Uma Abordagem para Classificação de Interações Sociais Dinâmicas a partir de seus Atributos

Thiago H. P. SIlva, Alberto H. F. Laender

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
{thps,laender}@dcc.ufmg.br

Abstract. Network analyses provide important information for understanding how a network evolves. In this context, some studies focus on classifying nodes and their relationships based on topological properties and centrality metrics. Instead, we discuss the importance of applying the notion of social capital to the classification process. Here, we propose a new approach to classify nodes and edges in temporal multigraphs based on the persistence of the edges' attributes. Overall, our results show that the social role of the nodes and the strength of their ties are statistically well-defined when compared with several traditional graph metrics.

Categories and Subject Descriptors: [**Information systems applications**]: Collaborative and social computing systems and tools

Keywords: Edge Classification, Node Classification, Social Networks

1. INTRODUÇÃO

Como classificar as interações dinâmicas em uma rede social? Diversos trabalhos têm investigado propriedades e padrões topológicos de redes sociais com a finalidade de definir o comportamento de seus atores, bem como mensurar a força de seus relacionamentos [Barabási 2009; Easley and Kleinberg 2010; Leão et al. 2018; Watts 2004]. Explorar comportamentos e dinâmicas dos atores em uma rede social é essencial para um bom entendimento de sua estrutura, o que é geralmente caracterizado por meio de grafos que capturam os aspectos sociais envolvidos [Easley and Kleinberg 2010].

Neste contexto, estudos têm explorado a noção de capital social dada pelo posicionamento estratégico de um determinado ator em uma estrutura social [Burt 2005; Granovetter 1973]. Por exemplo, Granovetter [1973] define o conceito de *weak ties* como sendo aquelas relações importantes que tornam uma rede mais coesa através da criação de pontes. De forma similar, Freire & Figueiredo [2011] exploraram tal conceito com o objetivo de mensurar a importância de grupos e indivíduos de acordo com a capacidade de conectar diferentes partes de uma rede.

Este artigo contribui para essa discussão ao mapear tais conceitos sociais nas relações entre nodos para relações do tipo nodo-atributo. Especificamente, aplicamos os conceitos de *closure* e *brokerage* que definem, respectivamente, a habilidade de agregar padrões similares e a capacidade de criar pontes com padrões diversificados [Burt 2005]. Sendo assim, nosso estudo objetiva classificar o papel social dos nodos, bem como de suas interações dinâmicas.

Em resumo, nossas contribuições são: (i) definição de classes sociais para classificar nós e suas

Work supported by project MASWeb (FAPEMIG/PRONEX grant APQ-01400-14) and by the authors' individual grants from CNPq and CAPES.

interações dinâmicas, bem como capturar o significado social dos relacionamentos; (ii) proposta de uma nova estratégia para classificar nós e arestas baseando-se nos relacionamentos entre nós e seus atributos; e (iii) caracterização das diferenças dos relacionamentos sociais em distintas redes sociais acadêmicas e sua avaliação através de métricas de rede.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta uma breve revisão de trabalhos relacionados, enquanto a Seção 3 apresenta a modelagem adotada para representação das redes sociais consideradas. Em seguida, a Seção 4 apresenta a metodologia adotada para avaliação da abordagem proposta e a Seção 5 analisa os resultados dos experimentos realizados nessa avaliação. Finalmente, a Seção 6 apresenta as nossas conclusões e algumas considerações sobre trabalhos futuros.

2. TRABALHOS RELACIONADOS

Recentemente, vários trabalhos têm estudado as características topológicas das redes sociais com o intuito de melhor entender as relações sociais envolvidas [Easley and Kleinberg 2010; Leão et al. 2018; Newman 2004]. Ao classificarem as interações sociais a partir de métricas de topologia de rede, Leão et al. [2018] analisam as interações sociais expressas pela estrutura topológica com o objetivo de filtrar ruídos devido às relações consideradas aleatórias. Já Brandão & Moro [2017] investigam a força dos relacionamentos sociais a partir de métricas topológicas em redes sociais acadêmicas.

Em outro contexto, Silva et al. [2015] exploram conceitos de capital social para mensurar o grau de influência de indivíduos através de vínculos sociais com suas comunidades. Tendo como base o compartilhamento de informação entre os nodos, Adamic & Adar [2003] mensuram a força dos relacionamentos a partir da análise de similaridade entre as mensagens trocadas entre indivíduos, enquanto Gilbert & Karahalios [2009] consideram também os aspectos temporais das interações.

O trabalho mais próximo do nosso é o algoritmo RECAST (*Random rElationship ClASsifier sTrategy*) que filtra relações aleatórias e designa classes sociais (amizade, conhecido, ponte e aleatório) para os relacionamentos em uma rede dinâmica [Vaz de Melo et al. 2015]. Similarmente, o modelo proposto analisa aspectos temporais e a regularidade das relações ao longo do tempo. Em contraste, a abordagem proposta neste artigo se diferencia ao classificar tanto nodos quanto múltiplas relações, definindo classes mais abrangentes vinculadas a conceitos sociais. Embora alguns trabalhos relacionados tendem a caracterizar redes com o intuito de evidenciar e discutir suas definições de forma empírica, a metodologia experimental deste estudo é expandida ao também correlacionar a classificação com algoritmos amplamente usados na literatura.

3. ABORDAGEM PROPOSTA

Nesta seção introduzimos o modelo de grafo proposto para possibilitar a mineração de múltiplas relações ao longo do tempo. Em seguida, apresentamos o processo de extração de atributos relevantes, bem como os algoritmos de classificação de arestas e nodos.

Grafo Temporal com Múltiplas Arestas. Definimos um multigrafo temporal como um conjunto de nodos e arestas formados em intervalos discretos k, adaptando os modelos empregados por Vaz de Melo et al. [2015] e Shah et al. [2016]. Formalmente, esse tipo de grafo é definido por $\mathcal{G}_k = (\mathcal{V}_k, \mathcal{E}_k, m)$, onde m: $\mathcal{E}_k \rightarrow \{\{u, v\} | u, v \in \mathcal{V}_k\}$ é a função que possibilita a existência de múltiplas arestas ao atribuir cada aresta $e \in \mathcal{E}_k$ a um par de nodos. Neste contexto, $G_t = \bigcup_1^t \mathcal{G}_i$ representa o grafo temporal agregado que compreende o conjunto de todos os nodos e suas interações dentro do intervalo de tempo [1,t]. Assim, dado o conjunto de atributos $\mathcal{A}$, a função Φ: $e \in \bigcup_{i=1}^{k} \mathcal{E}_i \rightarrow a$ realiza o mapeamento de cada aresta para um subconjunto de atributos. Note que cada subconjunto $a \subseteq \mathcal{A}$ tem o potencial, em uma aplicação real, de ser associado a diversos vértices em diferentes faixas temporais, de modo que é possível transformar o grafo $\mathcal{G}_k$ no grafo de atributos $\mathcal{H}_k = (\mathcal{V}_k \cup \mathcal{A}, \mathcal{E}_k^{'}, m')$, onde a função m': $\mathcal{E}_k^{'} \rightarrow \{\{u, i\} | u \in \mathcal{V}_k \wedge i \in \mathcal{A}\}$. Assim, $\mathcal{H}_k$ é uma abstração que possibilita transformar os atributos de

Algorithm 1 Extração de Atributos Relevantes

Require: H, t, k
Ensure: $\Gamma_k(u)$, $\forall u \in V_t$
1: **for all** $u \in V_t$ **do**
2: $\quad \mathcal{A}_{temp} \leftarrow \lambda$
3: $\quad$ **for all** $k \in [1,t]$ **do**
4: $\quad\quad \Gamma_k(u) \leftarrow \lambda$
5: $\quad\quad \mathcal{A}_{temp} \leftarrow \mathcal{A}_{temp} \cup \{a|(u,a) \in \mathcal{E}_k^{'}\}$
6: $\quad\quad vector \leftarrow \lambda$
7: $\quad\quad$ **for all** $i \in \mathcal{A}_{temp}$ **do**
8: $\quad\quad\quad vector.add(pers_k(u,i))$
9: $\quad\quad IQR \leftarrow p(vector,75) - p(vector,25)$
10: $\quad\quad$ **for all** $i \in \mathcal{A}_{temp}$ **do**
11: $\quad\quad\quad$ **if** $pers_k(u,i) > p(vector,75) + IQR * 1.5$ **then**
12: $\quad\quad\quad\quad \Gamma_k(u) \leftarrow \Gamma_k(u) \cup i$

Algorithm 2 Classificação das Arestas

Require: G_t, t, k, Φ e Γ
Ensure: $\Delta(e), \forall e \in \bigcup_{i=1}^{t} \mathcal{E}_i$
1: **for all** $k \in [1,t]$ **do**
2: $\quad$ **for all** $e \in \mathcal{E}_k$ **do**
3: $\quad\quad (u,v) \leftarrow getNodes(e)$
4: $\quad\quad$ **if** $|\Gamma_k(u)| \neq 0$ **then**
5: $\quad\quad\quad$ **if** $|\Gamma_k(u) \cap \Phi(e)| \neq 0$
6: $\quad\quad\quad\quad$ **then** $u_{state} \leftarrow closure$
7: $\quad\quad\quad$ **else** $u_{state} \leftarrow brokerage$
8: $\quad\quad$ **else if** $\sum_{j=1}^{k} \mathbb{1}()_{[u \in \mathcal{V}_j]} > 1$ **then**
9: $\quad\quad\quad u_{state} \leftarrow no\text{-}info$
10: $\quad\quad$ **else** $u_{state} \leftarrow sporadic$
11: $\quad\quad$ **if** $|\Gamma_k(v)| \neq 0$ **then**
12: $\quad\quad\quad$ **if** $|\Gamma_k(v) \cap \Phi(e)| \neq 0$ **then**
13: $\quad\quad\quad\quad v_{state} \leftarrow closure$
14: $\quad\quad\quad$ **else** $v_{state} \leftarrow brokerage$
15: $\quad\quad$ **else if** $\sum_{j=1}^{k} \mathbb{1}()_{[v \in \mathcal{V}_j]} > 1$ **then**
16: $\quad\quad\quad v_{state} \leftarrow no\text{-}info$
17: $\quad\quad$ **else** $v_{state} \leftarrow sporadic$
18: $\quad\quad \Delta(e) \leftarrow \delta(\{u_{state}, v_{state}\})$

cada aresta em um nodo adicional, permitindo, desta forma, que um nodo original possa ser ligado a esse novo nodo.

Extração de Atributos Relevantes. O próximo passo consiste em extrair os atributos relevantes definidos pelo próprio nodo ao considerar o seu histórico de interações. Para isso, usamos o conceito de *edge persistence* adaptado para o grafo de atributos como $pers_t(u,a) = \frac{1}{t}\sum_{k=1}^{t} \mathbb{1}()$, onde a função indicadora retorna o valor 1 se a aresta (u,a) existe em $\mathcal{E}_k^{'}$, ou 0 caso contrário. É importante ressaltar que a operação é realizada sobre cada grafo de atributos em intervalos discretos ($\{\mathcal{H}_1, ..., \mathcal{H}_k\}$) e não sobre o grafo agregado. O Algoritmo 1 detalha o processo de extração de atributos relevantes, recebendo como entrada o grafo agregado H ($\{\mathcal{H}_1, ..., \mathcal{H}_t\}$), o tempo t e suas subdivisões destacadas por k. Em resumo, o algoritmo inspeciona todos os atributos através da métrica *edge persistence* e os avalia conforme suas relevâncias por meio de percentis (função p nas linhas 9 e 11). Assim, ele constrói o conjunto $\Gamma_k(u)$ de atributos estatisticamente relevantes para cada nodo u na faixa temporal $[1,k]$. Como verificamos que os valores da métrica *edge persistence* não seguem uma distribuição normal, então optamos pela seleção de atributos relevantes com base na definição de detecção de *outliers* dada pelo *interquartile range* (IQR)[1]. No pior caso, quando estão presentes todas as arestas em todos os instantes com todos os atributos, a complexidade do Algoritmo 1 é $O(|V|\frac{t}{k}(|E| + |\mathcal{A}|))$.

Classificação das Arestas. As arestas múltiplas são classificadas de acordo com o grau de relevância de seus atributos dado pelo passado de cada um dos nodos de cada interação. Definimos como estados de interação dinâmica de cada nó o conjunto {*closure, brokerage, no-info, sporadic*}, cujos valores são atribuídos a cada instante em que uma aresta é inspecionada. Um nodo possui estado do tipo *closure* quando há um vínculo temporal forte com os atributos de sua competência que estão sendo compartilhados no exato momento da interação. Quando há atributos relevantes no histórico que não estão sendo empregados na aresta inspecionada, então atribui-se o estado *brokerage*. Caso não haja atributos relevantes e o nó possua interações em mais de um instante, então atribui-se o estado *no-info*. Por fim, considera-se o nó como *sporadic* quando há apenas um registro dele no grafo temporal. O Algoritmo 2 descreve o processo de classificação das arestas de acordo com os históricos de cada um dos nodos. Uma vez definidos os estados dinamicamente atribuídos em diferentes faixas temporais, as arestas são definidas de acordo com os estados dos seus nodos (linha 18). Quando todas as arestas estão presentes em todos os instantes, a complexidade do Algoritmo 2 é $O(\frac{t}{k}|E|)$. A Tabela I descreve

[1]Outra abordagem seria o uso do *z-score* modificado para a mesma finalidade. Como os resultados experimentais foram semelhantes para o IQR e para o *z-score* modificado, optamos pelo uso do IQR devido à possibilidade de se aplicar restrições através dos percentis.

Table I: Mapeamento dos estados dinâmicos dos nodos para determinação da classe da aresta.

$\delta(\{closure, closure\})$	$\rightarrow$ *very strong*
$\delta(\{closure, brokerage\})$	$\rightarrow$ *strong bridge*
$\delta(\{closure, no\text{-}info\})$	$\rightarrow$ *strong*
$\delta(\{closure, sporadic\})$	$\rightarrow$ *strong*
$\delta(\{brokerage, brokerage\})$	$\rightarrow$ *regular bridge*
$\delta(\{brokerage, no\text{-}info\})$	$\rightarrow$ *weak bridge*
$\delta(\{brokerage, sporadic\})$	$\rightarrow$ *weak bridge*
$\delta(\{no\text{-}info, no\text{-}info\})$	$\rightarrow$ *ordinary*
$\delta(\{no\text{-}info, sporadic\})$	$\rightarrow$ *ordinary*
$\delta(\{sporadic, sporadic\})$	$\rightarrow$ *sporadic*

o mapeamento δ de todos os pares do conjunto de estados para determinação das classes das arestas. São definidas sete classes: *very strong, strong, strong bridge, regular bridge, weak bridge, ordinary* e *sporadic*. Em resumo, têm-se papéis sociais e suas intensidades que passam a indicar (i) a presença de atributos relevantes nas interações (*very strong*, *strong bridge* e *strong*), (ii) arestas com potencial de transferência de atributos relevantes (todas indicados como *bridge*) e (iii) arestas sem nenhuma informação relevante ou corriqueiras (*ordinary* e *sporadic*).

Classificação dos Nodos. Para classificação dos nodos, consideramos três classes específicas: *hub*, *regular* e *sporadic*. Por *hub*, entende-se um nodo que tem autoridade para difundir atributos relevantes. Já *regular* indica que o nodo não possui um domínio específico. Por fim, *sporadic* considera a existência de apenas um registro para o nodo em uma faixa temporal. A função de classificação dos nodos (Ω), que pode ser obtida em O(1) pelo Algoritmo 2 por meio de *flags*, é dada por:

$$\Omega(u) = \begin{cases} hub, & \text{se } |\Gamma_t(u)| \neq 0 \\ regular, & \text{senão se } \sum_{k=1}^{t} \mathbb{1}()_{[u \in \mathcal{V}_k]} > 1 \\ sporadic, & \text{demais casos.} \end{cases}$$

4. METODOLOGIA

Como avaliar a classificação de interações sociais em uma rede social? Esse processo é bastante desafiador devido à inexistência de redes reais que definam claramente os papéis sociais de seus nodos e arestas. Alternativamente, a metodologia experimental para avaliar esse tipo de classificação pode ser realizada por meio da caracterização de redes sociais com diferentes propriedades. Outra alternativa consiste em comparar a classificação com algoritmos para grafos bem conhecidos.

Dados Utilizados. Para avaliação da abordagem proposta neste artigo, o grafo correspondente foi construído a partir da rede social acadêmica referente a um conjunto de 24 comunidades científicas[2] derivadas de alguns dos principais Grupos de Interesse Especial da ACM[3], cujos dados foram coletadas da DBLP[4] em Junho de 2018. A Tabela II lista as comunidades consideradas e algumas de suas estatísticas. Em resumo, tem-se subredes de características bem distintas que possibilitam contrastar o efeito da abordagem de classificação proposta sob cada uma delas.

A construção do multigrafo temporal leva em consideração o ano k em que cada conferência foi realizada. Assim, para cada artigo publicado no ano k, todos os pares de coautores formam arestas (u, v) de um grafo $\mathcal{G}_k$. Note que este modelo possibilita a existência de múltiplas arestas entre os seus nodos no instante k, conforme destacado na Seção 3. O conjunto de atributos compreende dados extraídos do título dos artigos após passarem pelo processo de remoção de palavras sem significado (*stop words*) e pela redução das palavras flexionadas para as suas respectivas raízes (*stemming*). Cada palavra do título filtrado torna-se parte do conjunto a e, assim, há uma relação de cada um dos coautores u com a no grafo de atributos $\mathcal{H}_k$.

[2]Este conjunto de conferências foi estudado por Alves et al. [2013] e um subconjunto dele por Silva et al. [2015].
[3]Association for Computing Machinery: `http://www.acm.org/sigs`
[4]DBLP: `https://dblp.uni-trier.de/`

Table II: Estatísticas das comunidades acadêmicas.

Comunidade	#nodos	#arestas	#arestas temporais	#triângulos	Transit.	CC	Densidade
SAC	10804	18066	19712	17734	65,4%	0,73	3,10E-04
DAC	10272	27800	31972	59736	49,1%	0,70	5,27E-04
CHI	8959	27587	32154	43749	35,1%	0,77	6,87E-04
CIKM	7342	16347	18822	18101	43,6%	0,75	6,07E-04
MMSys	7124	18728	22783	27810	37,4%	0,77	7,38E-04
SIGCSE	6247	15252	18232	23837	37,5%	0,66	7,82E-04
KDD	4998	13614	15150	27806	49,3%	0,77	1,09E-03
SIGIR	4905	11247	13595	13172	35,9%	0,69	9,35E-04
SIGMOD	4869	16042	18090	33527	45,9%	0,74	1,35E-03
CCS	2854	6851	7612	9009	49,8%	0,77	1,68E-03
SIGCOMM	2844	8653	9715	17469	43,4%	0,77	2,14E-03
ICSE	2829	4977	5354	5441	61,8%	0,67	1,24E-03
SIGUCCS	2517	2349	2734	2933	72,2%	0,32	7,42E-04
STOC	2500	5568	6608	4393	23,0%	0,49	1,78E-03
SIGMETRICS	2440	4500	4906	4377	49,9%	0,67	1,51E-03
SIGGRAPH	2439	4568	4935	5364	48,7%	0,64	1,54E-03
ISCA	2257	8748	9231	83390	90,6%	0,71	3,44E-03
MobiCom	2074	5056	5732	7025	58,0%	0,82	2,35E-03
PODC	1972	3573	4353	2776	29,5%	0,61	1,84E-03
POPL	1858	3129	3495	2896	45,9%	0,55	1,81E-03
SIGDOC	1570	1847	2048	1580	60,6%	0,53	1,50E-03
MICRO	1321	2907	3108	4651	71,0%	0,72	3,33E-03
ISSAC	1253	1705	2154	1126	33,8%	0,45	2,17E-03
HSCC	361	546	572	491	56,7%	0,60	8,40E-03
Média	4025,4	9569,2	10961,1	17433,0	49,8%	0,66	1,77E-03
Mediana	2673	6209,5	7110	8017,0	48,9%	0,70	1,51E-03
Desvio Padrão	2959,3	7929,9	9224,4	20620,5	15,4%	0,12	1,63E-03
Rede Completa	79684	221541	263067	417042	38,2%	0,67	6,98E-05

Propriedades das Redes. Para avaliar a classificação proposta, utilizamos três métricas de centralidade (*degree, closeness* e *betweenness*) para determinar quais classes tendem a ter papéis mais bem definidos em uma estrutura social, a métrica *clustering coefficient* para mensurar o grau de coesão de cada classe e a métrica $PageRank$ para avaliar a importância dos nodos [Easley and Kleinberg 2010].

5. EXPERIMENTOS

Para avaliação da nossa abordagem, dividimos os experimentos em três etapas: (i) caracterização e discussão da classificação para diferentes comunidades; (ii) comparação das classes atribuídas de acordo com métricas de topologia de rede; e (iii) análise de sensibilidade da classificação proposta.

5.1 Caracterização

A Tabela III apresenta a distribuição das classes de nodos e arestas múltiplas para as 24 diferentes comunidades. A classificação dos nodos evidencia uma presença expressiva de instâncias da classe *sporadic* (média de 57,8%). De fato, uma rede social acadêmica possui uma forte presença de nodos novos como, por exemplo, estudantes ou colaboradores em relações interdisciplinares. No entanto, destaca-se também uma forte presença de instâncias da classe *hub* com percentagens acima de 30% para comunidades mais estabelecidas como CIKM, KDD, SIGIR, SIGMOD, STOC, SIGMETRICS, ISCA, PODC, POPL e MICRO. Ou seja, boa parte dos membros dessas comunidades tende a manter uma coerência nos tópicos de pesquisa ao longo de suas trajetórias acadêmicas. Em contraste, comunidades como SAC, SIGUCCS e SIGDOC possuem percentuais bem baixos para a classe *hub*, em razão da pouca sinergia entre os seus membros. Em geral, tais percentuais podem ser vistos como evidências das características de cada comunidade. Por exemplo, a comunidade STOC apresenta a tendência de seus membros terem competência em tópicos bem específicos da área de Teoria da Computação. Em contrapartida, SAC é uma comunidade com foco em computação aplicada englobando um leque bem diverso de temas.

Table III: Composição das classificações de nodos e arestas.

	Classificação dos Nodos			Classificação das Arestas Múltiplas						
				very	strong	bridge			ordinary	sporadic
Comunidade	hub	regular	sporadic	strong		strong	regular	weak		
SAC	17,6%	12,5%	69,9%	5,4%	16,4%	3,4%	4,4%	17,4%	17,9%	35,0%
DAC	21,2%	12,0%	66,8%	9,9%	20,9%	9,2%	6,6%	17,8%	11,0%	24,6%
CHI	21,8%	13,0%	65,2%	10,7%	21,7%	10,1%	8,1%	19,6%	11,9%	18,0%
CIKM	35,7%	17,4%	46,9%	13,6%	22,1%	12,5%	9,0%	18,1%	11,9%	12,7%
MMSys	26,2%	14,1%	59,6%	12,6%	11,6%	19,3%	7,9%	18,4%	12,0%	18,3%
SIGCSE	24,1%	12,6%	63,3%	13,9%	22,3%	11,5%	7,3%	17,1%	9,0%	19,0%
KDD	32,3%	16,5%	51,2%	11,3%	21,4%	11,4%	7,5%	18,4%	11,8%	18,1%
SIGIR	37,3%	14,4%	48,3%	18,2%	20,8%	15,6%	9,7%	17,2%	7,7%	10,9%
SIGMOD	32,2%	17,6%	50,2%	10,8%	21,5%	12,2%	8,2%	18,5%	12,6%	16,3%
CCS	29,4%	18,8%	51,8%	7,4%	18,9%	7,7%	8,9%	25,1%	16,8%	15,2%
SIGCOMM	29,6%	18,5%	51,9%	10,2%	21,1%	11,1%	7,2%	20,3%	14,5%	15,6%
ICSE	22,8%	14,7%	62,5%	5,5%	17,2%	5,9%	6,4%	22,0%	17,5%	25,5%
SIGUCCS	12,4%	13,1%	74,5%	5,8%	13,6%	3,0%	6,5%	15,7%	21,0%	34,4%
STOC	41,3%	19,5%	39,2%	17,7%	17,3%	22,6%	14,5%	16,0%	7,6%	4,4%
SIGMETRICS	35,0%	18,4%	46,6%	10,9%	21,0%	11,3%	8,6%	22,6%	13,5%	12,1%
SIGGRAPH	17,2%	12,8%	70,1%	5,5%	17,4%	6,9%	5,6%	20,5%	14,5%	29,6%
ISCA	31,4%	16,5%	52,1%	6,5%	17,3%	7,3%	5,6%	16,9%	15,9%	30,5%
MobiCom	28,5%	16,2%	55,3%	9,4%	20,6%	7,8%	6,7%	18,9%	14,1%	22,5%
PODC	42,1%	17,4%	40,5%	14,9%	19,3%	18,7%	12,7%	19,7%	7,6%	7,1%
POPL	30,2%	21,1%	48,7%	9,4%	19,3%	8,1%	9,2%	25,9%	18,6%	9,6%
SIGDOC	14,6%	12,5%	72,9%	5,5%	15,9%	4,5%	5,1%	16,8%	14,9%	37,3%
MICRO	36,2%	16,2%	47,6%	9,6%	20,3%	10,5%	8,4%	19,5%	14,4%	17,3%
ISSAC	25,7%	15,1%	59,2%	16,6%	22,5%	13,1%	8,3%	16,3%	11,2%	11,9%
HSCC	22,4%	10,2%	67,3%	7,3%	24,3%	8,2%	4,2%	14,3%	12,8%	28,8%
Média	27,0%	15,2%	57,8%	9,9%	19,5%	9,7%	7,3%	18,3%	13,7%	21,5%
Mediana	27,1%	14,9%	57,2%	9,7%	20,7%	9,6%	7,3%	17,9%	13,8%	18,5%
Desvio Padrão	8,5%	2,8%	10,8%	4,2%	2,8%	4,8%	2,8%	2,8%	3,8%	10,8%
Rede Completa	18,8%	12,2%	69,0%	11,2%	20,3%	10,7%	7,8%	18,6%	12,4%	19,1%

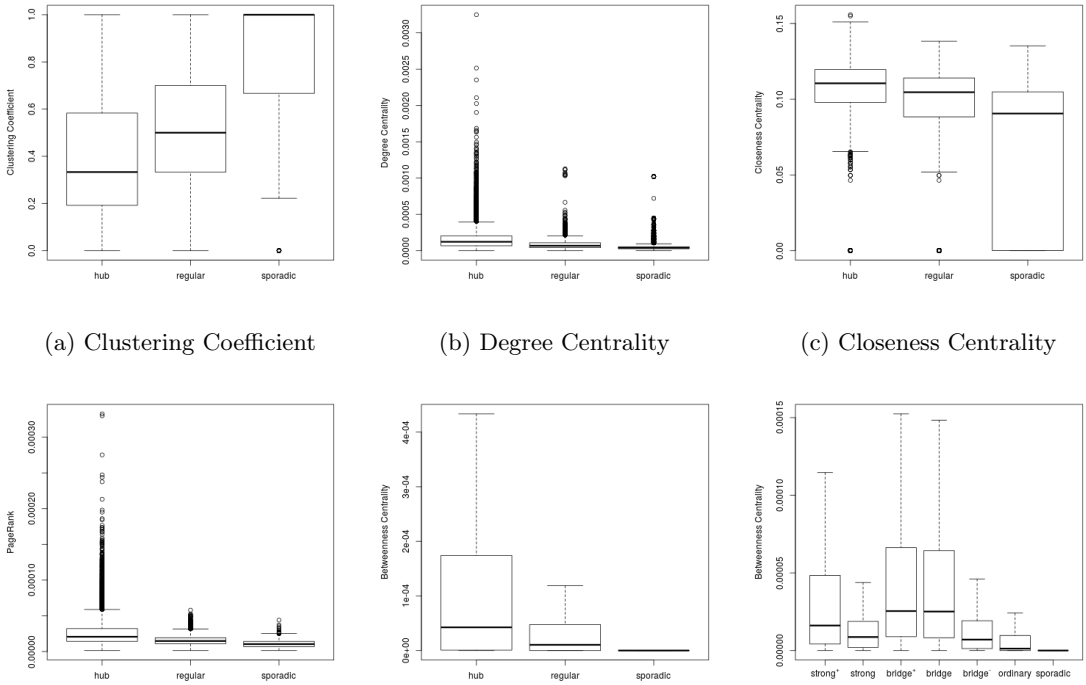
A segunda parte da tabela destaca que, quando se desconsidera as classes *ordinary* e *sporadic*, a maioria dos relacionamentos "carrega" algum tipo de informação (média de 64,4%), demonstrando um forte vínculo social nodo-atributo. Por outro lado, em média, 13,7% das relações não são representativas e 21,5% são corriqueiras. Novamente, há comunidades específicas com comportamento único como, por exemplo, SIGIR com a maior presença de arestas da classe *very strong* (18,2% das múltiplas arestas). As comunidades SIGUCCS, SIGDOC e CCS também se destacam por terem um número expressivo de arestas corriqueiras, reforçando a existência de um fraco vínculo entre seus membros. Outro destaque é a composição das arestas do tipo *regular bridge* que são as menos representativas, exceto nas comunidades STOC e PODC. A rede completa (última linha) segue a tendência apontada pelas médias de todas as comunidades.

5.2 Classificação versus Propriedades de Redes

Antes de apresentarmos nossos resultados, destacamos que as distribuições dos valores não passaram nos testes de normalidade. Desta forma, avaliamos a distinção estatística[5] entre as classes par a par por meio do teste não paramétrico Mann-Whitney-Wilcoxon e entre todas as classes por meio de sua extensão dada pelo teste de Kruskal-Wallis, conforme descritos por Hollander et al. [2013].

Classificação dos Nodos (Figura 1(a-e)). A Figura 1(a) evidencia uma característica interessante da classe *sporadic* que tende a ser muito dependente de sua vizinhança, enquanto nodos dos tipos *regular* e *hub* tendem a diversificar seus relacionamentos. Na Figura 1(b), confirma-se que quanto maior é o grau de um nodo, maior a sua tendência de ser um *hub*. Os demais gráficos (Figura 1(c-e)) demonstram uma correlação forte entre a classificação e as propriedades de rede das classes. Nodos da classe *hub* tendem a ser mais centrais e próximos aos demais (Figura 1(c)), têm maior importância na rede (Figura 1(d)) e possuem maior fluxo de informação (Figura 1(e)). Todas as distribuições são estatisticamente diferentes de acordo com os testes de Kruskal-Wallis e Mann-Whitney-Wilcoxon.

[5]Todos os experimentos foram realizados com o nível de significância $\alpha = 0,05$.



(a) Clustering Coefficient (b) Degree Centrality (c) Closeness Centrality

(d) PageRank (e) Betweenness Centrality (f) Betweenness Centrality

Fig. 1: Distribuição de diversas propriedades de rede para nodos (a-e) e arestas (f) de acordo com suas classificações. Os *outliers* foram suprimidos de (e) e (f) para uma melhor visualização.

Classificação das Arestas (Figura 1(f)). Os resultados mostram as arestas das classes *strong bridge* (bridge$^+$) e *regular bridge* (bridge) como as mais importantes, em concordância com o conceito da força dos relacionamentos com troca de informação (*brokerage*). Baseando também no conceito de detenção de conhecimento (*closure*), as classes *very strong* (strong$^+$) e *strong* também se destacam. As arestas das classes *ordinary* e *sporadic* possuem os menores valores, uma vez que não são consideradas como interações usadas para transferência de informação (i.e., *social ties* [Granovetter 1973]). As distribuições por classe são estatisticamente diferentes de acordo com os testes de Kruskal-Wallis, já o teste Mann-Whitney-Wilcoxon não diferencia as classes *strong bridge* e *regular bridge*.

5.3 Análise de Sensibilidade

Dois fatores sensíveis no processo de classificação é o tempo de existência do nodo no grafo e o poder discriminativo do conjunto de atributos a ele vinculado.

Tempo de Existência. O teste de sensibilidade a seguir consiste em investigar a robustez da nossa abordagem para diferenciar nodos com tempos de existência similares. Para isso, dividimos os nodos nas seguintes faixas temporais anuais: $[1, 5)$, $[5, 10)$, $[10, 15)$ e $[15, \infty)$. De acordo com o teste de Kruskal-Wallis, nossa abordagem distinguiu as distribuições de todas as métricas de rede por classes para todas as faixas temporais. Já o teste Mann-Whitney-Wilcoxon não diferenciou as distribuições das classes *hub* e *regular* para as métricas *betweenness* e *clustering coefficient* para a faixa $[1, 5)$.

Seleção de Atributos Estatisticamente Válidos. O Algoritmo 1 de construção de Γ considera que todos os atributos contidos nas arestas são relevantes (Seção 3). De fato, se um atributo é associado em várias oportunidades, então pode-se inferir sua importância. No entanto, um rigor estatístico pode ser adicionado ao processo de forma a excluir atributos que, mesmo aleatoriamente distribuídos,

são erroneamente atribuídos como relevantes. Assim, uma etapa adicional consiste em tornar a função Φ, que associa cada aresta e a um conjunto de atributos a, em uma associação aleatória Φ'. Em seguida, obtemos Γ a partir de diferentes instâncias Φ' para medir a probabilidade de cada atributo i ter sido erroneamente classificado como sendo relevante. Por fim, excluímos de $\mathcal{A}$ os atributos que foram destacados como relevantes com probabilidade significativamente superior ao nível de significância α. Ambas as configurações (sem exclusão e com a etapa de exclusão de atributos que não são estatisticamente válidos quando aleatoriamente distribuídos) possuem distribuições estatisticamente equivalentes[6]. Na prática, esta etapa elimina informação natural de evolução da rede e, portanto, não é acoplada ao processo de classificação.

6. CONCLUSÕES

Neste trabalho exploramos o papel dos atores em uma rede social acadêmica para classificar suas interações dinâmicas. Para isso, consideramos a importância do vínculo de atributos ao longo do tempo nesse processo de classificação. Tal classificação foi confrontada com métricas de grafos amplamente usadas na literatura. Como resultado, as distribuições se mostraram estatisticamente diferentes e em concordância com o significado social esperado. Além disso, foi mostrada a robustez da classificação ao lidar com atributos estatisticamente válidos e também ao considerar diferentes tempos de existência na rede. Como trabalhos futuros, pretendemos adaptar a noção de classificação social proposta para mensurar a influência dos nodos, bem como para uso no problema de detecção de comunidades.

REFERENCES

Adamic, L. A. and Adar, E. Friends and neighbors on the web. *Social Networks* 25 (3): 211–230, 2003.

Alves, B. L., Benevenuto, F., and Laender, A. H. F. The Role of Research Leaders on the Evolution of Scientific Communities. In *Proc. of the 22nd Int'l Conf. on the World Wide Web (Comp. Volume)*. pp. 649–656, 2013.

Barabási, A.-L. Scale-free networks: a decade and beyond. *Science* 325 (5939): 412–413, 2009.

Brandão, M. A. and Moro, M. M. The strength of co-authorship ties through different topological properties. *Journal of the Brazilian Computer Society* 23 (1): 5, 2017.

Burt, R. S. *Brokerage and closure: An introduction to social capital*. Oxford University Press, 2005.

Easley, D. and Kleinberg, J. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

Freire, V. P. and Figueiredo, D. R. Ranking in collaboration networks using a group based metric. *Journal of the Brazilian Computer Society* 17 (4): 255–266, Nov, 2011.

Gilbert, E. and Karahalios, K. Predicting tie strength with social media. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. pp. 211–220, 2009.

Granovetter, M. S. The Strength of Weak Ties. *American Journal of Sociology* 78 (6): 1360–1380, 1973.

Hollander, M., Wolfe, D. A., and Chicken, E. *Nonparametric statistical methods*. Vol. 751. John Wiley & Sons, 2013.

Leão, J. C., Brandão, M. A., Vaz de Melo, P., and Laender, A. H. F. Who is really in my social circle? Mining Social Relationships to Improve Detection of Real Communities (to appear). *Journal of Internet Services and Applications*, 2018.

Newman, M. E. pp. 337–370. In , *Who Is the Best Connected Scientist? A Study of Scientific Coauthorship Networks*. Springer Berlin Heidelberg, pp. 337–370, 2004.

Shah, N., Beutel, A., Hooi, B., Akoglu, L., Gunnemann, S., Makhija, D., Kumar, M., and Faloutsos, C. EdgeCentric: Anomaly Detection in Edge-Attributed Networks. In *In Proc. of the IEEE 16th International Conference on Data Mining Workshops*. pp. 327–334, 2016.

Silva, T. H. P., Rocha, L. M., Silva, A. P. C., and Moro, M. M. 3c-index: Research Contribution across Communities as an Influence Indicator. *Journal of Information and Data Management* 6 (3): 192, 2015.

Vaz de Melo, P. O. S., Viana, A. C., Fiore, M., Jaffrès-Runser, K., Mouël, F. L., Loureiro, A. A. F., Addepalli, L., and Chen, G. RECAST: Telling Apart Social and Random Relationships in Dynamic Networks. *Perform. Eval.* vol. 87, pp. 19–36, 2015.

Watts, D. J. The "New" Science of Networks. *Annual Review of Sociology* vol. 30, pp. 243–270, 2004.

[6]Os gráficos para as configurações desta etapa foram omitidos devido à limitação de páginas, mas ressalta-se uma melhor distinção entre as classes quando há exclusão de tais atributos não discriminativos.