson | **Tomoyuki Nishita** | James P. Callan | Jiawei Han |
| Ramesh Govindan | **Andrew P. Witkin** | Chris Buckley | Raghu Ramakrishnan |
| Srinivasan Seshan | Norman I. Badler | **C. J. van Rijsbergen** | Jeffrey F. Naughton |
| David Wetherall | **Peter Schröder** | Justin Zobel | **Jim Gray**$^\star$ |
| Yin Zhang | Steven Feiner | Ellen M. Voorhees | Hans-Peter Kriegel |
| Jennifer Rexford | **Hugues Hoppe** | Mark Sanderson | Gerhard Weikum |
| Jia Wang | **Jessica K. Hodgins** | **Norbert Fuhr** | Philip S. Yu |
| J. J. Garcia-Luna-Aceves | **Greg Turk** | Nicholas J. Belkin | Divesh Srivastava |
| Randy H. Katz | **Marc Levoy** | Chengxiang Zhai | Joseph M. Hellerstein |
| Albert G. Greenberg | **P. Prusinkiewicz** | Charles L. A. Clarke | Krithi Ramamritham |
| Mark Handley | Eihachiro Nakamae | Alan F. Smeaton | Nick Roussopoulos |
| **Simon S. Lam** | Dimitris N. Metaxas | Gordon V. Cormack | **Surajit Chaudhuri** |

pesquisadores melhores colocados, conforme destacado nas Tabelas 3.2 e 3.3. Como podemos notar, vários pesquisadores renomados aparecem no topo dessas listas, incluindo os palestrantes convidados das últimas edições das respectivas conferências e pesquisadores premiados por suas contribuições naquelas comunidades, cujos nomes aparecem em negrito, incluindo alguns que também receberam o ACM *A.M. Turing Award*[5] (marcados com $^\star$). De fato, analisando os pesquisadores premiados de cada comunidade, descobrimos que grande parte deles apareceu no núcleo da comunidade pelo menos uma vez na história da conferência. Mais especificamente, estas frações são de 58% dos membros premiados da CHI[6], 65% para a ICSE[7], 75% para a KDD[8], 26% para a POPL[9], 35% para a SIGCOMM[10], 50% para a SIGGRAPH[11], 70% para a SIGIR[12], e 85% para a SIGMOD[13]. Exceto para as conferências SIGCOMM e POPL (SIGPLAN), cujos SIGs patrocinam outros eventos que não foram considerados em nosso conjunto de dados, as outras seis comunidades apresentam um número muito

[5]http://amturing.acm.org/byyear.cfm
[6]http://www.sigchi.org/about/awards
[7]http://www.sigsoft.org/awards/outResAwd.htm
[8]http://www.sigkdd.org/awards_innovation.php
[9]http://www.sigplan.org/Awards/Achievement/Main
[10]http://www.sigcomm.org/awards/sigcomm-awards
[11]http://www.siggraph.org/participate/awards
[12]http://www.sigir.org/awards/awards.html
[13]http://www.sigmod.org/sigmod-awards

alto de membros premiados que aparecem pelo menos uma vez em seus respectivos núcleos. Além disso, podemos observar que a comunidade POPL possui pelo menos três pesquisadores que receberam o ACM *A.M. Turing Award*, um dos mais importantes prêmios da comunidade científica. Estas observações fornecem evidências que nossa abordagem captura corretamente a noção do núcleo de uma comunidade científica.

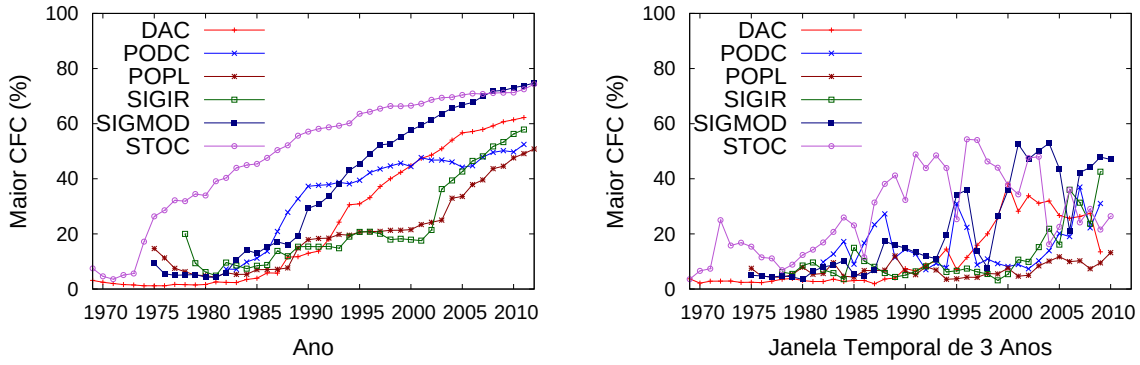
# Capítulo 4

# Análise das Comunidades

Neste capítulo apresentamos uma série de análises sobre as comunidades científicas. Primeiro, analisamos como as propriedades dessas comunidades evoluem. Em seguida, comparamos, ao longo do tempo, as propriedades dos membros do núcleo com os demais membros das comunidades. Finalmente, calculamos a média do *CoScore* das comunidades para investigar variações nas propriedades dos núcleos e correlacionar essas variações com as propriedades das comunidades.

## 4.1 Evolução das Comunidades

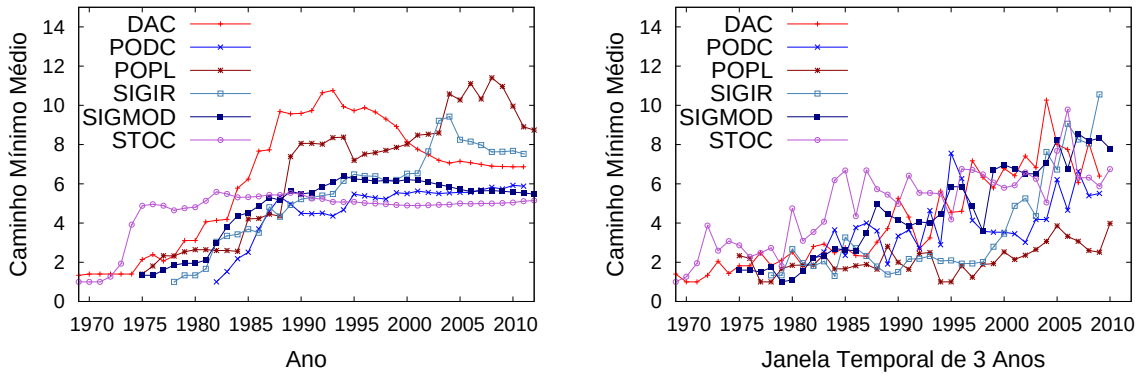
A fim de estudar a evolução das principais propriedades estruturais das comunidades científicas, examinamos várias métricas de redes para cada uma das comunidades consideradas. Apresentamos aqui cinco métricas populares: assortatividade, caminho mínimo médio (CMM), coeficiente de agrupamento (CA), tamanho do maior componente fracamente conectado (CFC) e grau médio dos nodos. As Figuras 4.1, 4.2, 4.3, 4.4 e 4.5 mostram como cada uma dessas cinco métricas variam ao longo do tempo. Apresentamos essas métricas para um conjunto de seis comunidades científicas selecionadas entre aquelas que mais se estendem ao longo do tempo em nosso conjunto de dados. Nossas análises são realizadas sob duas perspectivas. A primeira consiste em analisar a evolução da rede ano a ano, acumulando nodos e arestas da instância final do grafo. Essa perspectiva nos permite observar a estrutura final de uma comunidade em função do tempo. A segunda perspectiva consiste em analisar instâncias construídas com base em nodos e arestas criados em uma janela de tempo predefinida (três anos, tal como discutido na Subsecção 3.2.2). Esta análise nos permite investigar as variações da rede com potencial para impactar a sua estrutura final. Os resultados da análise são semelhantes para as outras comunidades e podem ser observados no Apêndice B.




(a) Maior CFC final

(b) Maior CFC por janela

Figura 4.1: Maior CFC das comunidades científicas

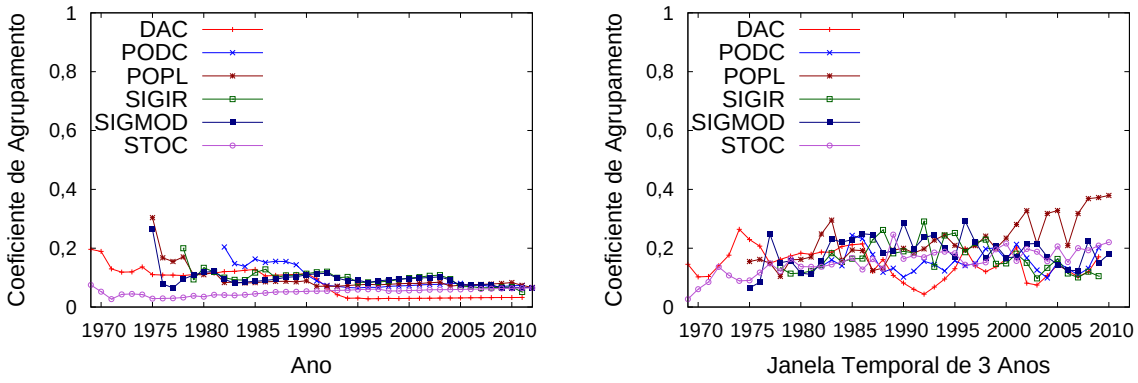



(a) CMM final

(b) CMM por janela

Figura 4.2: Caminho mínimo médio das comunidades científicas

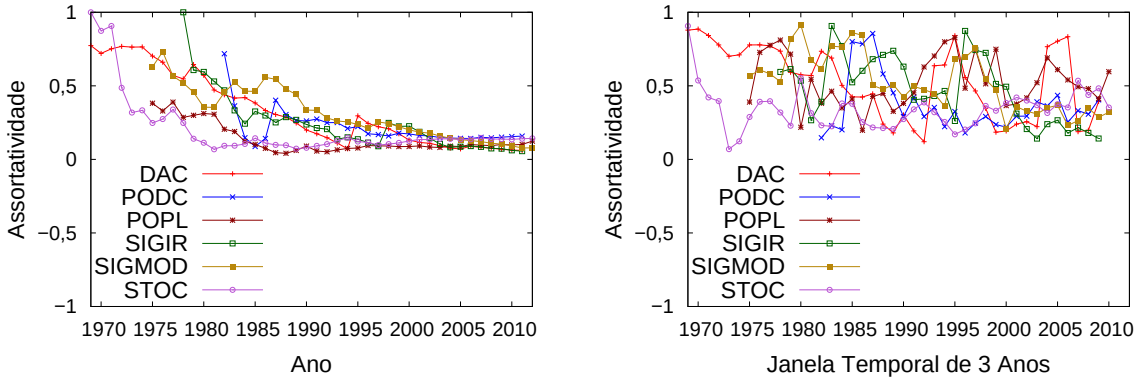
Notamos a partir da Figura 4.1 que o maior CFC tende a aumentar largamente em função do tempo. Isto sugere que na fase inicial, as comunidades científicas são formadas por vários grupos de pesquisa pequenos e segregados. Com o tempo, alguns pesquisadores (e.g., estudantes) deixam suas instituições e começam a colaborar com outros grupos de pesquisa. Além disso, como a comunidade evolui, líderes de grupos de pesquisa tendem a colaborar com outros colegas da mesma comunidade. Assim, com o tempo, pesquisadores de diferentes grupos tendem a colaborar e aumentar o tamanho do maior CFC. Como consequência, o caminho mínimo médio, calculado apenas sobre o maior CFC, tende a aumentar, se tornando estável em torno de valores semelhantes aos de redes de mundo pequeno (ou seja, caminhos que contenham de 4




(a) CC final

(b) CC por janela

Figura 4.3: Coeficiente de agrupamento das comunidades científicas

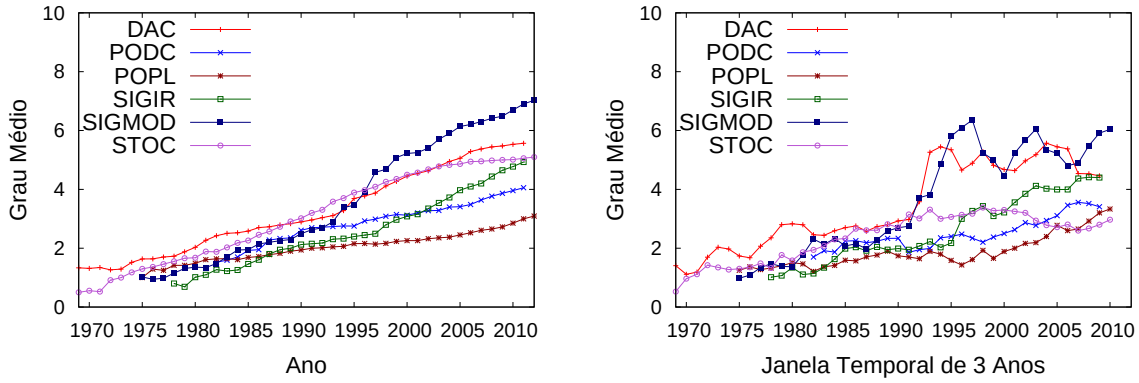



(a) Assortatividade final

(b) Assortatividade por janela

Figura 4.4: Assortatividade das comunidades científicas

a 10 arestas) [Mislove et al., 2007; Backstrom et al., 2012], conforme Figura 4.2. Também podemos observar na Figura 4.3 que o coeficiente de agrupamento médio tende a valores entre 0,1 e 0,2, sugerindo que os coautores de um pesquisador têm entre 10% e 20% de chance de serem conectados entre si. Esses valores tendem a diminuir ligeiramente ao longo do tempo, tal como pequenos componentes tendem a se conectar para formarem componentes maiores, reduzindo a média do coeficiente de agrupamento. Quando se trata de assortatividade, observamos na Figura 4.4 que esta medida tende a 0, mas ainda assim é positiva. Isso significa que há uma ligeira tendência nessas comunidades de nodos se conectarem com outros de grau similar. Um valor positivo para a assortatividade é uma característica típica das redes sociológicas [Newman & Park,




(a) Grau médio final

(b) Grau médio por janela

Figura 4.5: Grau médio das comunidades científicas

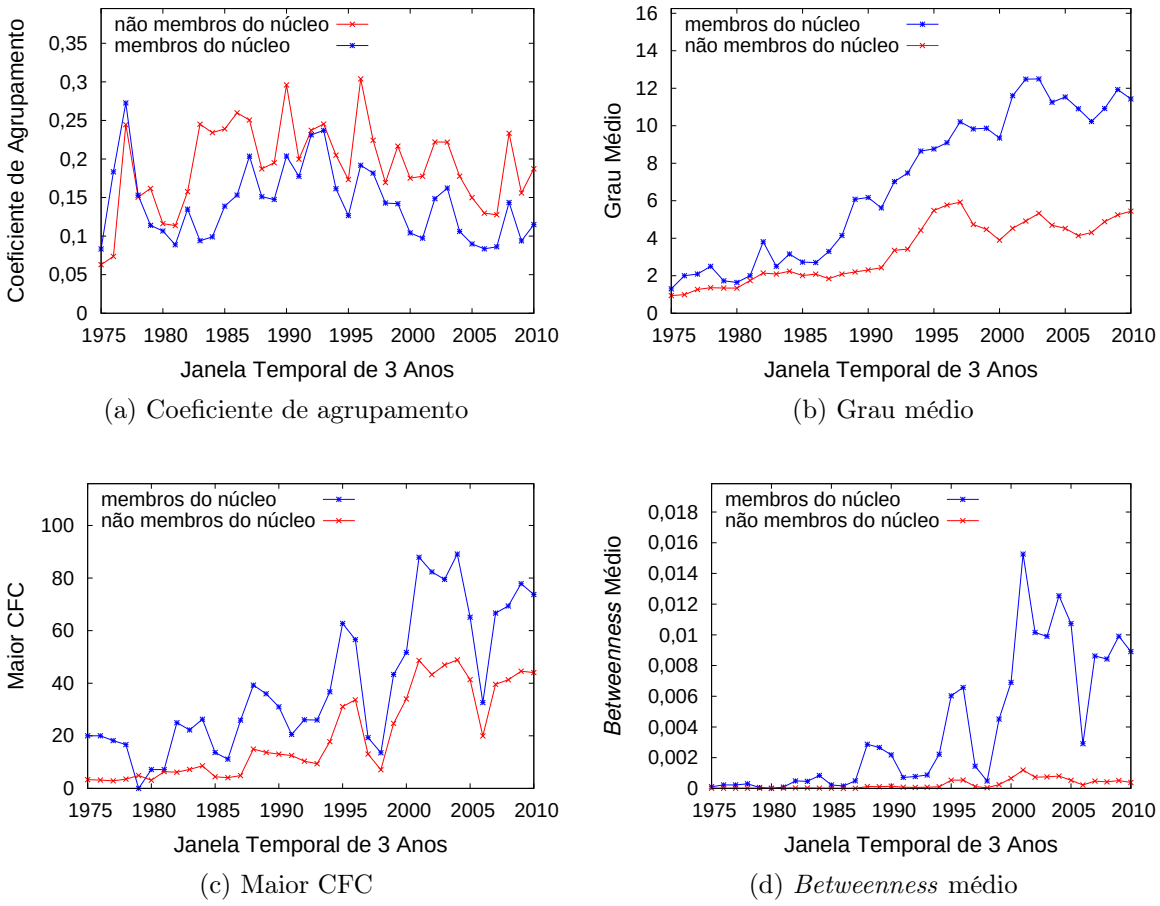
2003]. Finalmente, podemos observar na Figura 4.5 que o grau médio dos nodos tende a aumentar ao longo do tempo, mesmo com a assortatividade tendendo a 0. Isto pode indicar uma renovação, em que novos pesquisadores se juntam à rede através de pesquisadores mais experientes, por exemplo, estudantes com seus orientadores.

Em geral, podemos observar que as comunidades científicas têm características de evolução semelhantes e que essas propriedades são dinâmicas, mudando ao longo do tempo. Mais importante ainda, nossas observações sugerem que um pequeno grupo de pesquisadores que fazem parte do núcleo são responsáveis por criar caminhos entre grupos de pesquisa menores e mais conectados. A fim de investigar melhor os pesquisadores que fazem parte do núcleo, na próxima seção comparamos membros e não membros dos núcleos das comunidades.

## 4.2 Caracterização dos Núcleos das Comunidades

*Até que ponto as propriedades dos membros do núcleo diferem dos demais membros das comunidades?* Para responder a essa pergunta, calculamos as propriedades de rede para os membros e não membros dos núcleos das comunidades. Consideramos a análise de janelas de tempo para compreender as variações que essas duas classes podem ter na medida global. A Figura 4.6 mostra o grau médio e o coeficiente de agrupamento médio calculados para membros e não membros do núcleo da comunidade SIGMOD, os resultados obtidos para as demais comunidades podem ser encontrados no Apêndice C, sendo que nossas observações são válidas para todas elas. Além disso, também medimos a fração de membros do núcleo, bem como dos não membros que estão no maior CFC

e calculamos o *betweenness* médio de cada um desses grupos de membros.




(a) Coeficiente de agrupamento

(b) Grau médio

(c) Maior CFC

(d) *Betweenness* médio

Figura 4.6: Propriedades da comunidade SIGMOD para os membros e não membros do núcleo

Podemos fazer observações importantes a partir dessas análises. Primeiro, podemos observar que o grau médio dos membros dos núcleos é consideravelmente maior em comparação com o dos não membros, que tendem a estabelecer mais e mais conexões em função do tempo. No entanto, o coeficiente de agrupamento dos membros do núcleo tendem a ser ligeiramente menor quando comparados com o dos não membros, o que significa que eles podem atuar como *hubs*, conectando diferentes grupos com pequenas interseções. Ao analisar a fração de membros do núcleo que fazem parte do maior CFC, podemos notar que é muito maior do que a fração de não membros, sugerindo que eles podem estar conectando componentes menores. Confirmamos essas observações, analisando a centralidade desses grupos de pesquisadores através da métrica *betweenness*. Podemos notar que o *betweenness* médio do núcleo da comunidade é muito maior, o

que significa que um maior número de caminhos mais curtos incluem esses nodos.

Na próxima seção investigamos como aspectos dos membros dos núcleos podem impactar a estrutura geral das comunidades.

## 4.3 Impacto dos Membros dos Núcleos na Estrutura Topológica das Comunidades

Agora analisamos o quanto as variações no núcleo da comunidade afetam a estrutura da rede. Para isso, calculamos a média do *CoScore* dos membros de cada comunidade ao longo do tempo. Intuitivamente, essa medida captura a prolificidade global e o grau de participação dos membros do núcleo em uma comunidade científica. A Figura 4.7 mostra o *CoScore* médio para um conjunto específico de comunidades em função do tempo. Nossa análise abrange todas as comunidades, podendo os gráficos das demais serem observados no Apêndice D. Plotamos em duas figuras distintas para facilitar a visualização. Podemos observar em todas as comunidades que, apesar de pequenas quedas, de forma geral o valor aumenta ao longo do tempo sofrendo variações. Podemos especular inúmeros fatores que são capazes de explicar essas pequenas variações, incluindo a expansão ou redução do número de artigos publicados, a ascensão e queda de temas com capacidade de atrair pesquisadores ou a perda de membros importantes do núcleo, membros envolvidos na organização de outras conferências, etc. No entanto, desconsiderando o que causou essas variações, queremos investigar se tais variações podem impactar diretamente a estrutura da rede.

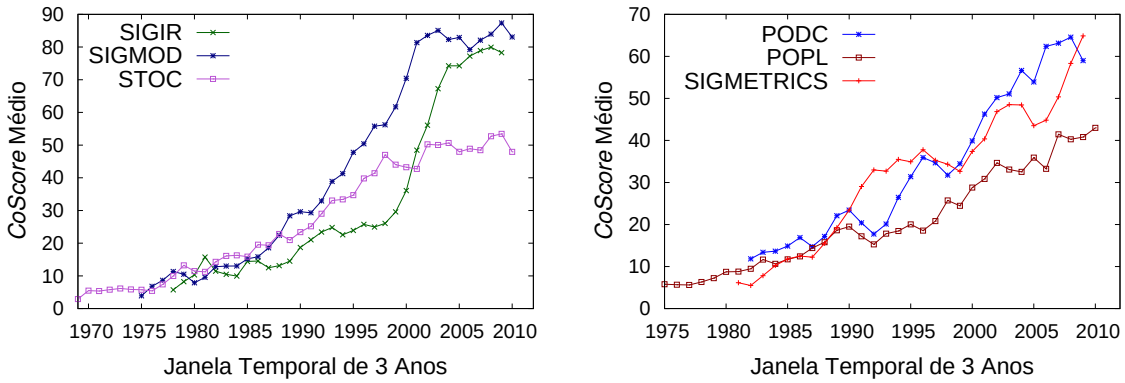



Figura 4.7: *CoScore* médio das comunidades científicas

Nossa abordagem para investigar esta questão consiste em calcular o coeficiente de correlação de Pearson entre a média do *CoScore* e as métricas de redes para cada

comunidade. A Tabela 4.1 apresenta esses valores.

Tabela 4.1: Correlação entre a média do *CoScore* e as métricas de redes

| Comunidade | Diâmetro | C. Mín. Méd. | Coef. Agrup. | Assort. | Maior CFC | Grau Méd. |
|---|---|---|---|---|---|---|
| CCS | 0,81 | 0,78 | -0,36 | -0,67 | 0,49 | 0,88 |
| CHI | 0,91 | 0,91 | -0,82 | -0,84 | 0,95 | 0,85 |
| CIKM | 0,79 | 0,79 | -0,53 | -0,82 | 0,65 | 0,97 |
| DAC | 0,94 | 0,94 | -0,41 | -0,52 | 0,90 | 0,90 |
| HSCC | 0,38 | 0,65 | -0,72 | -0,34 | 0,80 | 0,58 |
| ICSE | 0,73 | 0,75 | -0,39 | -0,72 | 0,44 | 0,98 |
| ISCA | 0,57 | 0,58 | 0,62 | -0,33 | 0,67 | 0,85 |
| KDD | 0,61 | 0,69 | -0,11 | -0,94 | 0,76 | 0,74 |
| MICRO | 0,51 | 0,48 | 0,49 | -0,23 | 0,38 | 0,86 |
| MM | 0,89 | 0,88 | -0,89 | -0,90 | 0,91 | 0,93 |
| MOBICOM | 0,73 | 0,81 | 0,15 | -0,41 | 0,81 | 0,80 |
| PODC | 0,56 | 0,56 | -0,18 | -0,25 | 0,33 | 0,94 |
| POPL | 0,70 | 0,67 | 0,83 | -0,04 | 0,54 | 0,92 |
| SAC | 0,76 | 0,77 | 0,06 | -0,60 | -0,57 | 0,76 |
| SIGCOMM | 0,63 | 0,67 | -0,15 | -0,94 | 0,90 | 0,88 |
| SIGCSE | 0,78 | 0,71 | -0,18 | -0,40 | 0,92 | 0,93 |
| SIGDOC | 0,90 | 0,92 | -0,21 | -0,92 | 0,88 | 0,91 |
| SIGGRAPH | 0,82 | 0,90 | -0,38 | -0,73 | 0,92 | 0,88 |
| SIGIR | 0,91 | 0,94 | -0,43 | -0,75 | 0,81 | 0,92 |
| SIGMETRICS | 0,57 | 0,50 | 0,57 | -0,58 | 0,45 | 0,94 |
| SIGMOD | 0,94 | 0,95 | 0,01 | -0,75 | 0,91 | 0,91 |
| STOC | 0,73 | 0,78 | 0,70 | 0,03 | 0,62 | 0,84 |
| **Média** | **0,73** | **0,76** | **-0,10** | **-0,58** | **0,66** | **0,87** |

A partir dessa análise, podemos notar que o diâmetro da rede de uma conferência possui uma forte correlação positiva com a média do *CoScore*, sendo este 0,73. Isto significa que, quando a média do *CoScore* de uma comunidade aumenta ou diminui, o diâmetro tende a seguir a mesma tendência. Isto sugere que os membros do núcleo podem conectar componentes menores, criando pontes entre eles, o que contribui para aumentar o diâmetro total da rede. Esta conjectura é também suportada pelo alto valor do coeficiente de correlação para o caminho mínimo médio (em média 0,76) e o tamanho do maior CFC (em média de 0,66).

Em seguida, podemos observar um alto valor positivo do coeficiente de correlação entre a média do *CoScore* das comunidades e o grau médio da rede. Também, podemos observar uma forte correlação negativa com a assortatividade da rede. Isto sugere que um aumento na média do *CoScore* aumenta o conjunto de nodos densamente conectados na rede. Entretanto, embora criem caminhos entre os componentes, esses nodos também tendem a se conectar principalmente com nodos de menor grau, diminuindo a assortatividade da rede. De fato, um pesquisador sênior tende a ser coautor de um grande número de estudantes e jovens pesquisadores, além de manter colaborações com pesquisadores seniores de outros grupos.

Finalmente, apesar das variações esperadas, observamos uma tendência clara para a maioria das comunidades em cada uma das métricas analisadas (i.e., claras correlações

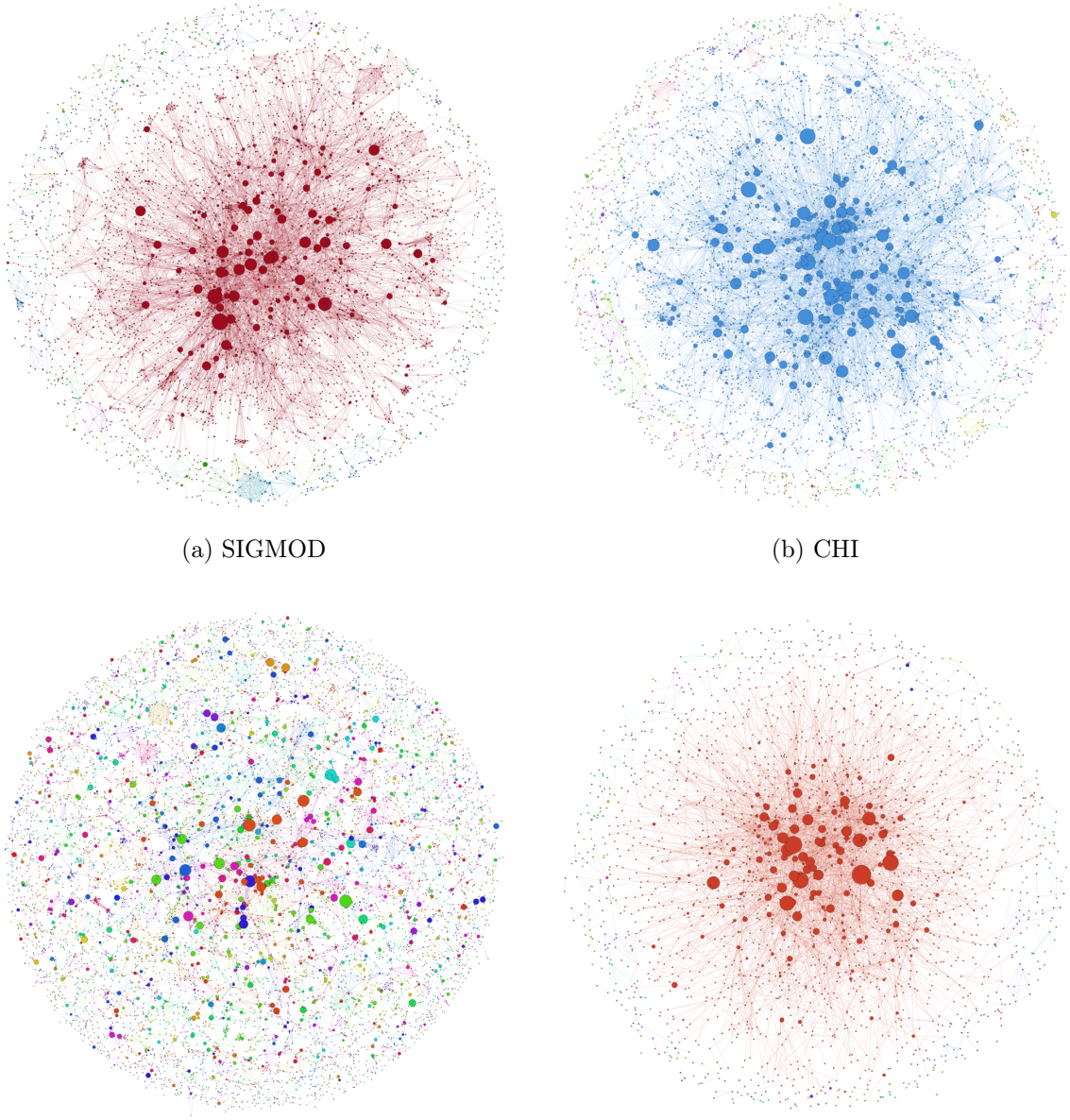
positivas ou negativas para a maioria das comunidades). Isso reforça que as nossas observações são válidas para um número significativo de comunidades científicas.

## 4.4 Visualização das Comunidades

Em complemento a nossas análises, plotamos as comunidades científicas acumulando todos os seus nodos e arestas ao longo do tempo. A Figura 4.8 apresenta a plotagem das comunidades SIGMOD, CHI, SAC, e STOC. Cada cor representa um componente conectado diferente e o tamanho dos nodos indica o número de vezes que o pesquisador apareceu no núcleo ao longo de todo o tempo de vida daquela comunidade. As demais comunidades podem ser observadas no Apêndice E.

A maioria das comunidades apresenta as mesmas características que as comunidades SIGMOD e CHI, possuindo um grande CFC bem definido chamado de maior CFC, sendo este composto por membros e não membros do núcleo. Observamos claramente que um grande número de membros do núcleo se encontra no maior CFC, com algumas pequenas exceções, sendo esses membros responsáveis por conectar grupos de outros pesquisadores, conforme apontando em nossas análises anteriormente. As comunidades SAC e STOC apresentam um comportamento atípico. A primeira não possui um grande CFC bem definido. Isto acontece porque essa conferência abrange várias áreas ao longo do tempo, dificultando assim a fixação dos pesquisadores na comunidade. Outro detalhe sobre a SAC é que ela não possui pesquisadores que participaram várias vezes do núcleo, o que indica uma frequente renovação dessa comunidade. Com relação à comunidade STOC, uma conferência teórica, é possível verificar o oposto da SAC. A comunidade possui um grande CFC muito bem definido e com grande parte dos membros do seu núcleo situada no centro da rede, indicando que os pesquisadores dessa comunidade interagem frequentemente, podendo, de certa forma, dificultar a entrada de novos membros.

Por fim, podemos notar que existe um grupo de pesquisadores que persiste no núcleo ao longo da vida de uma dada comunidade e a forma como os membros desse núcleo se organizam na rede reforça nossa teoria que esses pesquisadores atuam como pontes que interligam grupos de pesquisa dentro da comunidade.



(a) SIGMOD

(b) CHI

(c) SAC

(d) STOC

Figura 4.8: Instância final das comunidades científicas

# Capítulo 5

# Conclusões e Trabalhos Futuros

## 5.1 Revisão do Trabalho

Nesta dissertação apresentamos a caracterização de várias comunidades científicas presentes na DBLP, uma biblioteca digital da área de Ciência da Computação da qual extraímos e construímos 22 redes de colaboração científica. Em seguida, realizamos uma investigação profunda sobre os papéis que os membros dessas comunidades desempenham na formação e evolução topológica dessas redes.

Trabalhos anteriores estudam comunidades inteiras e são baseados em abordagens algorítmicas, tais como agrupamento hierárquico e *k-means*, que buscam identificar grupos de nodos mais agrupados entre si do que ao restante da rede ou nodos que possuem alguma característica topológica. Diferentemente, focamos aqui em identificar os principais membros de uma comunidade que chamamos de núcleo da comunidade. Para determinar este núcleo, primeiramente definimos uma nova métrica chamada de *CoScore*, derivada do índice h. Como nossas análises são focadas na evolução temporal dessas comunidades, realizamos um estudo para estimar o índice h de um dado pesquisador ao longo do tempo. Utilizamos dados coletados pelo projeto SHINE e comparamos várias estratégias de evolução para este índice, demonstrando que a estratégia final escolhida possui forte correlação positiva com os respectivos valores do Google Citations, uma ferramenta largamente utilizada pela comunidade científica para estimar o índice h de um pesquisador. Assim, nossa métrica é capaz de capturar tanto a prolificidade de um pesquisador, quanto o seu envolvimento em uma comunidade ao longo do tempo.

Após quantificar a importância de cada membro de uma comunidade, definimos dois importantes limiares para nosso estudo, o tamanho do núcleo da comunidade e o tamanho da janela temporal, sendo estes 10% e 3 anos, respectivamente. Nossos resul-

tados indicam que membros dos núcleos são pesquisadores que contribuem ativamente com publicações para suas comunidades e possuem posição de destaque nelas. Desta forma, mostramos que vários desses membros foram premiados por suas contribuições à sua comunidade, incluindo até mesmo alguns que receberam o ACM *A.M. Turing Award*.

Nossas análises mostraram, ainda, que o coeficiente de agrupamento dos membros do núcleo tende, de forma geral, a ser menor do que os dos demais membros da rede. No entanto, o valor da métrica *betweenness* dos membros do núcleo tende a ser consideravelmente maior do que a dos não membros, indicando que membros dos núcleos atuam como pontes que interligam pequenos grupos de pesquisa. Além disso, observamos que os membros dos núcleos tendem a aumentar o grau médio, diâmetro, caminho mínimo médio e o tamanho do maior componente conectado da rede, e diminuir a assortatividade e o coeficiente de agrupamento. Mais importante, observamos que as variações observadas nos conjuntos de membros dos núcleos das comunidades estão fortemente correlacionadas com as variações nas propriedades topológicas da rede.

Finalmente, fornecemos uma representação visual dessas comunidades que reforça nossas análises e observações. Nossos resultados também destacam a importância de se estudar a relevância dos membros dos núcleos das comunidades e esperamos que nossas observações possam inspirar futuros modelos de formação de comunidade.

## 5.2 Trabalhos Futuros

A partir dos resultados do nosso estudo algumas oportunidades de trabalhos futuros foram identificadas e são listadas a seguir:

- **Aplicação do estudo em outros contextos.** Nossas análises do núcleo das comunidades são aplicáveis a outros contextos, como jogos multijogador massivo, OSNs e repositórios de outras naturezas, como filmes e livros.

- **Utilização de outras métricas de prolificidade.** Existem outras métricas capazes de medir a prolificidade de um pesquisador além do índice h que poderiam também serem utilizadas no cálculo do *CoScore*, como o índice g [Egghe, 2006].

- **Avaliação do *CoScore* em outros contextos.** Nossa métrica quantifica a importância dos membros das comunidades. Desta forma, também poderíamos utilizá-la em outros contextos, e.g., para predição de *links* e em sistemas de recomendação.

- **Geração de modelos de formação de comunidades.** O *CoScore* pode ser combinado a outras métricas para mapear o comportamento de como nodos e arestas surgem na rede, possibilitando a geração de modelos de formação de comunidades, conforme resultados prévios apresentados por Leskovec et al. [2005, 2008].

- **Aplicação da abordagem proposta para o estudo de *clusters*.** Vários trabalhos na literatura utilizam abordagens algorítmicas para identificação de *clusters* e de nodos importantes na topologia da rede. No entanto, essas abordagens possuem um custo computacional considerável. Assim sendo, seria interessante aplicar a nossa abordagem em estudos semelhantes.

- **Análise do impacto da migração de pesquisadores entre comunidades.** A migração dos membros do núcleo de uma comunidade pode ser utilizada para prever o sucesso ou declínio dessa comunidade. Sendo assim, nossa métrica poderia ser aplicada em estudos que visem caracterizar a migração de membros de um comunidade ou até inspirar modelos capazes de predizer tais migrações.

# Referências Bibliográficas


Acuna, D. E.; Allesina, S. & Kording, K. P. (2012). Future impact: Predicting scientific success. *Nature*, 489(7415):201–202.

Ahn, Y.-Y.; Han, S.; Kwak, H.; Moon, S. & Jeong, H. (2007). Analysis of Topological Characteristics of Huge Online Social Networking Services. In *Proceedings of the 16$^{th}$ International Conference on World Wide Web*, pp. 835–844, Banff, Alberta, Canada.

Alves, B. L.; Benevenuto, F. & Laender, A. H. (2013). The Role of Research Leaders on the Evolution of Scientific Communities. In *Proceedings of the 22$^{nd}$ International Conference on World Wide Web (Companion Volume)*, pp. 649–656, Rio de Janeiro, Brazil.

Backstrom, L.; Boldi, P.; Rosa, M.; Ugander, J. & Vigna, S. (2012). Four Degrees of Separation. In *Proceedings of the 3$^{rd}$ Annual ACM Web Science Conference*, pp. 33–42, Evanston, Illinois.

Backstrom, L.; Huttenlocher, D.; Kleinberg, J. & Lan, X. (2006). Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *Proceedings of the 12$^{th}$ ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 44–54, Philadelphia, PA, USA.

Barabási, A.-L. & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512.

Benevenuto, F.; Almeida, J. & Silva, A. (2012). Coleta e Análise de Grandes Bases de Dados de Redes Sociais Online. In *Jornada de Atualizações em Informática 2012*.

Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M. & Hwang, D. (2006). Complex Networks: Structure and Dynamics. *Physics Reports*, 424(4-5):175–308.

Braitenberg, V. & Schüz, A. (1998). *Cortex: Statistics and Geometry of Neuronal Connectivity*. Springer.

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A. & Wiener, J. (2000). Graph Structure in the Web. *Computer Networks*, 33(1-6):309–320.

Cha, M.; Haddadi, H.; Benevenuto, F. & Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. In *In Proceedings of the 4$^{th}$ International AAAI Conference on Weblogs and Social Media*, Washington DC, USA.

Chakrabarti, D.; Kumar, R. & Tomkins, A. (2006). Evolutionary Clustering. In *Proceedings of the 12$^{th}$ ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 554–560, Philadelphia, PA, USA.

Ducheneaut, N.; Yee, N.; Nickell, E. & Moore, R. J. (2007). The Life and Death of Online Gaming Communities: A Look at Guilds in World of Warcraft. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 839–848, San Jose, California, USA.

Dunbar, R. I. M. (1992). Neocortex Size as a Constraint on Group Size in Primates. *Journal of Human Evolution*, 22(6):469–493.

Easley, D. & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

Ebel, H.; Mielsch, L.-I. & Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Physical Review E*, 66(3).

Egghe, L. (2006). An Improvement of the H-index: The G-index. *ISSI Newsletter*, pp. 8–9.

Euler, L. (1956). *The Seven Bridges of Konigsberg*. Wm. Benton.

Faloutsos, M.; Faloutsos, P. & Faloutsos, C. (1999). On Power-Law Relationships of the Internet Topology. In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pp. 251–262, Cambridge, Massachusetts, USA.

Figueiredo, D. R. (2011). Introdução a Redes Complexas. In *Jornada de Atualizações em Informática 2011*, capítulo 7, pp. 303–358.

Getoor, L. & Diehl, C. P. (2005). Link Mining: A Survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12.

Gonçalves, B.; Perra, N. & Vespignani, A. (2011). Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number. *PLoS ONE*, 6(8).

Hirsch, J. E. (2005). An Index to Quantify an Individual's Scientific Research Output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572.

Hopcroft, J.; Khan, O.; Kulis, B. & Selman, B. (2004). Tracking Evolving Communities in Large Linked Networks. *Proceedings of the National Academy of Sciences*, 101:5249–5253.

Huang, J.; Zhuang, Z.; Li, J. & Giles, C. L. (2008). Collaboration Over Time: Characterizing and Modeling Network Evolution. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 107–116, Palo Alto, California, USA.

Kempe, D.; Kleinberg, J. & Tardos, E. (2003). Maximizing the Spread of Influence through a Social Network. In *Proceedings of the 9$^{th}$ ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146, Washington, D.C.

Kempe, D.; Kleinberg, J. & Tardos, E. (2005). Influential Nodes in a Diffusion Model for Social Networks. In *Proceedings of the 32$^{nd}$ International Conference on Automata, Languages and Programming*, pp. 1127–1138, Lisbon, Portugal.

Kleinberg, J. (2008). The Convergence of Social and Technological Networks. *Communications of the ACM*, 51(11):66–72.

Kumar, R.; Novak, J. & Tomkins, A. (2006). Structure and Evolution of Online Social Networks. In *Proceedings of the 12$^{th}$ ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 611–617, Philadelphia, PA, USA.

Laender, A. H. F.; de Lucena, C. J. P.; Maldonado, J. C.; de Souza e Silva, E. & Ziviani, N. (2008). Assessing the Research and Education Quality of the Top Brazilian Computer Science Graduate Programs. *SIGCSE Bull.*, 40(2):135–145.

Leskovec, J.; Backstrom, L.; Kumar, R. & Tomkins, A. (2008). Microscopic Evolution of Social Networks. In *Proceedings of the 14$^{th}$ ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 462–470, Las Vegas, Nevada, USA.

Leskovec, J.; Kleinberg, J. & Faloutsos, C. (2005). Graphs Over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In *Proceedings of the 11$^{th}$ ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 177–187, Chicago, Illinois, USA.

Leskovec, J.; Lang, K. J. & Mahoney, M. (2010). Empirical Comparison of Algorithms for Network Community Detection. In *Proceedings of the 19$^{th}$ International Conference on World Wide Web*, pp. 631–640, Raleigh, North Carolina, USA.

Ley, M. (2009). Dblp: Some lessons learned. *Proceedings of the VLDB Endowment*, 2(2):1493–1500.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P. & Bhattacharjee, B. (2007). Measurement and Analysis of Online Social Networks. In *Proceedings of the 7$^{th}$ ACM SIGCOMM Conference on Internet Measurement*, pp. 29–42, San Diego, California, USA.

Moreno, J. L. (1953,1978). *Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama.* Beacon House.

Newman, M. & Park, J. (2003). Why Social Networks are Different from Other Types of Networks. *Physical Review E*, 68.

Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, 45:167–256.

Patil, A.; Liu, J.; Price, B.; Sharara, H. & Brdiczka, O. (2012). Modeling Destructive Group Dynamics in On-Line Gaming Communities. In *Proceedings of the 6$^{th}$ International AAAI Conference on Weblogs and Social Media.*

Rogers, E. M. (1962). *Diffusion of Innovations.* Free Press.

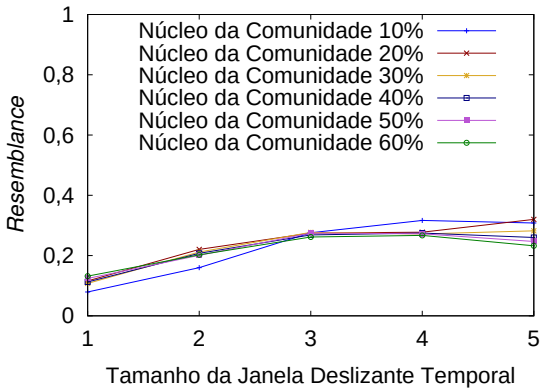
Seifi, M. & Guillaume, J.-L. (2012). Community Cores in Evolving Networks. In *Proceedings of the 21$^{st}$ International Conference on World Wide Web (Companion Volume)*, pp. 1173–1180, Lyon, France.

Viswanath, B.; Mislove, A.; Cha, M. & Gummadi, K. P. (2009). On the Evolution of User Interaction in Facebook. In *Proceedings of the 2$^{nd}$ ACM Workshop on Online Social Networks*, pp. 37–42, Barcelona, Spain.

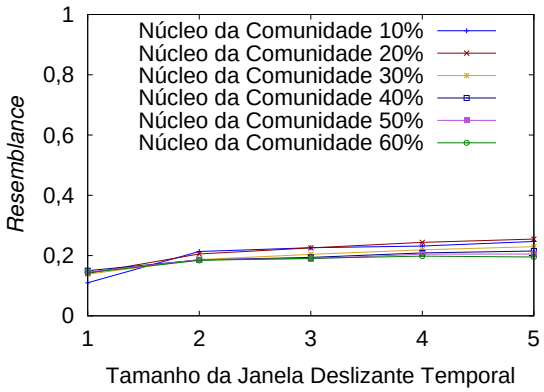
Watts, D. & Dodds, P. (2007). Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research*, 34(4):441–458.

# Apêndice A

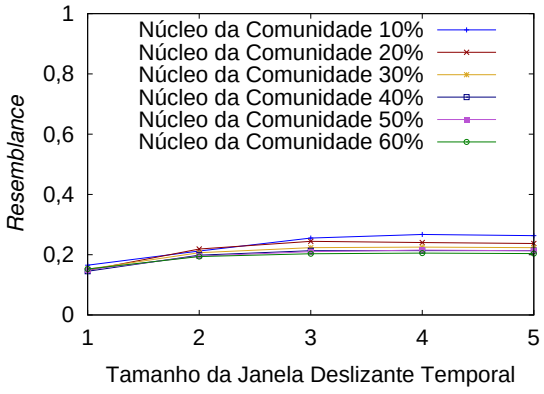
# Média dos Valores de *Resemblance*



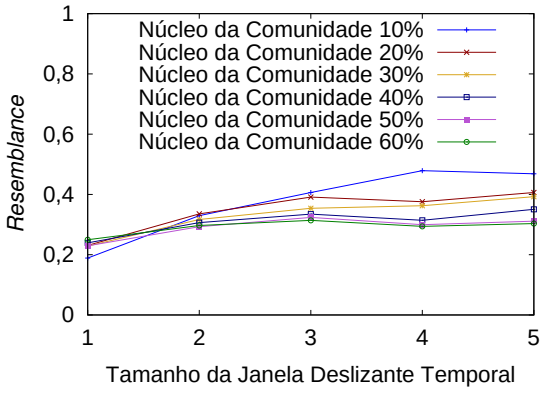

(a) CCS



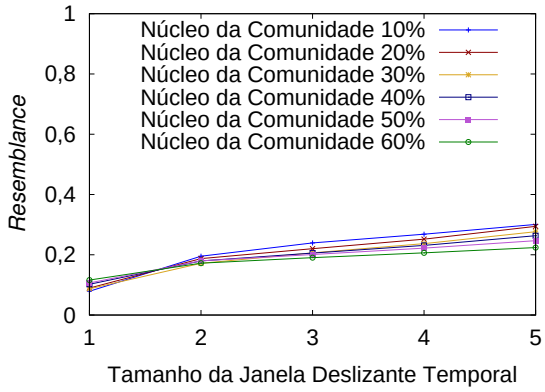

(b) CIKM



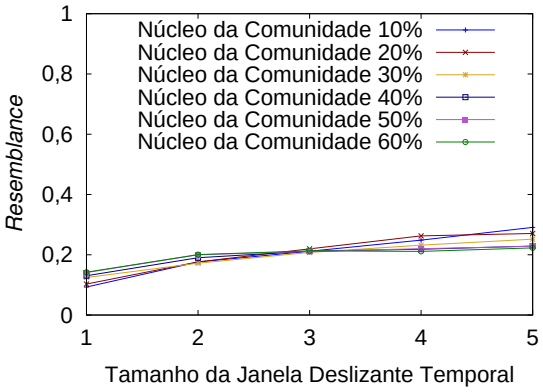

(c) DAC



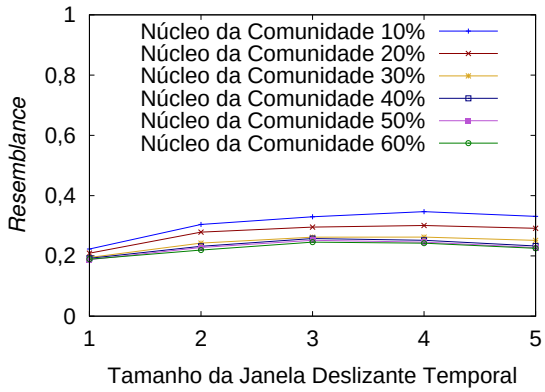

(d) HSCC



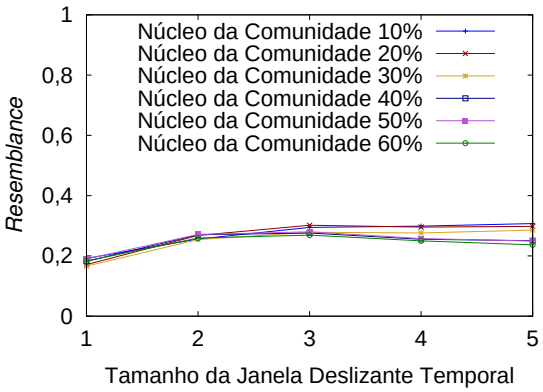

(e) ICSE



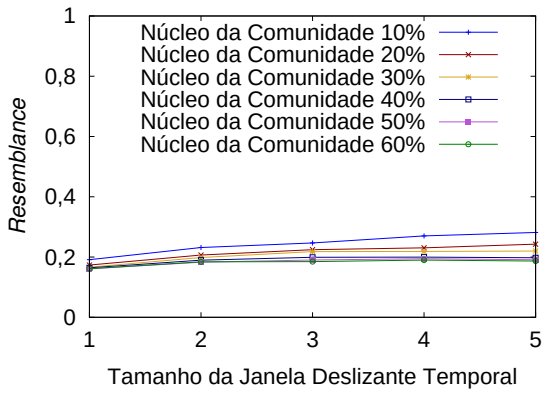

(f) ISCA



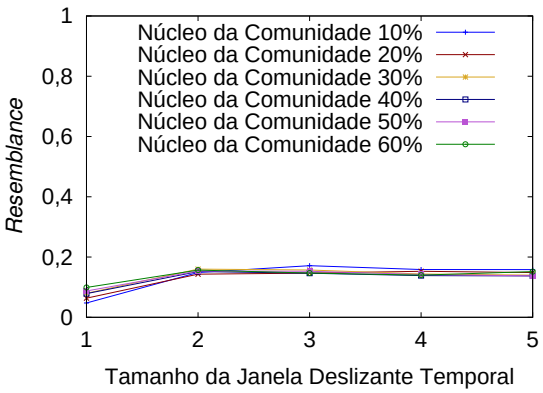

(g) KDD



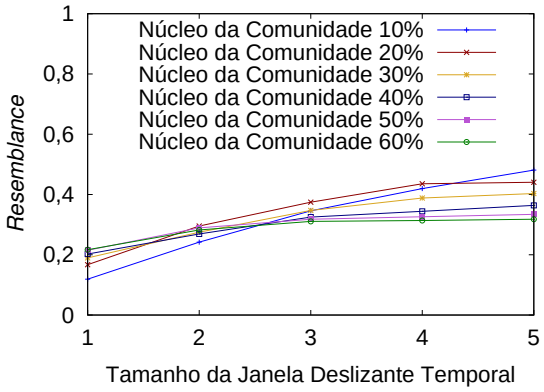

(h) MICRO



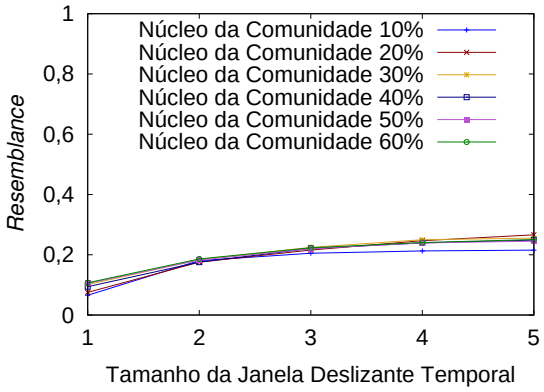

(i) MM



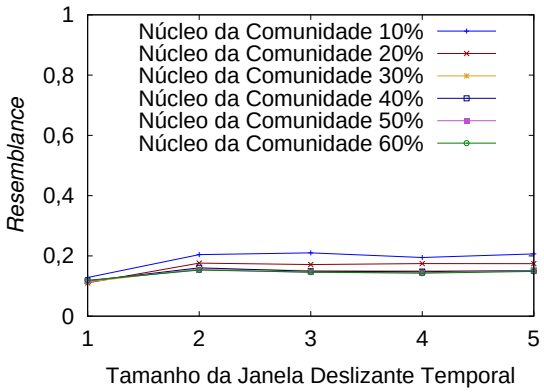

(j) MOBICOM



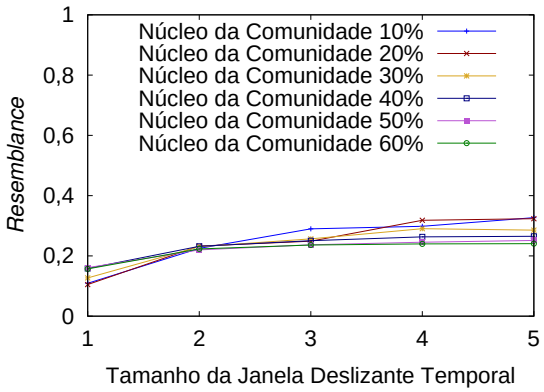

(k) PODC



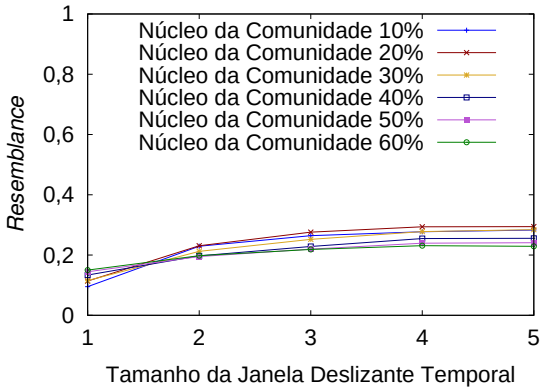

(l) POPL



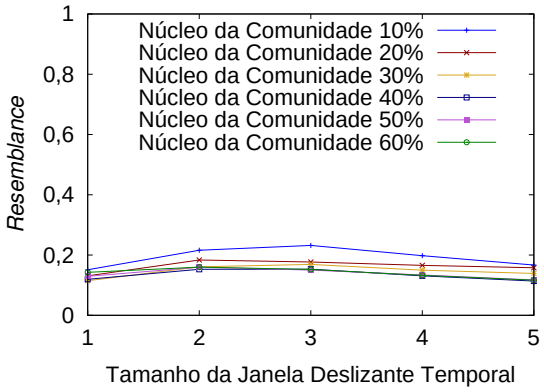

(m) SAC



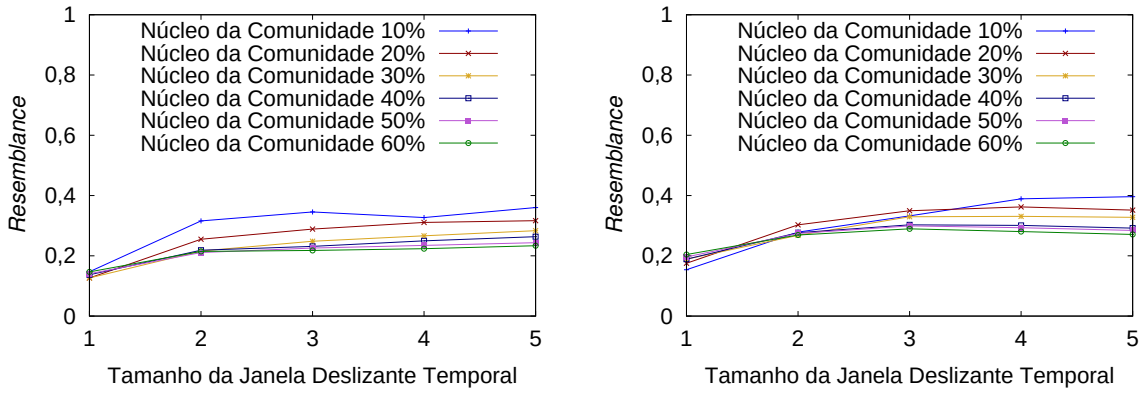

(n) SIGCOMM



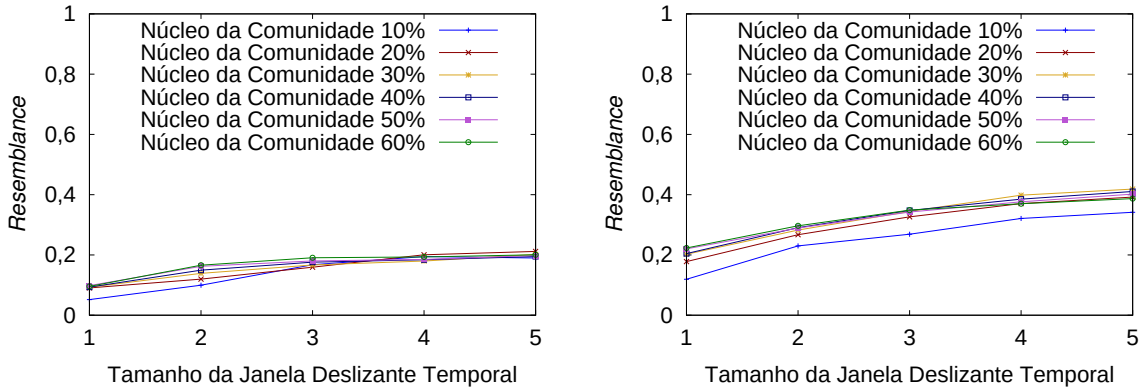

(o) SIGCSE




(p) SIGDOC




(q) SIGGRAPH (r) SIGIR



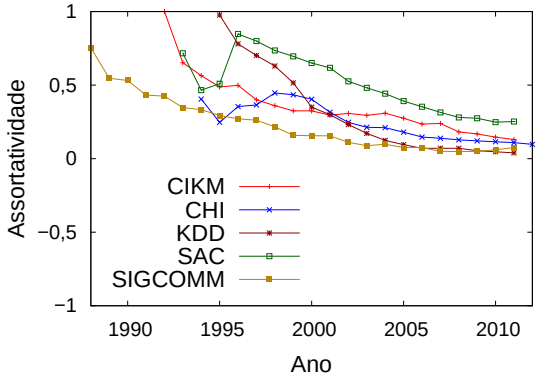

(s) SIGMETRICS (t) STOC

Figura A.-1: Média dos valores de *resemblance*

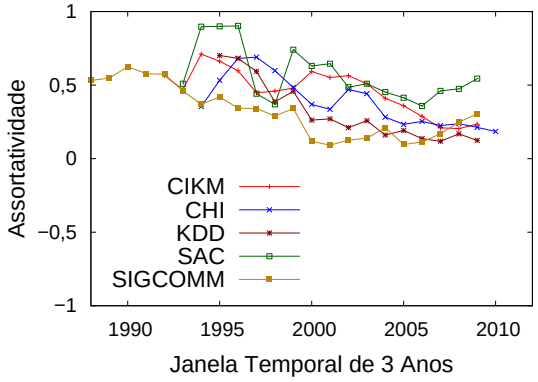
# Apêndice B

# Métricas de Evolução das Comunidades Científicas

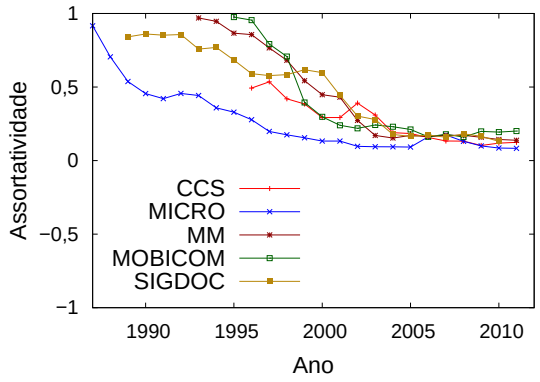
Neste apêndice apresentamos as métricas das demais conferências, separada entre três grupos. O grupo A é constituído pelas conferências CIKM, CHI, KDD, SAC e SIGCOMM, o grupo B pelas conferências CSS, MICRO, MM, MOBICOM e SIGDOC, e o grupo C pelas conferências HSCC, ICSE, ISCA, SIGCSE, SIGGRAPH e SIGMETRICS.



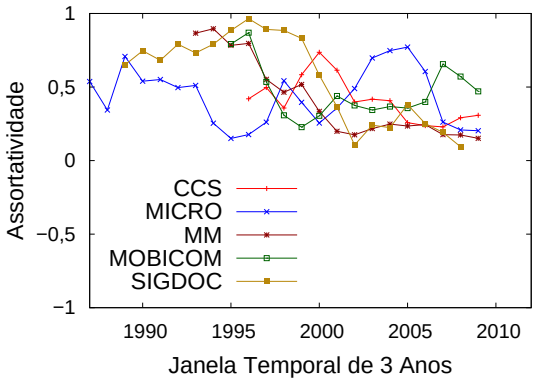

(a) Assortatividade final - Grupo A



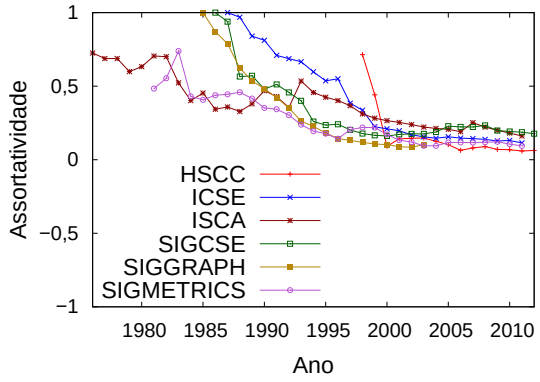

(b) Assortatividade por janela - Grupo A



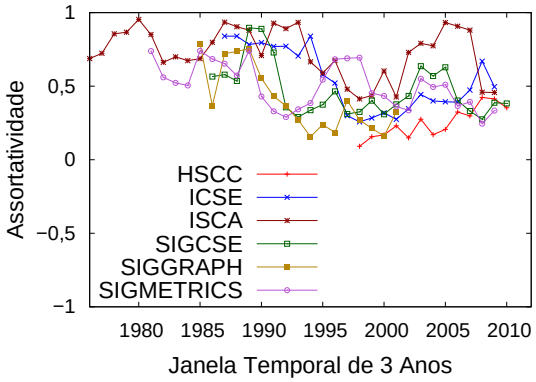

(c) Assortatividade final - Grupo B




(d) Assortatividade por janela - Grupo B

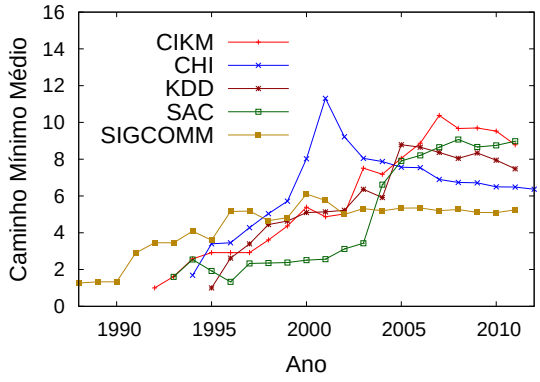



(e) Assortatividade final - Grupo C

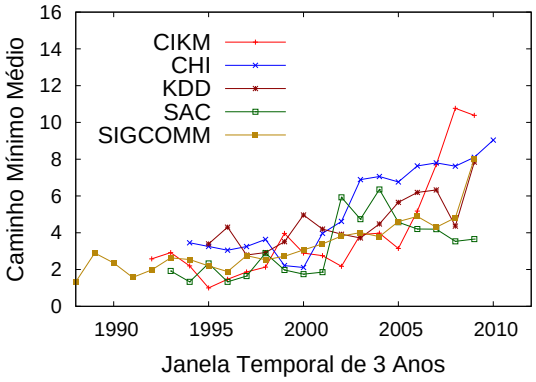



(f) Assortatividade por janela - Grupo C

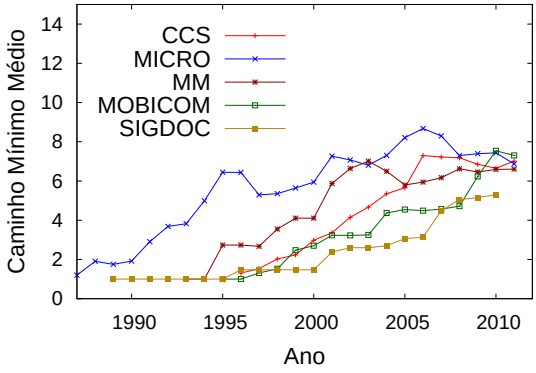
Figura B.1: Assortatividade das comunidades científicas



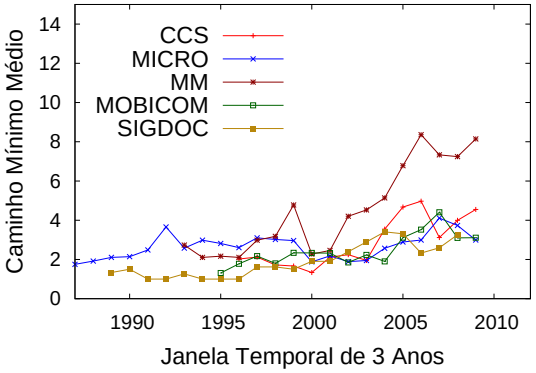

(a) CMM final - Grupo A



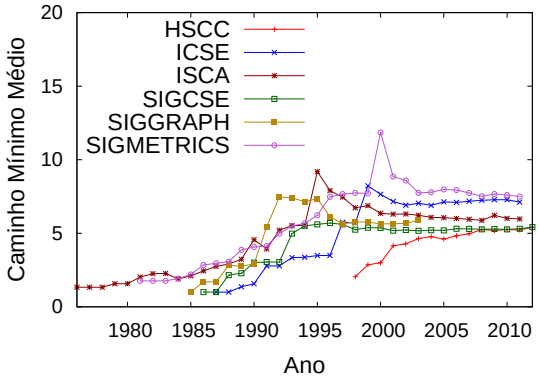

(b) CMM por janela - Grupo A



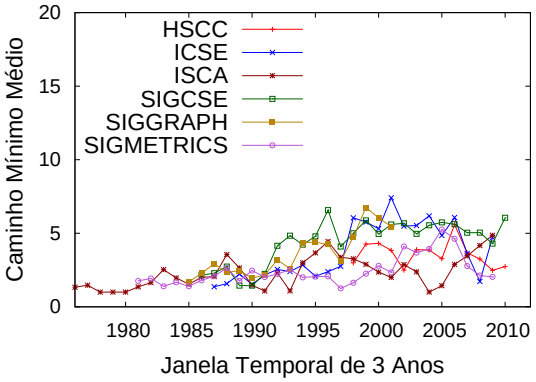

(c) CMM final - Grupo B




(d) CMM por janela - Grupo B

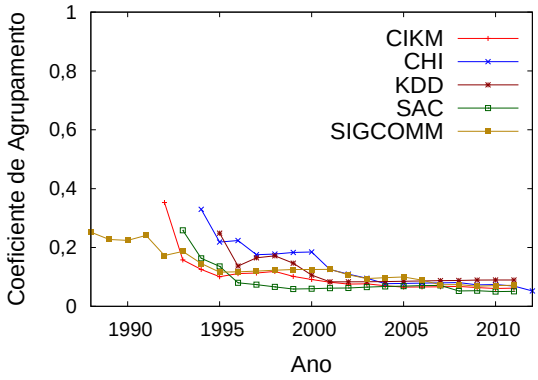



(e) CMM final - Grupo C

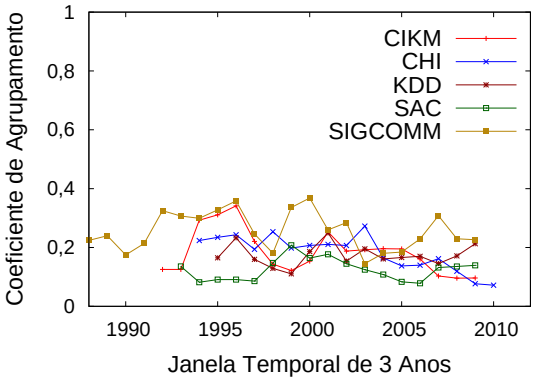



(f) CMM por janela - Grupo C

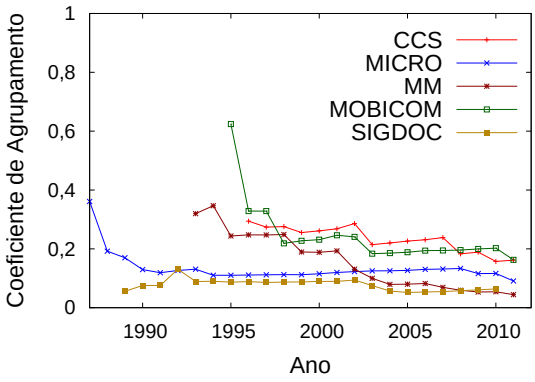
Figura B.2: Caminho mínimo médio das comunidades científicas



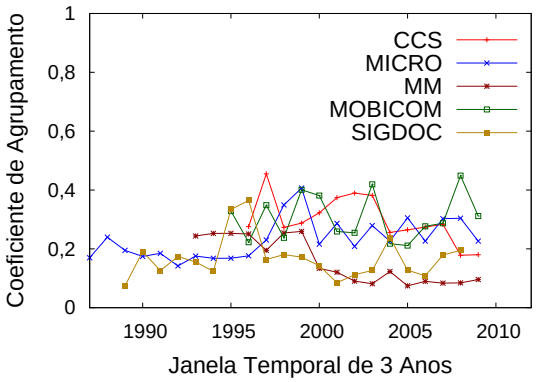

(a) CA final - Grupo A



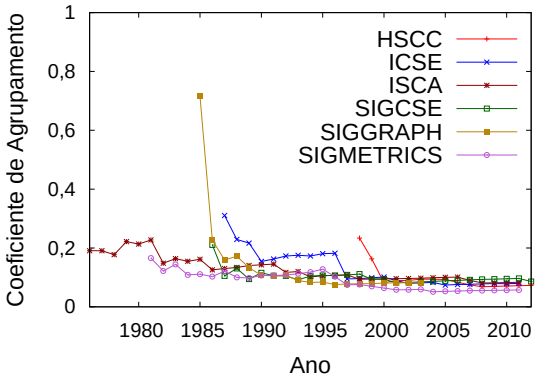

(b) CA por janela - Grupo A



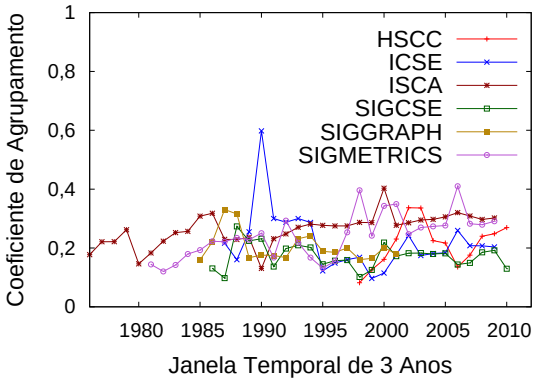

(c) CA final - Grupo B




(d) CA por janela - Grupo B

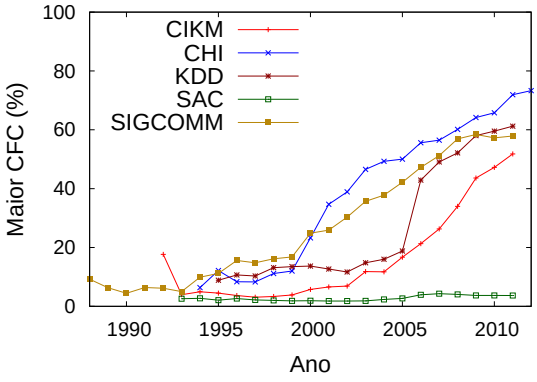



(e) CA final - Grupo C

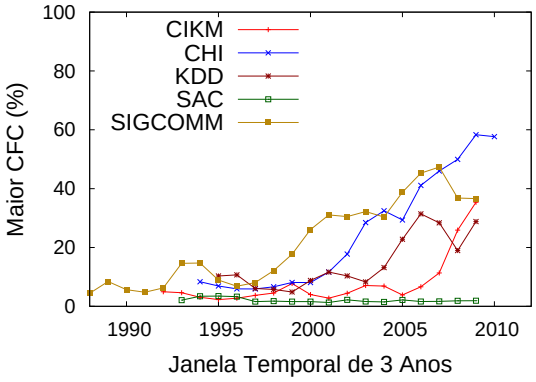



(f) CA por janela - Grupo C

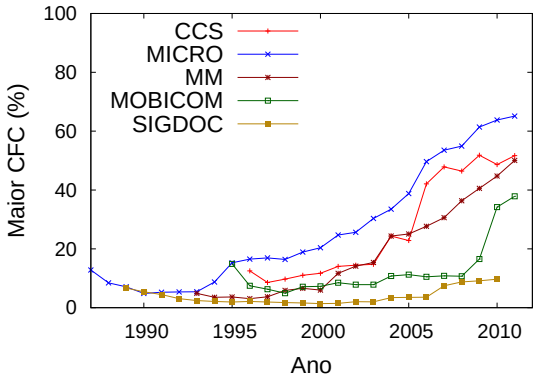
Figura B.3: Coeficiente de agrupamento das comunidades científicas



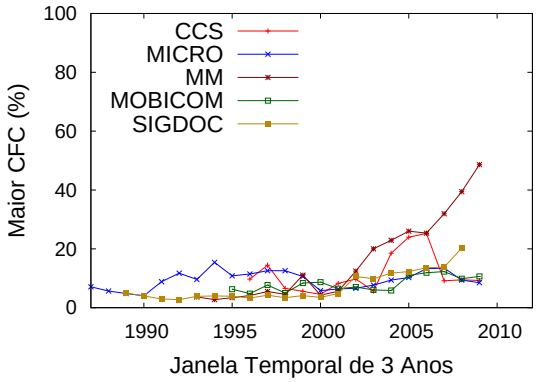

(a) Maior CFC final - Grupo A



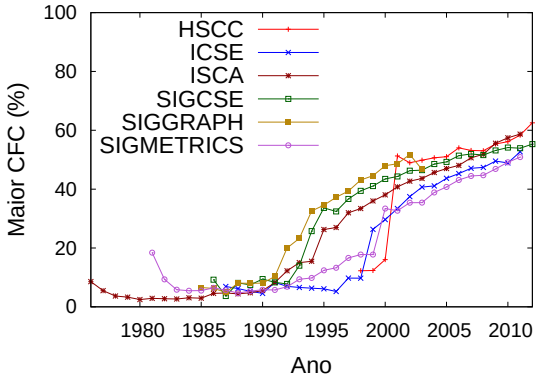

(b) Maior CFC por janela - Grupo A



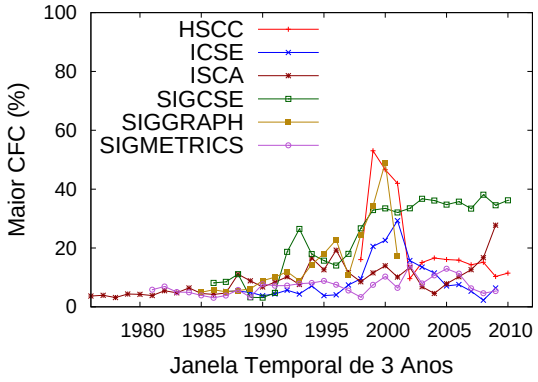

(c) Maior CFC final - Grupo B




(d) Maior CFC por janela - Grupo B

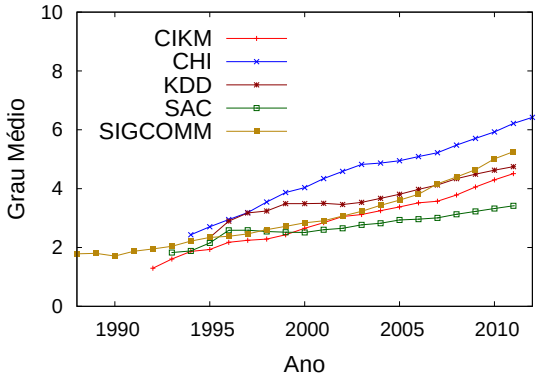



(e) Maior CFC final - Grupo C

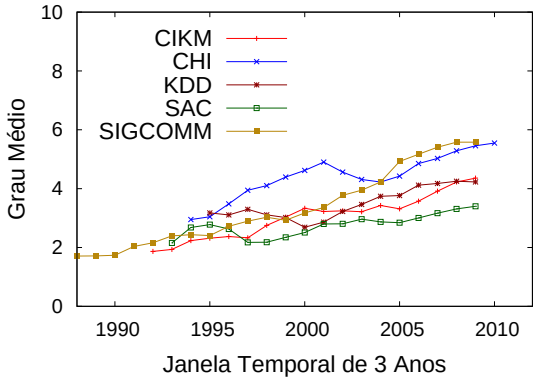



(f) Maior CFC por janela - Grupo C

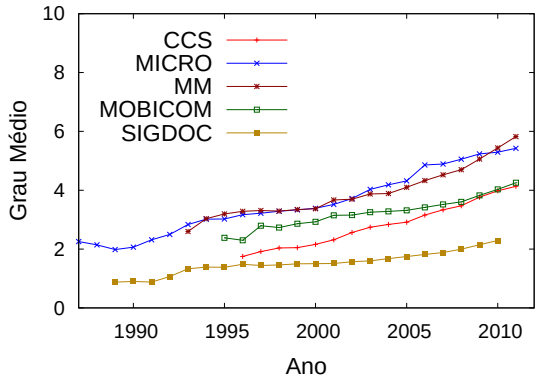
Figura B.4: Maior CFC das comunidades científicas



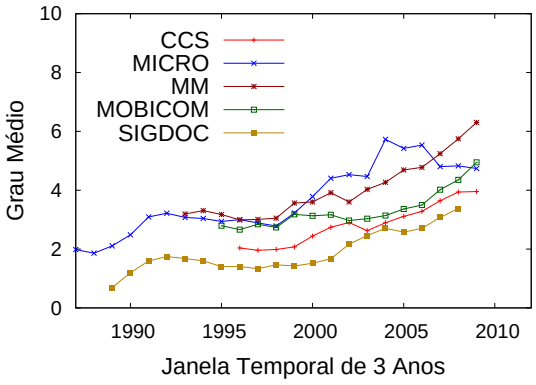

(a) Grau médio final - Grupo A



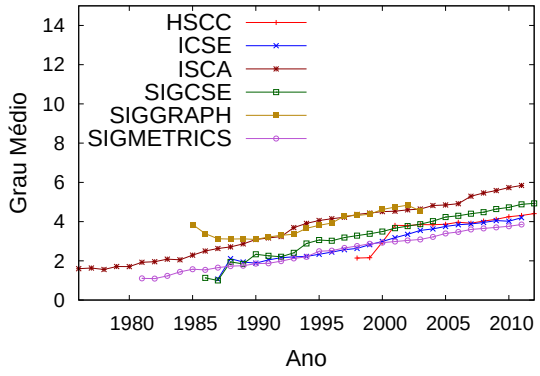

(b) Grau médio por janela - Grupo A



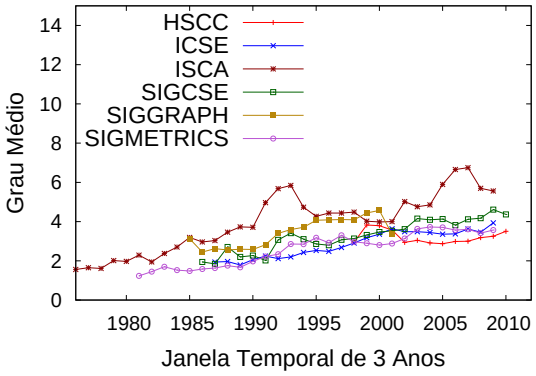

(c) Grau médio final - Grupo B




(d) Grau médio por janela - Grupo B




(e) Grau médio final - Grupo C

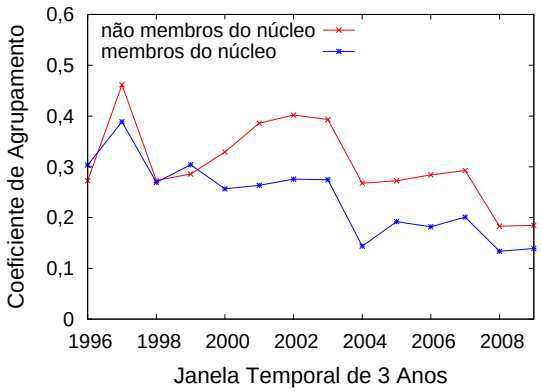



(f) Grau médio por janela - Grupo C

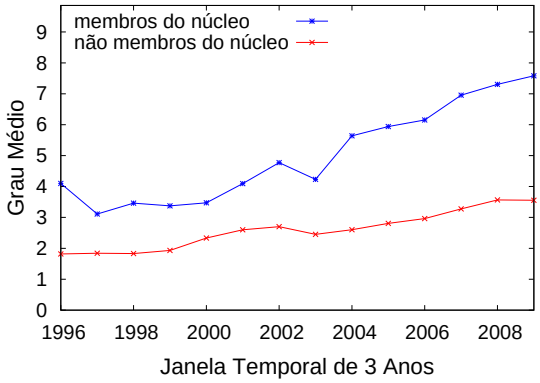
Figura B.5: Grau médio das comunidades científicas

# Apêndice C

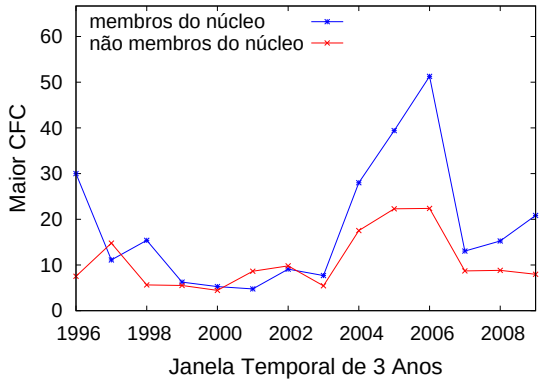
# Comparação entre Membros e Não Membros



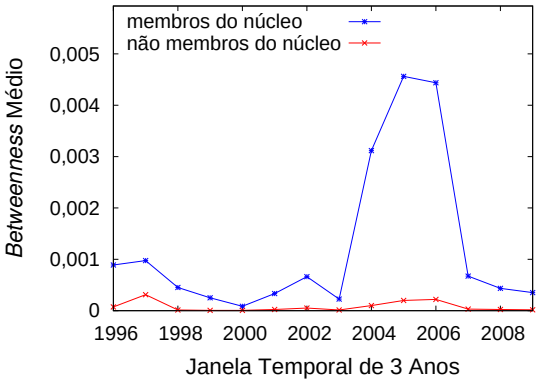

(a) Coeficiente de agrupamento




(b) Grau médio

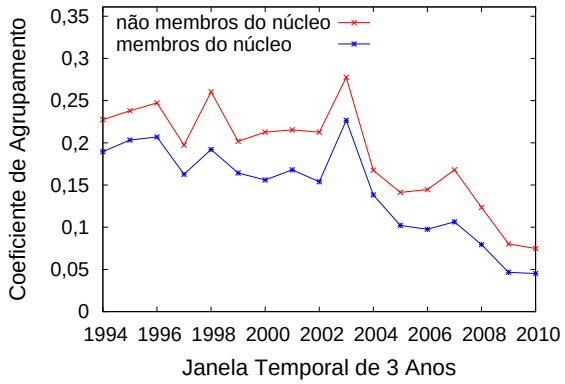



(c) Maior CFC

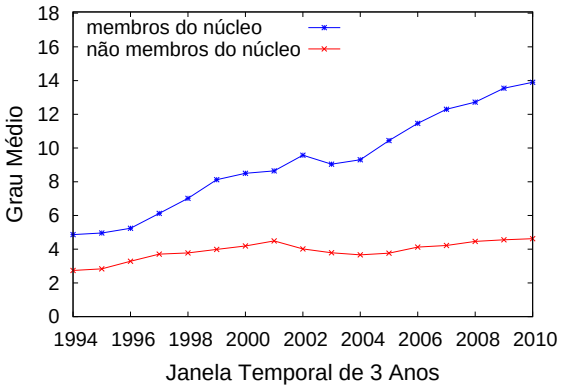



(d) *Betweenness* médio

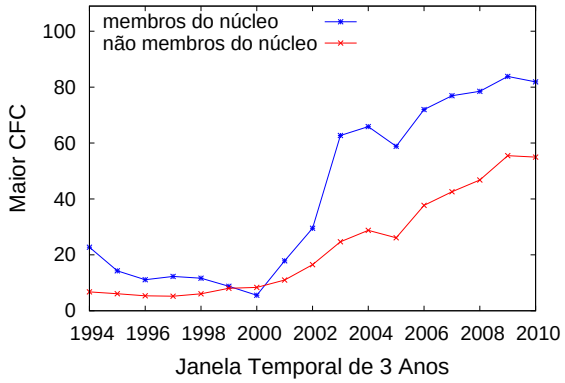
Figura C.1: Propriedades da comunidade CCS para os membros e não membros do núcleo



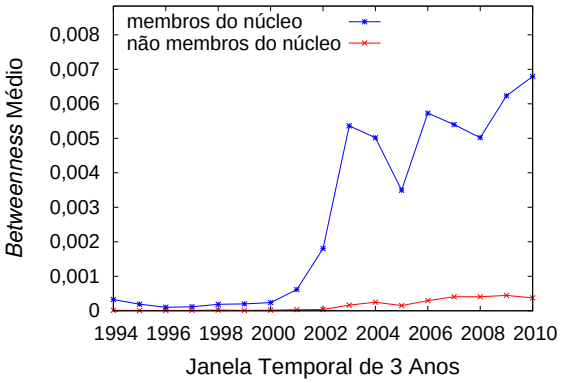

(a) Coeficiente de agrupamento




(b) Grau médio

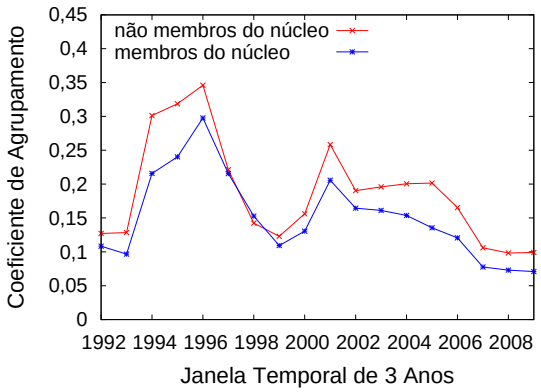



(c) Maior CFC

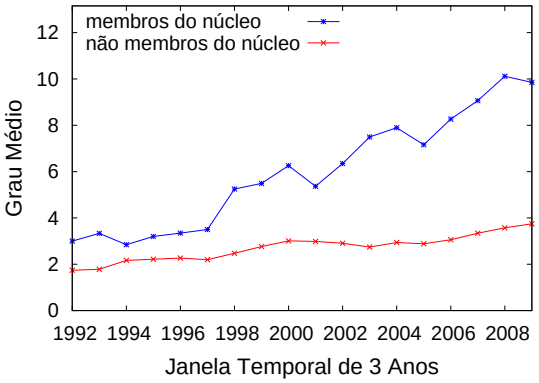



(d) *Betweenness* médio

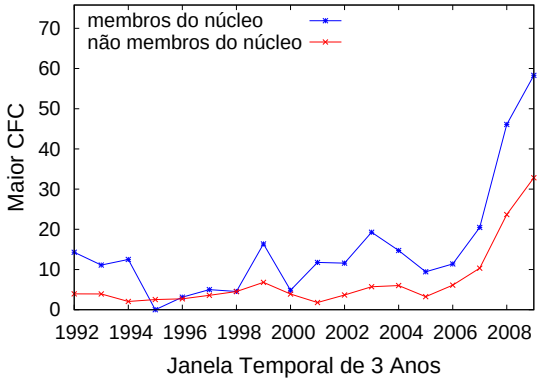
Figura C.2: Propriedades da comunidade CHI para os membros e não membros do núcleo



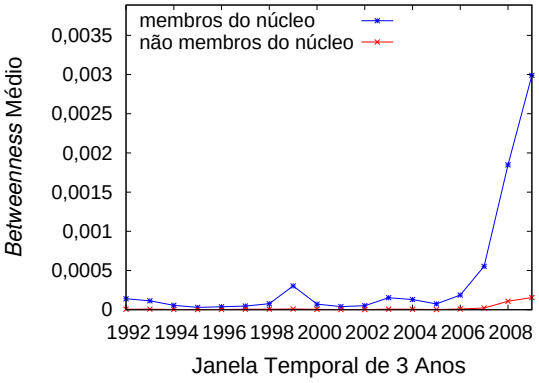

(a) Coeficiente de agrupamento




(b) Grau médio

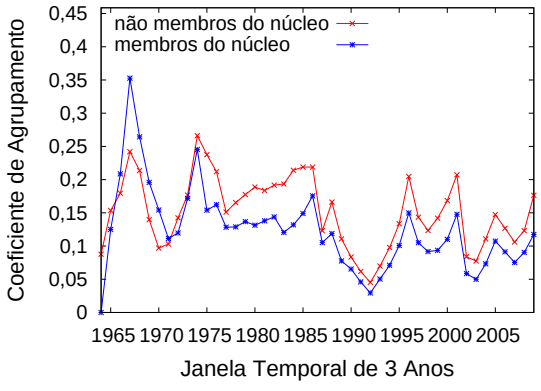



(c) Maior CFC

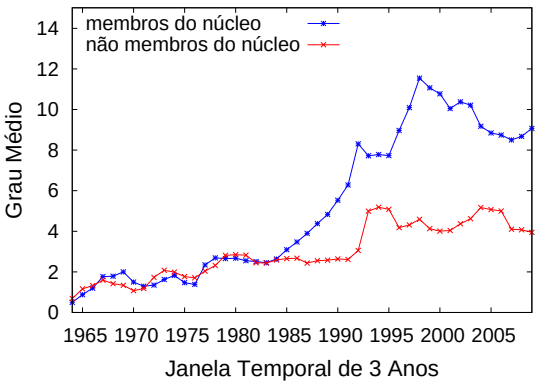



(d) *Betweenness* médio

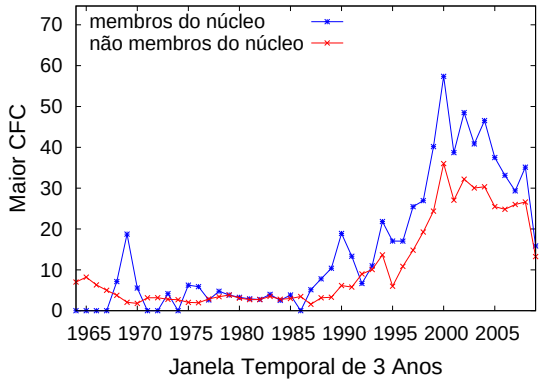
Figura C.3: Propriedades da comunidade CIKM para os membros e não membros do núcleo



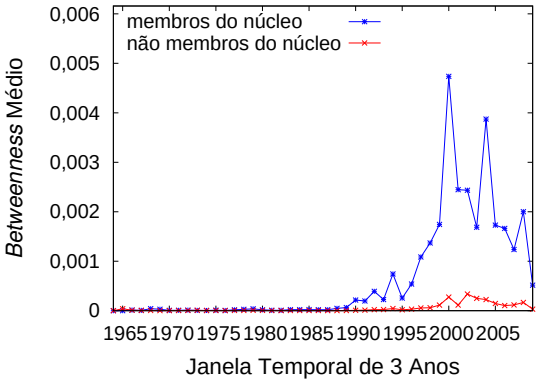

(a) Coeficiente de agrupamento




(b) Grau médio

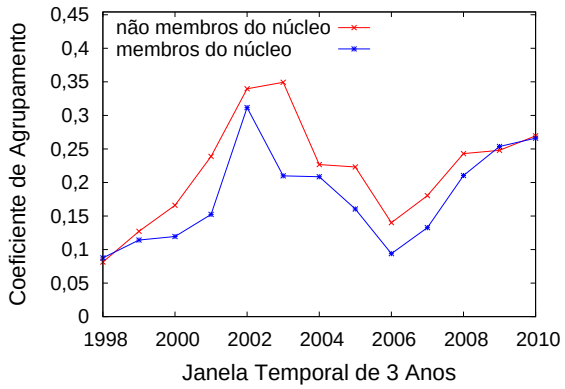



(c) Maior CFC

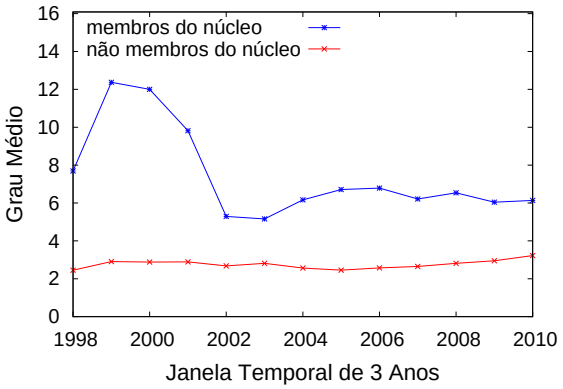



(d) *Betweenness* médio

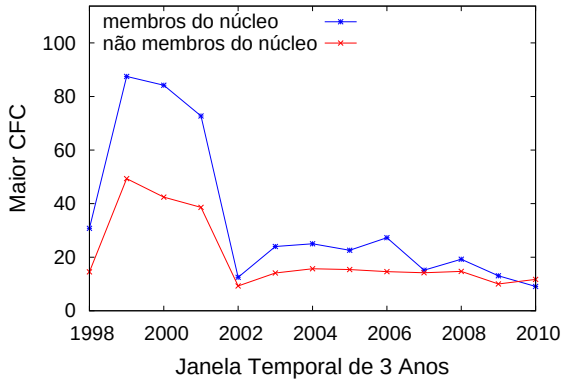
Figura C.4: Propriedades da comunidade DAC para os membros e não membros do núcleo



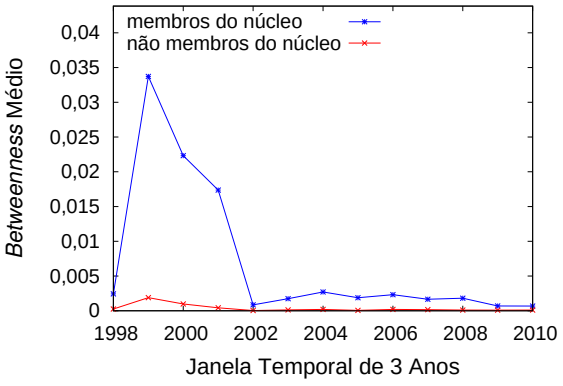

(a) Coeficiente de agrupamento




(b) Grau médio

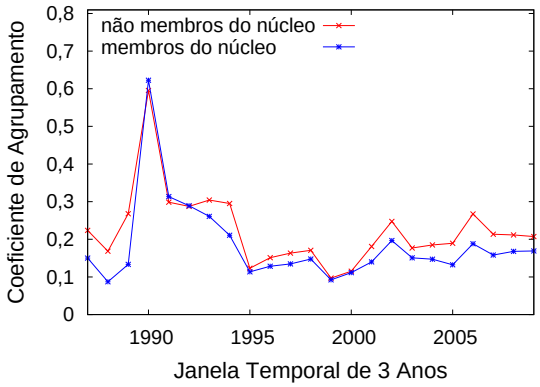



(c) Maior CFC

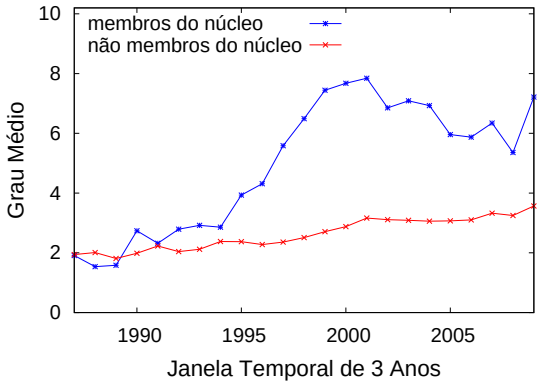



(d) *Betweenness* médio

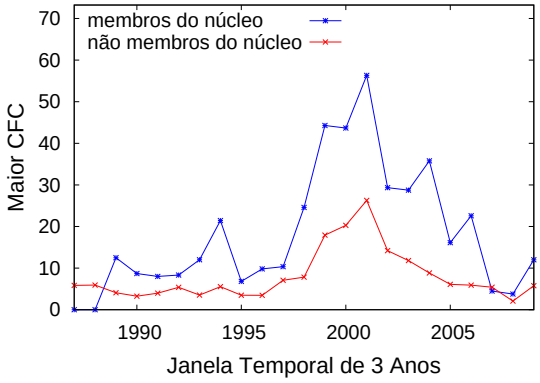
Figura C.5: Propriedades da comunidade HSCC para os membros e não membros do núcleo



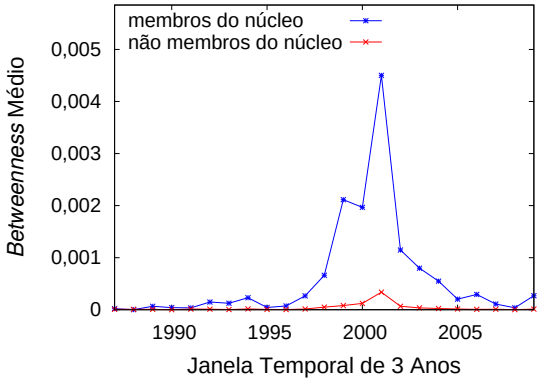

(a) Coeficiente de agrupamento




(b) Grau médio

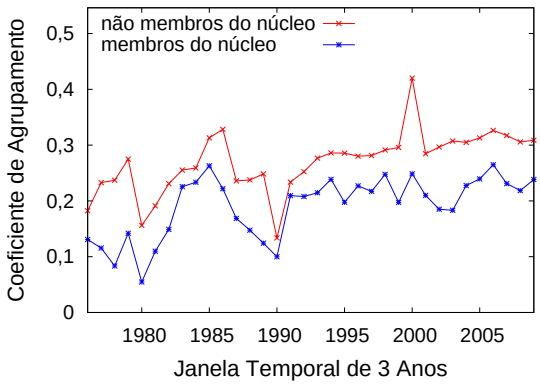



(c) Maior CFC

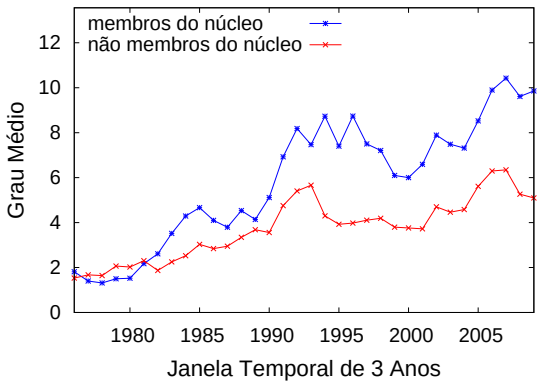



(d) *Betweenness* médio

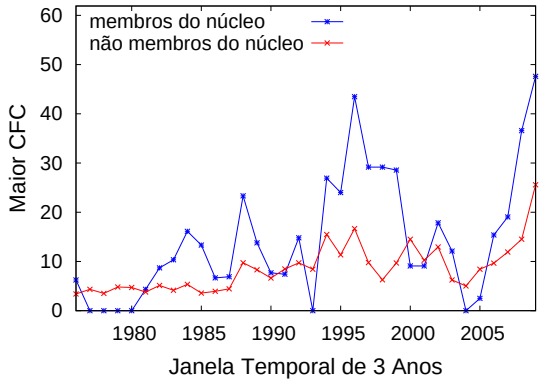
Figura C.6: Propriedades da comunidade ICSE para os membros e não membros do núcleo



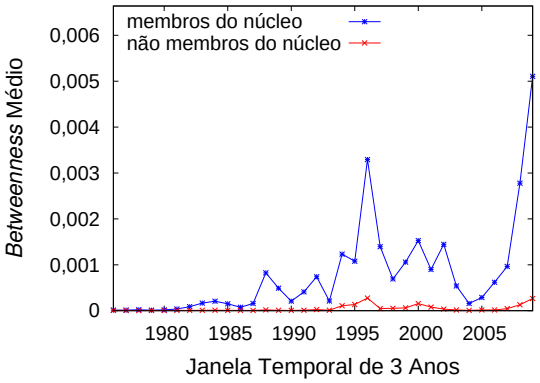

(a) Coeficiente de agrupamento




(b) Grau médio

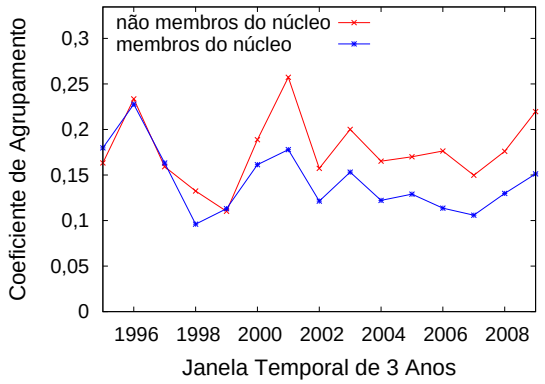



(c) Maior CFC

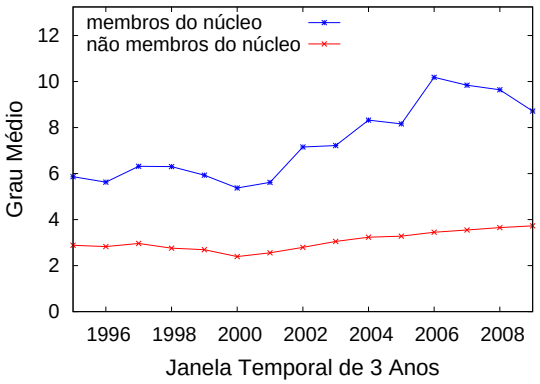



(d) *Betweenness* médio

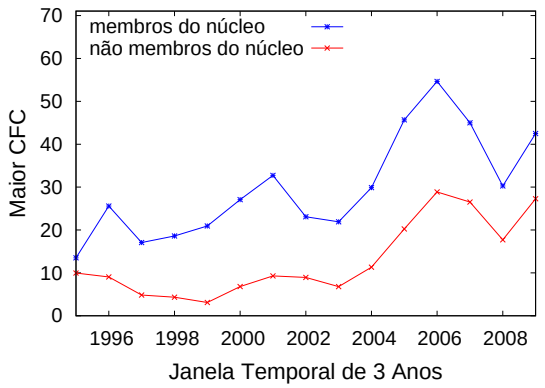
Figura C.7: Propriedades da comunidade ISCA para os membros e não membros do núcleo



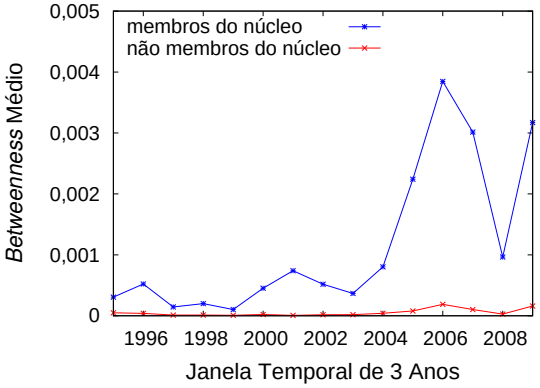

(a) Coeficiente de agrupamento




(b) Grau médio

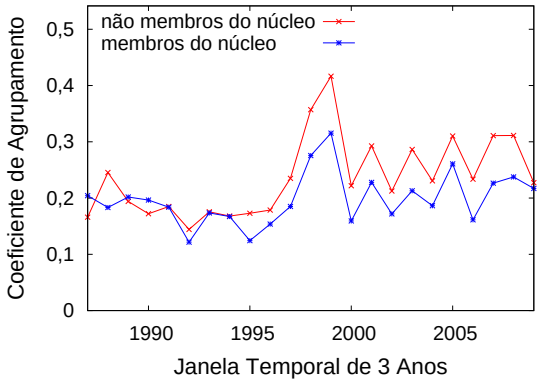



(c) Maior CFC

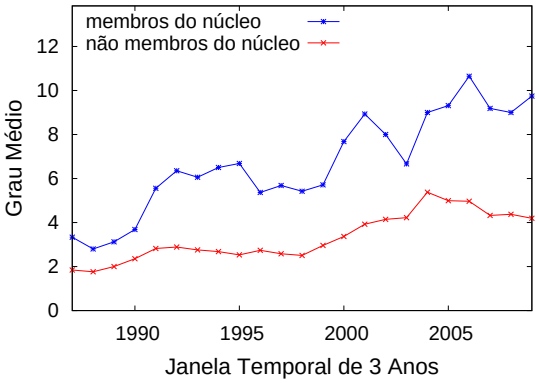



(d) *Betweenness* médio

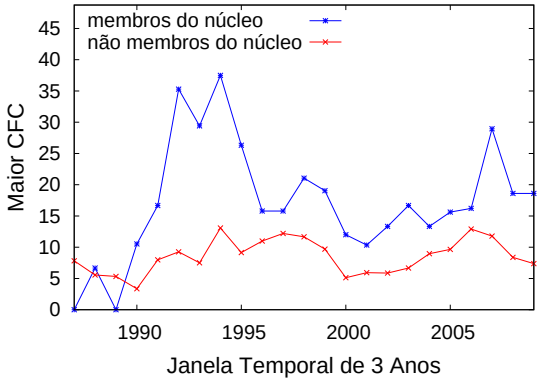
Figura C.8: Propriedades da comunidade KDD para os membros e não membros do núcleo



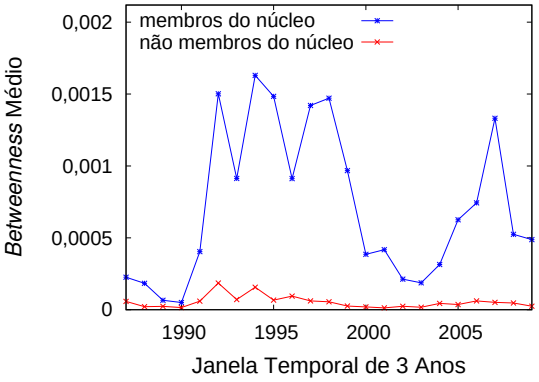

(a) Coeficiente de agrupamento




(b) Grau médio

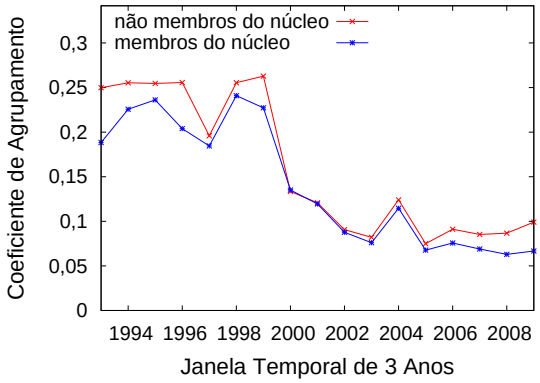



(c) Maior CFC

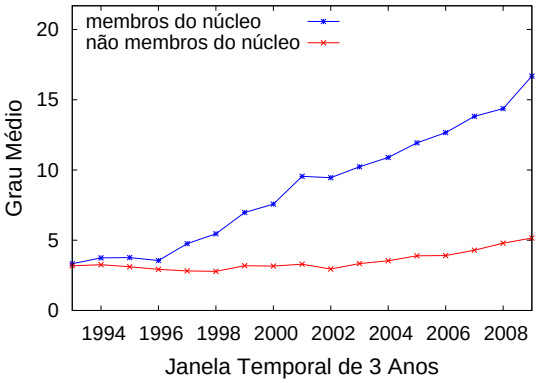



(d) *Betweenness* médio

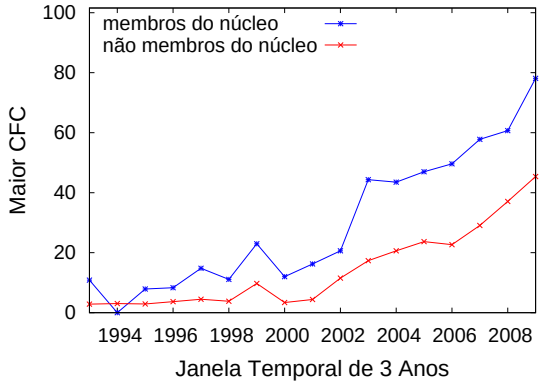
Figura C.9: Propriedades da comunidade MICRO para os membros e não membros do núcleo



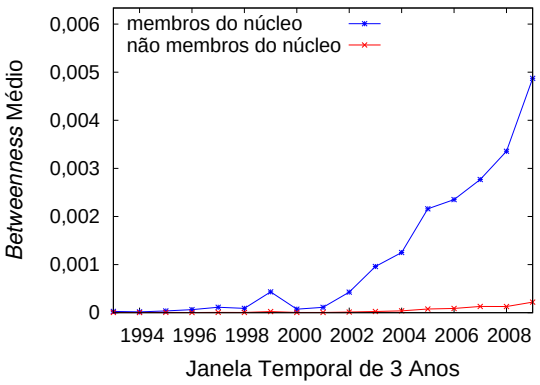

(a) Coeficiente de agrupamento




(b) Grau médio

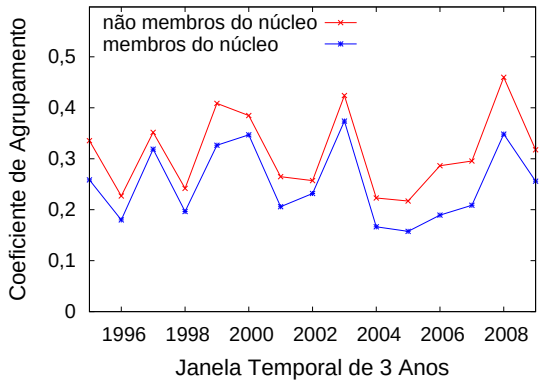



(c) Maior CFC

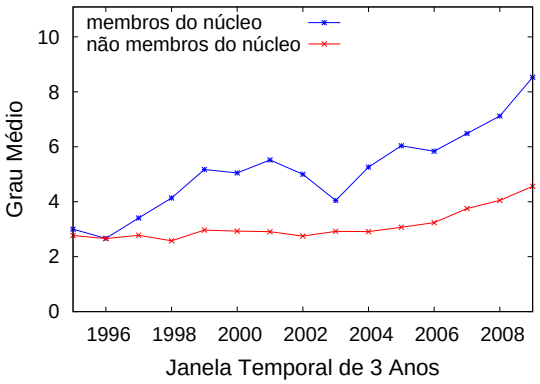



(d) *Betweenness* médio

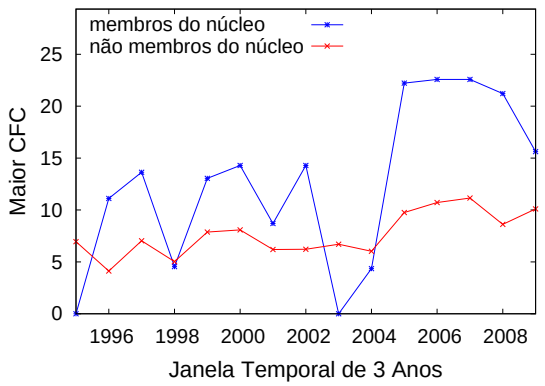
Figura C.10: Propriedades da comunidade MM para os membros e não membros do núcleo



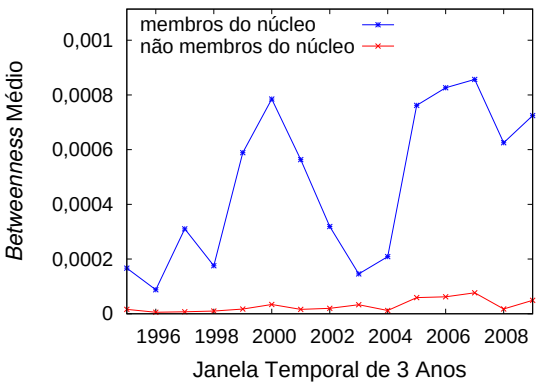

(a) Coeficiente de agrupamento




(b) Grau médio

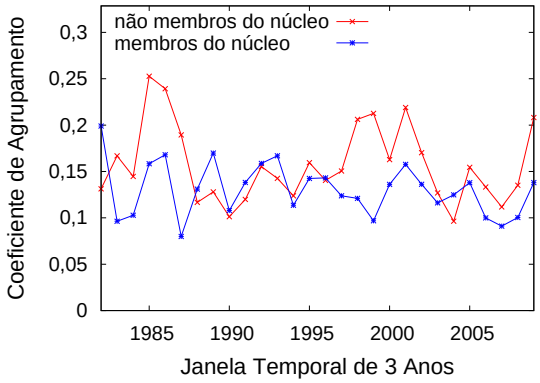



(c) Maior CFC

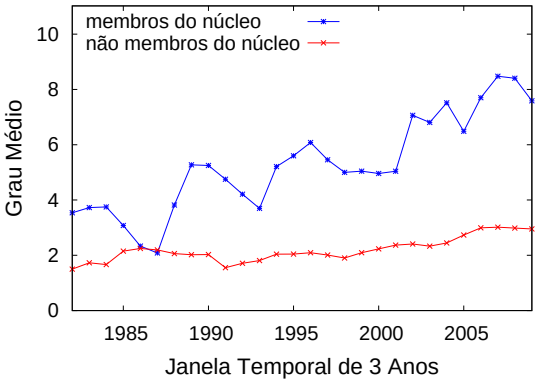



(d) *Betweenness* médio

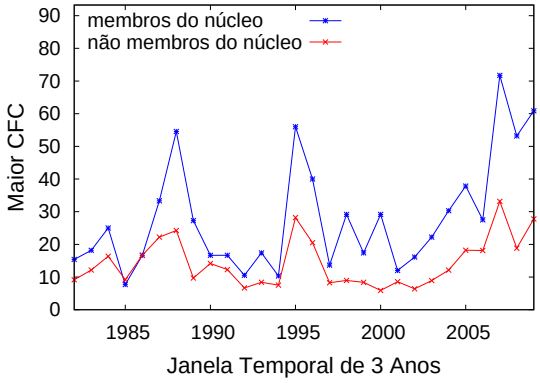
Figura C.11: Propriedades da comunidade MOBICOM para os membros e não membros do núcleo



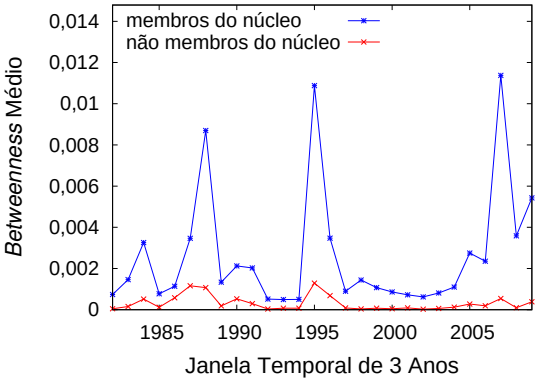

(a) Coeficiente de agrupamento




(b) Grau médio

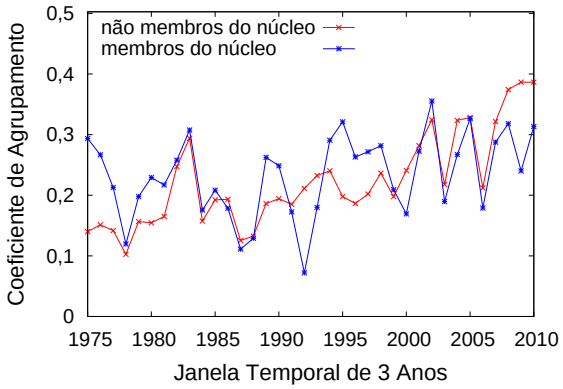



(c) Maior CFC

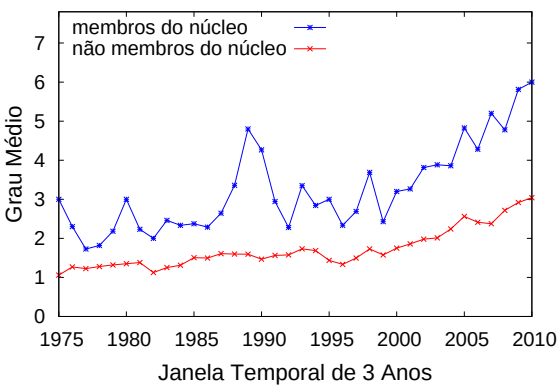



(d) *Betweenness* médio

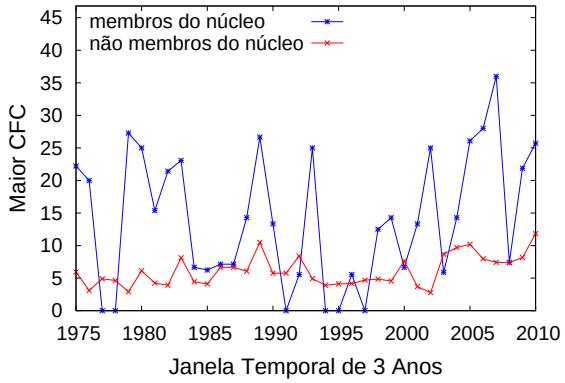
Figura C.12: Propriedades da comunidade PODC para os membros e não membros do núcleo



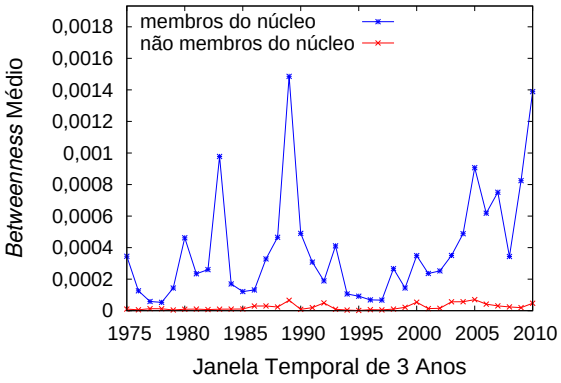

(a) Coeficiente de agrupamento




(b) Grau médio

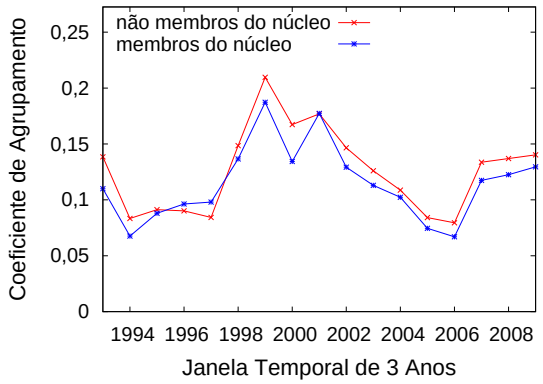



(c) Maior CFC

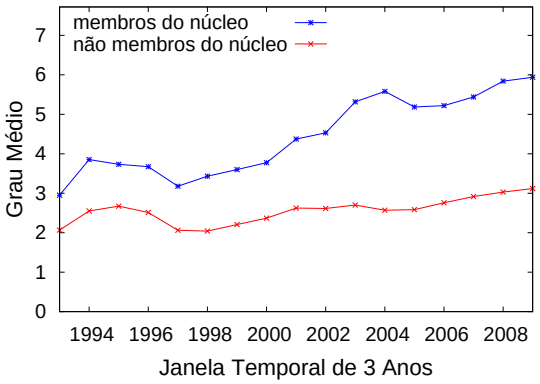



(d) *Betweenness* médio

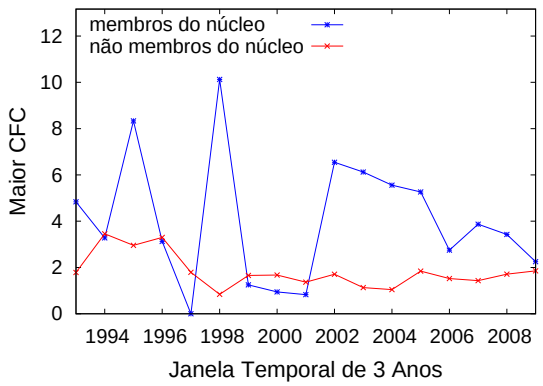
Figura C.13: Propriedades da comunidade POPL para os membros e não membros do núcleo



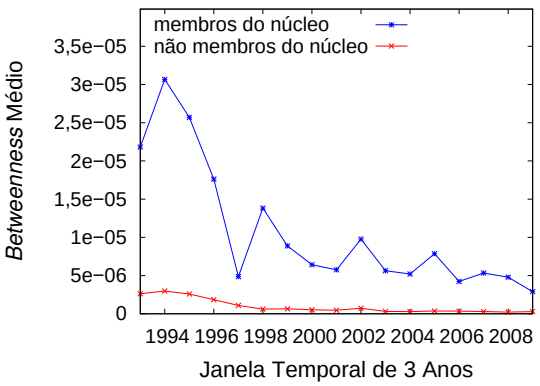

(a) Coeficiente de agrupamento




(b) Grau médio

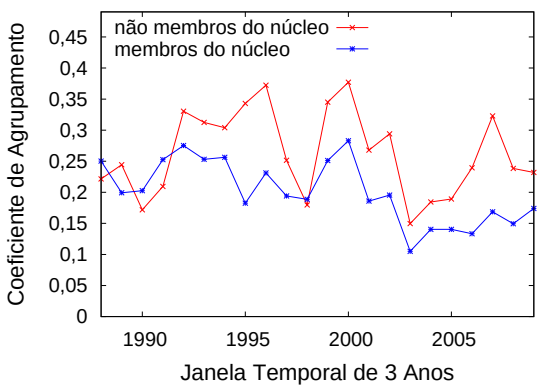



(c) Maior CFC

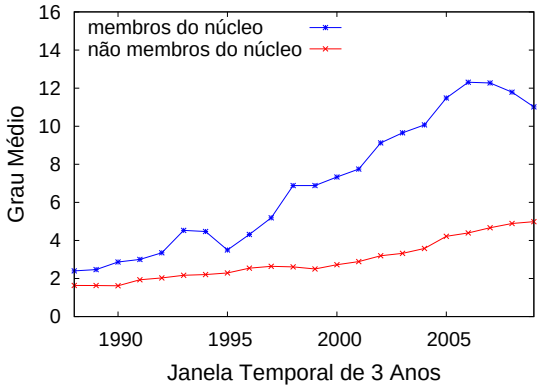



(d) *Betweenness* médio

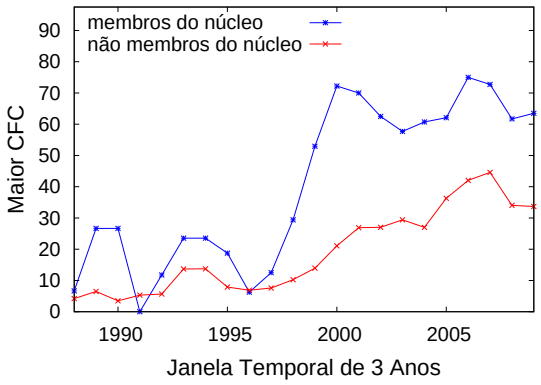
Figura C.14: Propriedades da comunidade SAC para os membros e não membros do núcleo



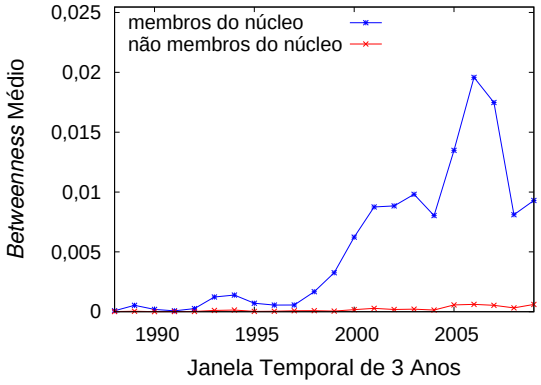

(a) Coeficiente de agrupamento




(b) Grau médio

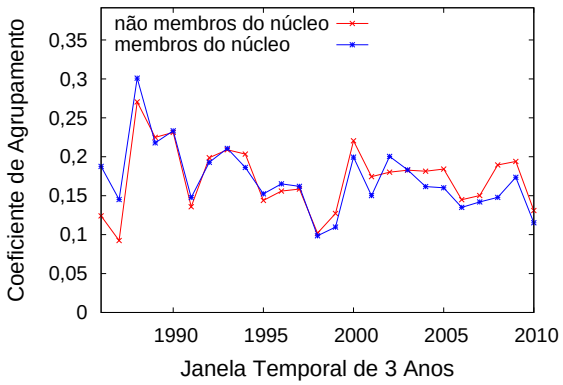



(c) Maior CFC

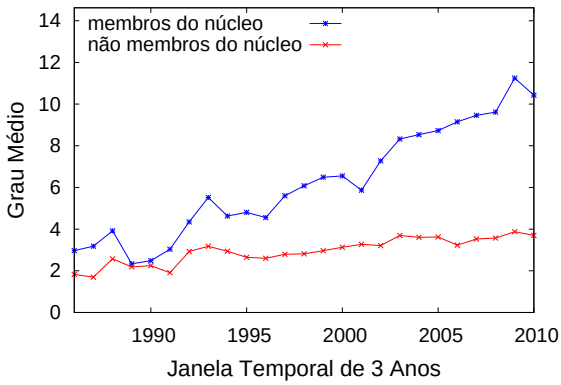



(d) *Betweenness* médio

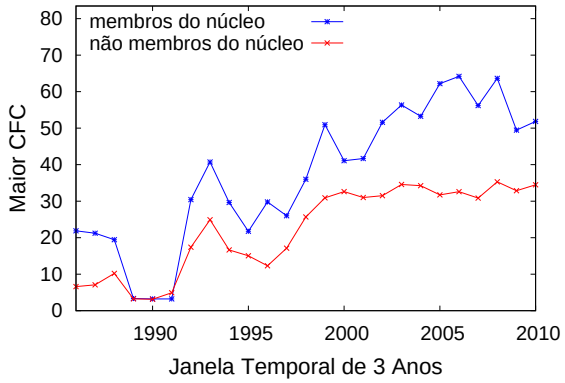
Figura C.15: Propriedades da comunidade SIGCOMM para os membros e não membros do núcleo



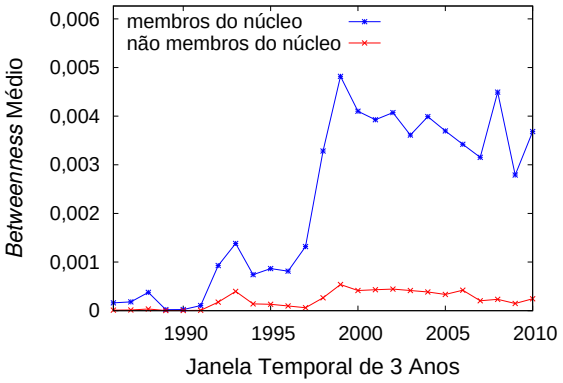

(a) Coeficiente de agrupamento




(b) Grau médio

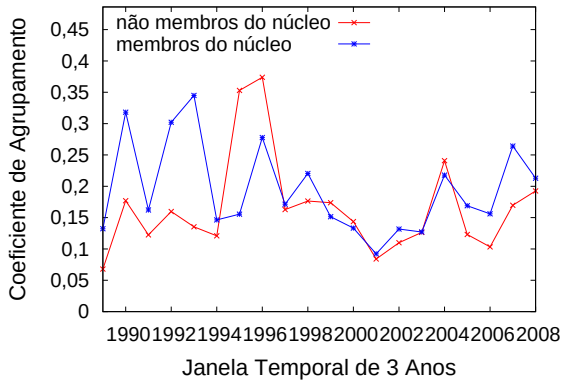



(c) Maior CFC

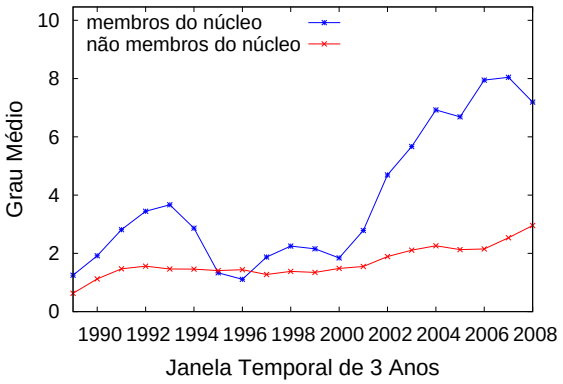



(d) *Betweenness* médio

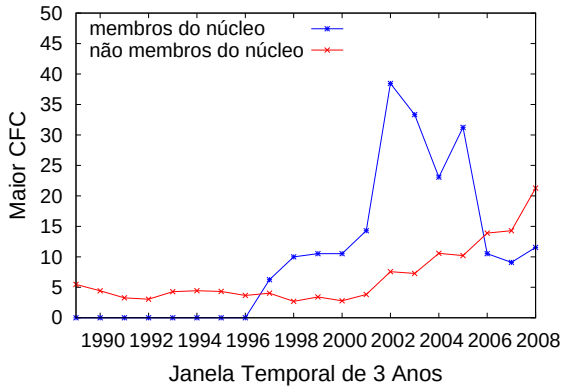
Figura C.16: Propriedades da comunidade SIGCSE para os membros e não membros do núcleo



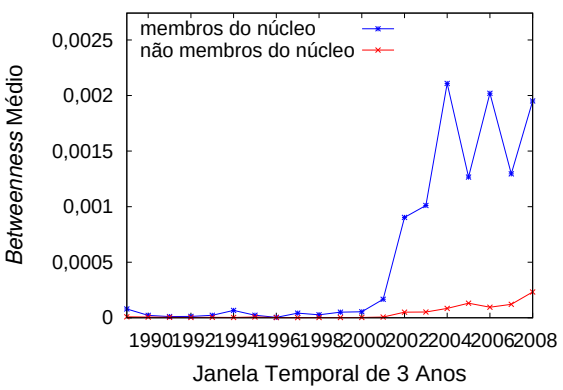

(a) Coeficiente de agrupamento




(b) Grau médio

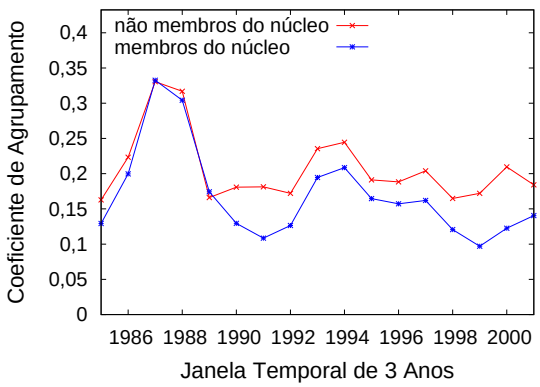



(c) Maior CFC

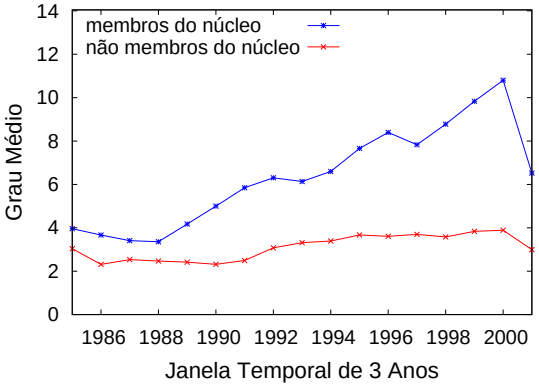



(d) *Betweenness* médio

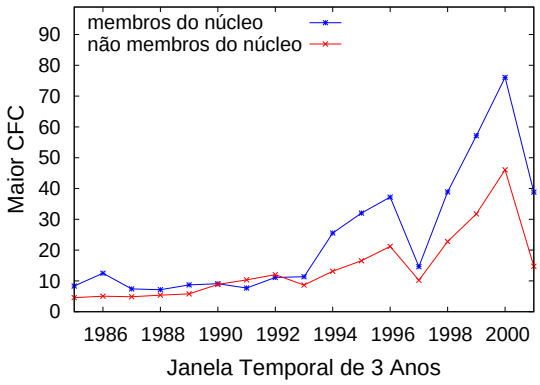
Figura C.17: Propriedades da comunidade SIGDOC para os membros e não membros do núcleo



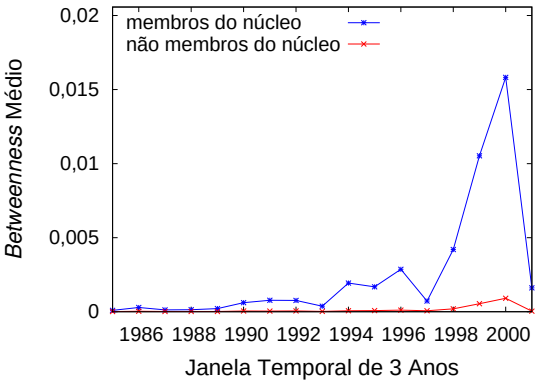

(a) Coeficiente de agrupamento




(b) Grau médio

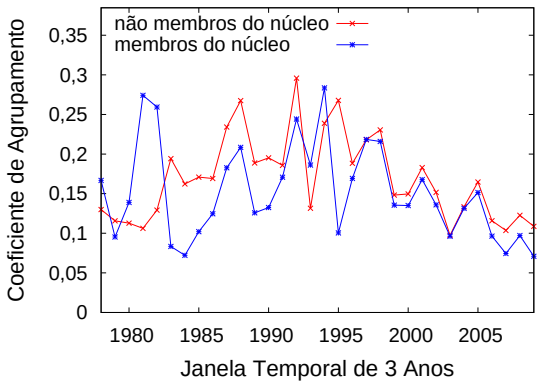



(c) Maior CFC

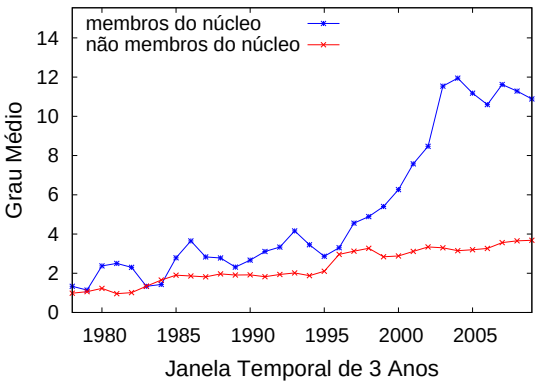



(d) *Betweenness* médio

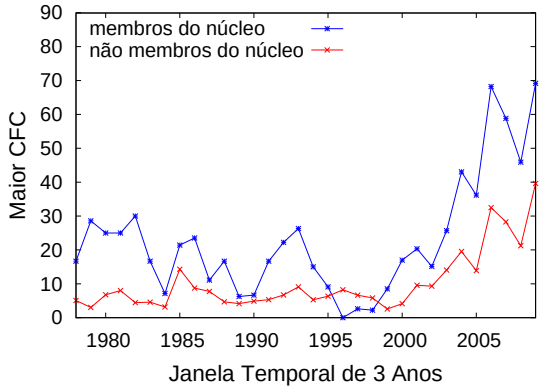
Figura C.18: Propriedades da comunidade SIGGRAPH para os membros e não membros do núcleo



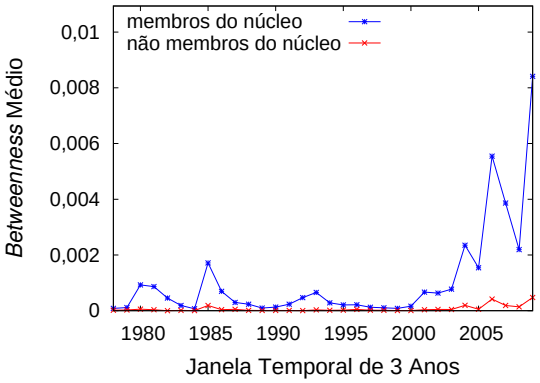

(a) Coeficiente de agrupamento




(b) Grau médio

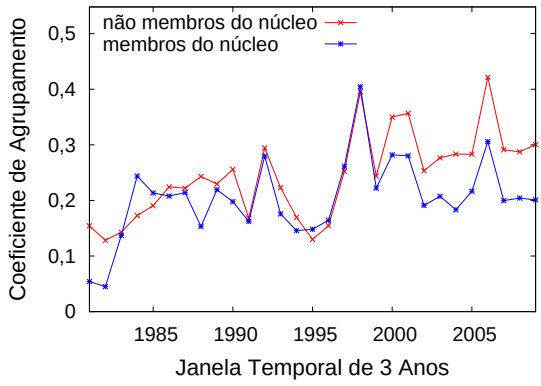



(c) Maior CFC

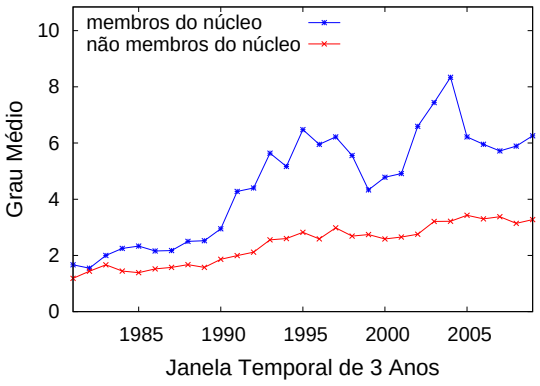



(d) *Betweenness* médio

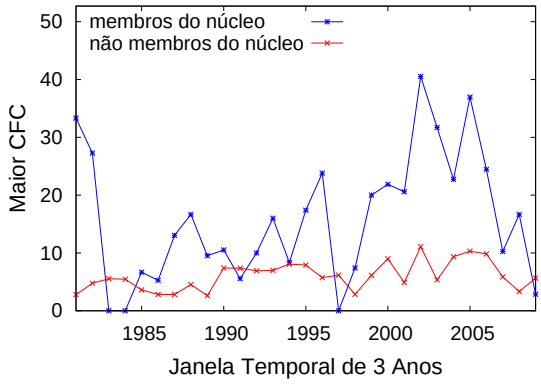
Figura C.19: Propriedades da comunidade SIGIR para os membros e não membros do núcleo



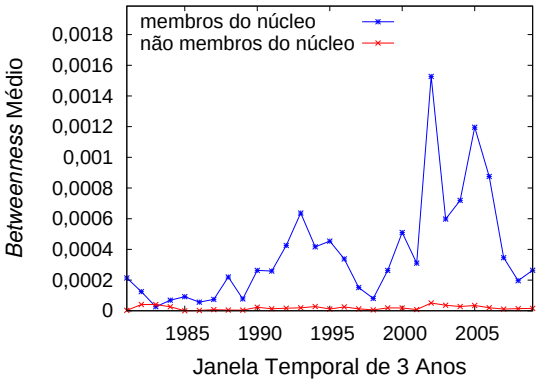

(a) Coeficiente de agrupamento




(b) Grau médio

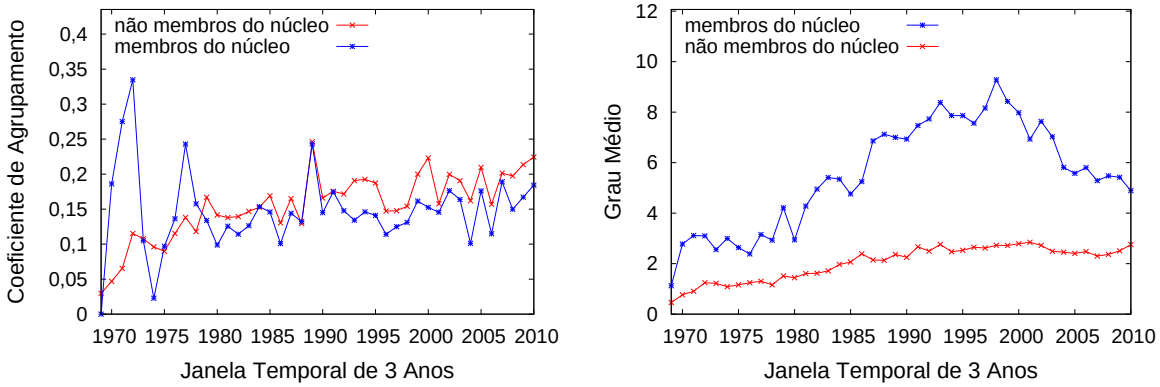



(c) Maior CFC




(d) *Betweenness* médio

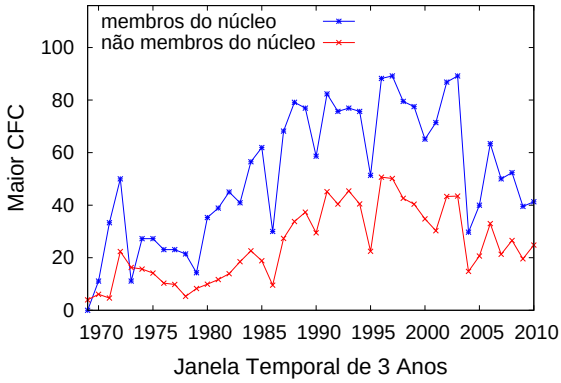
Figura C.20: Propriedades da comunidade SIGMETRICS para os membros e não membros do núcleo



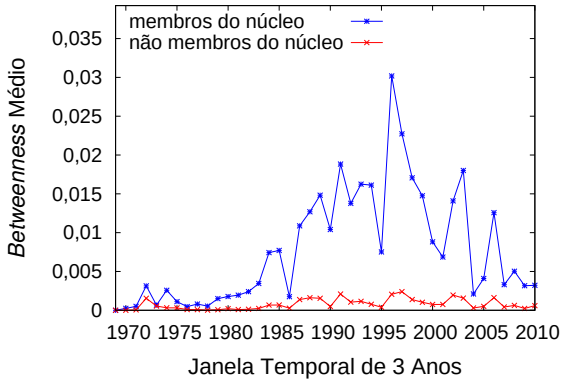

(a) Coeficiente de agrupamento




(b) Grau médio




(c) Maior CFC




(d) *Betweenness* médio

Figura C.21: Propriedades da comunidade STOC para os membros e não membros do núcleo

# Apêndice D

# Evolução do *CoScore*

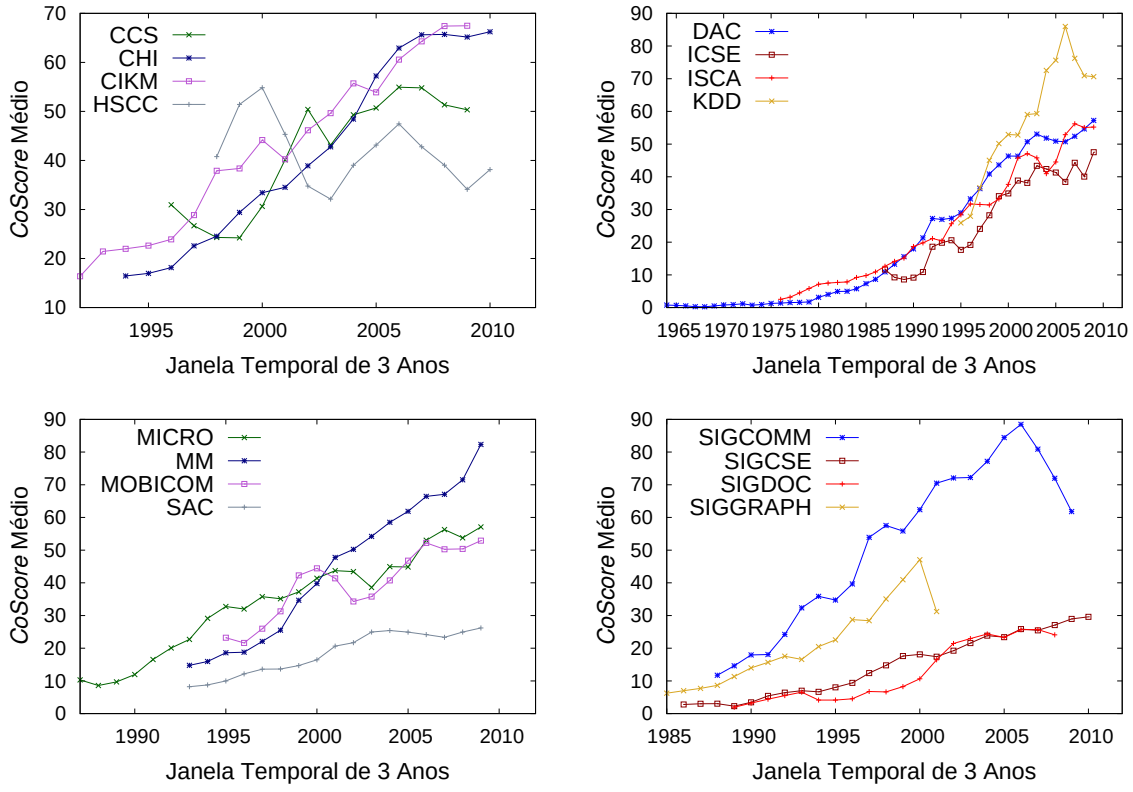
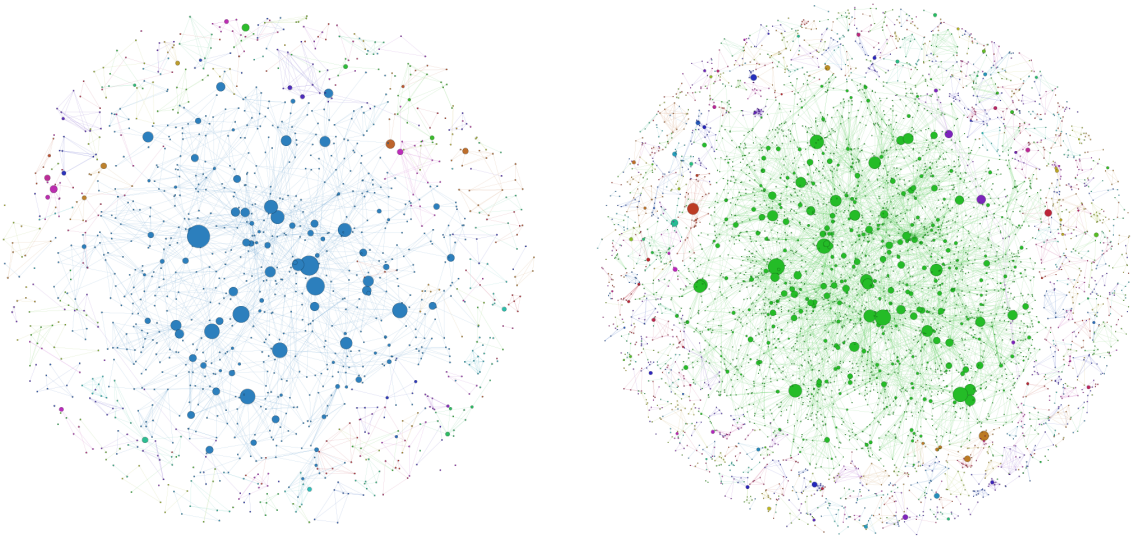



Figura D.1: *CoScore* médio das comunidades científicas

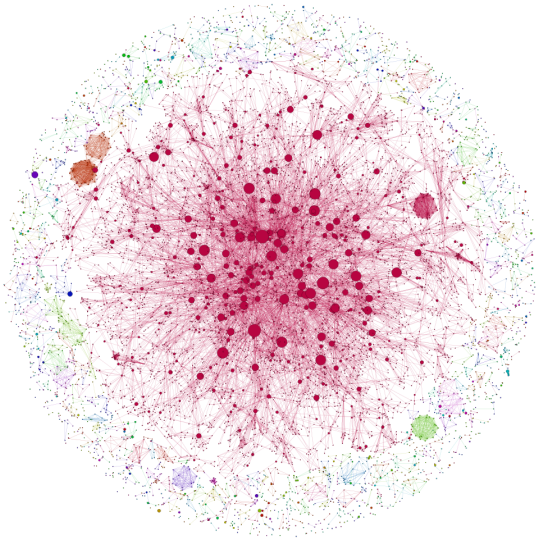
# Apêndice E

# Visualização das Comunidades Científicas

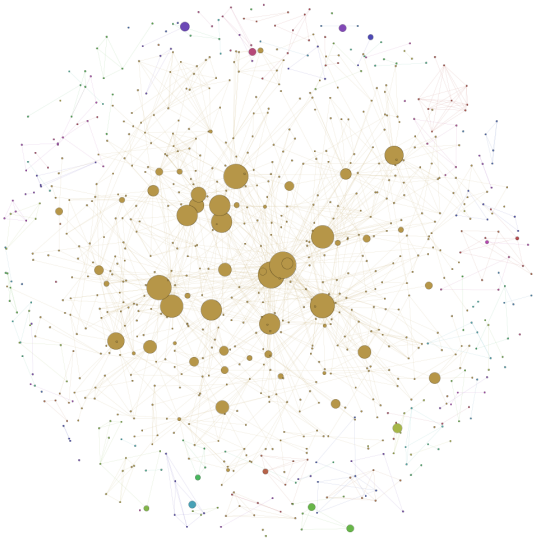


(a) CCS

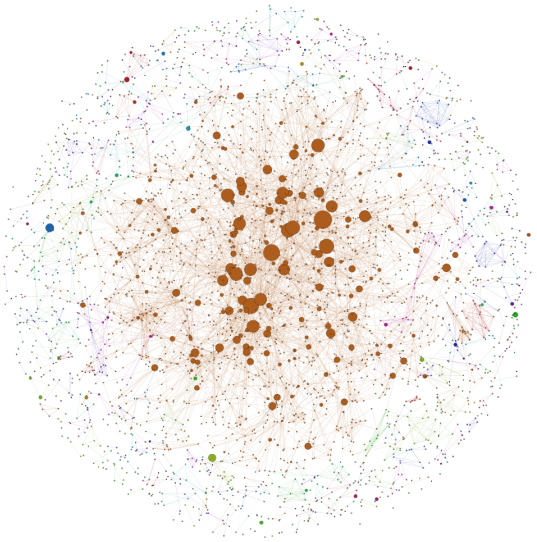
(b) CIKM



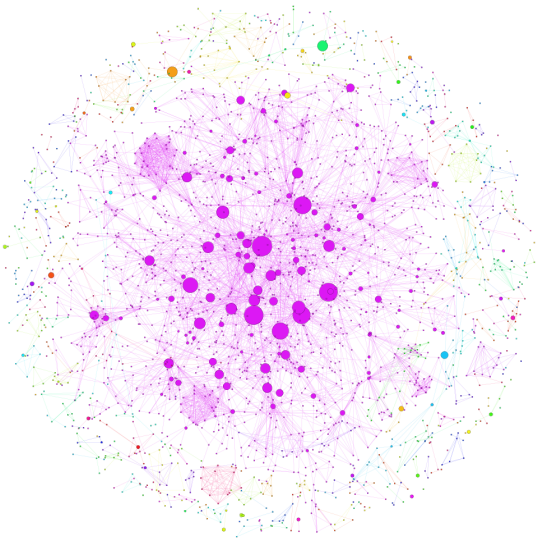
(c) DAC



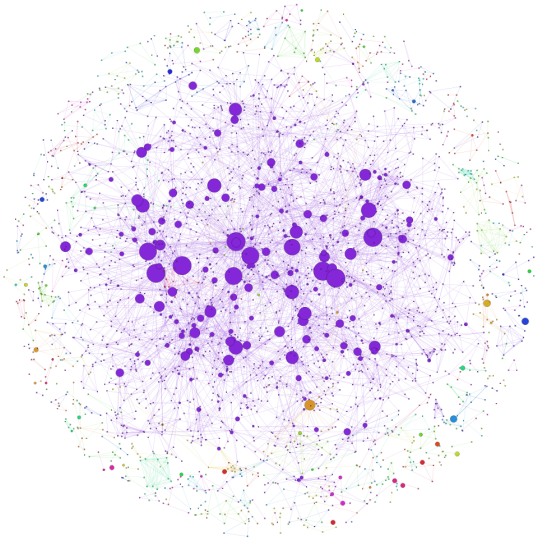
(d) HSCC



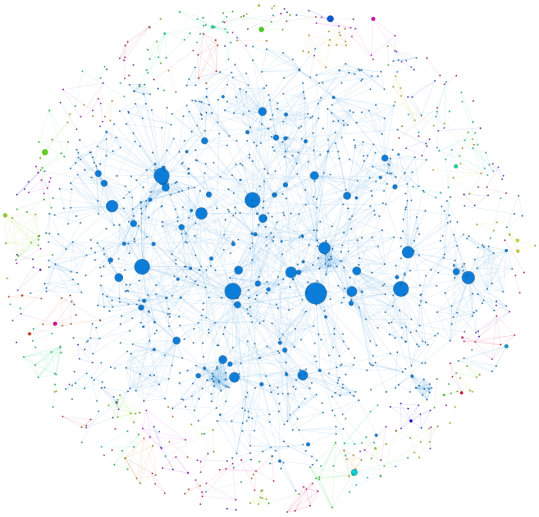
(e) ICSE



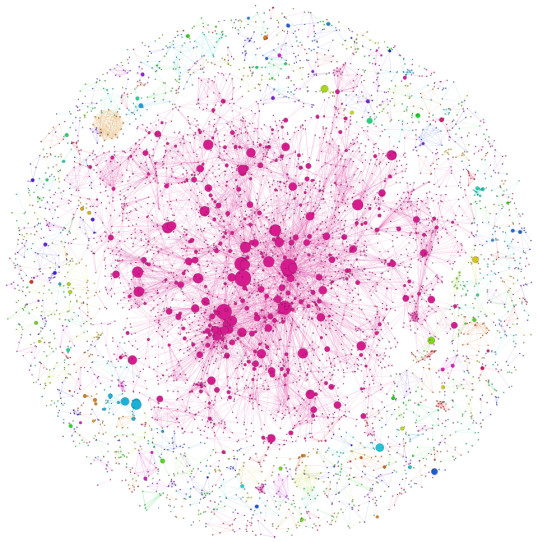
(f) ISCA



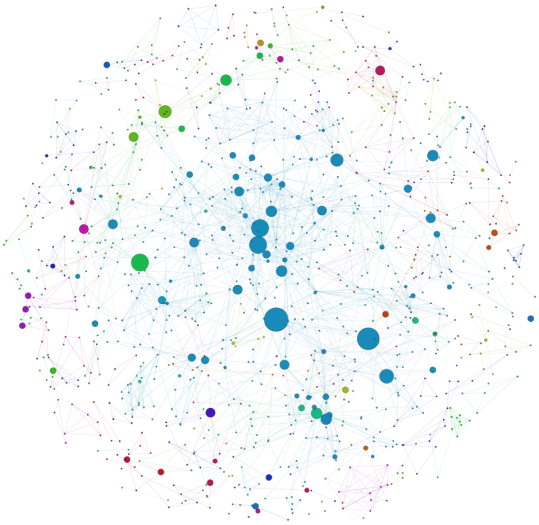
(g) KDD



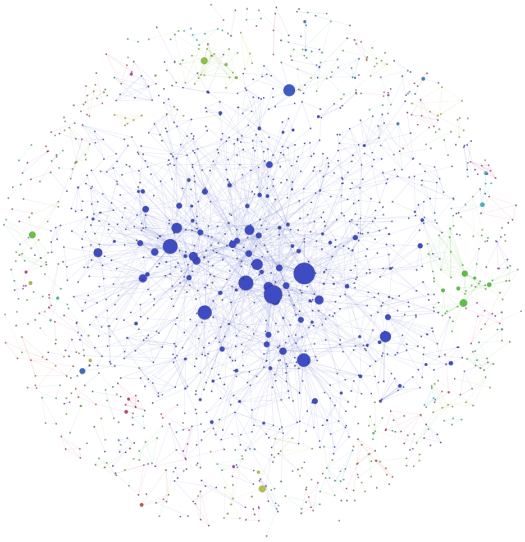
(h) MICRO



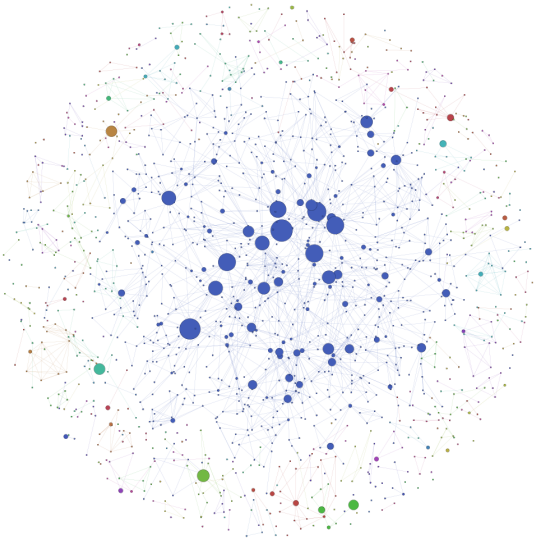
(i) MM



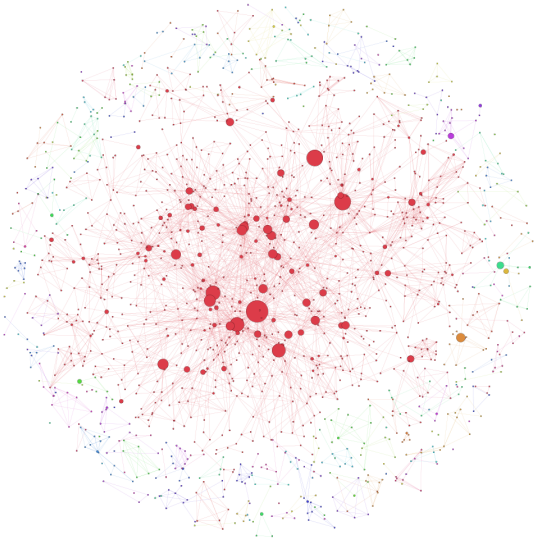
(j) MOBICOM



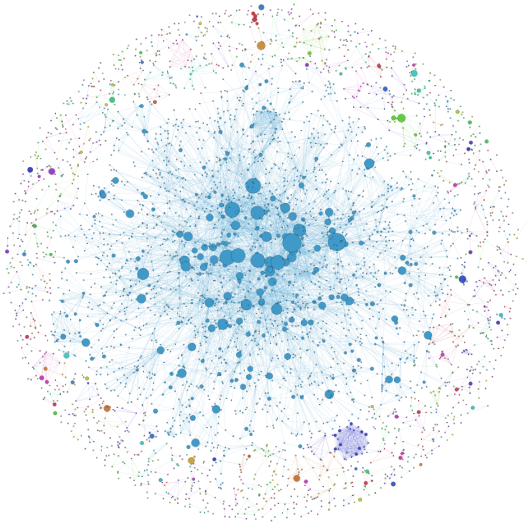
(k) PODC



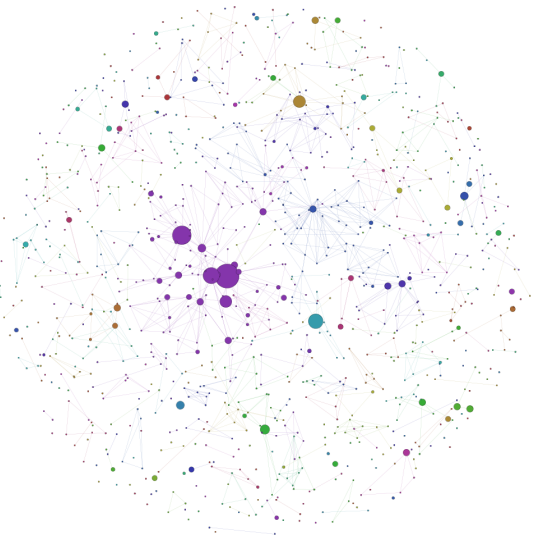
(l) POPL



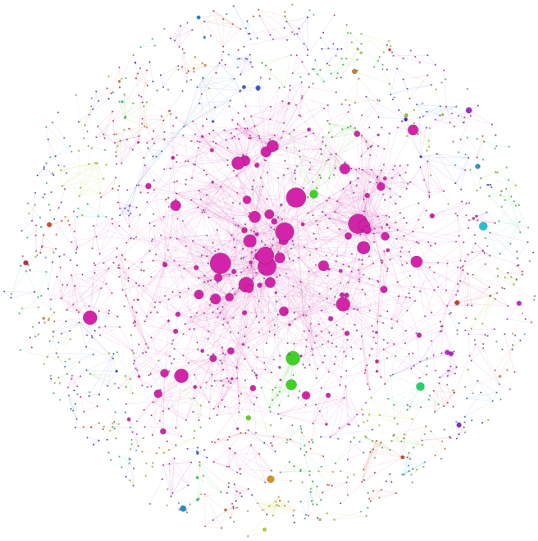
(m) SIGCOMM



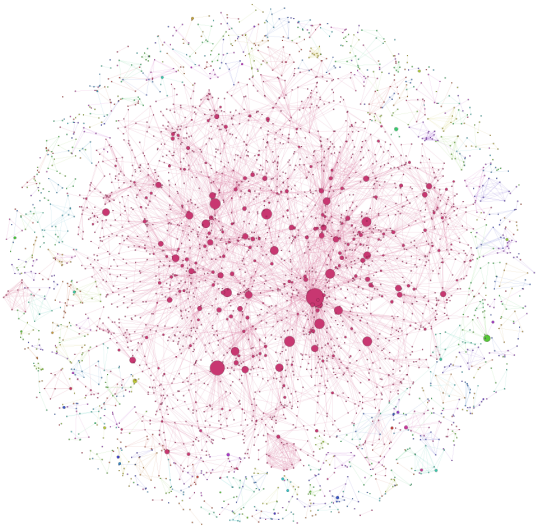
(n) SIGCSE



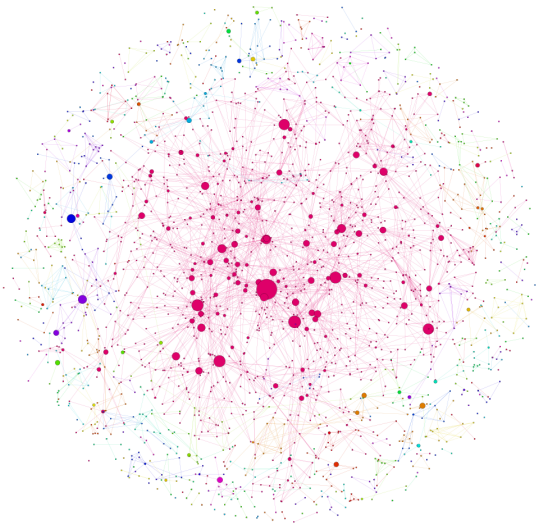
(o) SIGDOC



(p) SIGGRAPH



(q) SIGIR



(r) SIGMETRICS

Figura E.1: Instância final das comunidades científicas