STACY: Um Novo Algoritmo para Automaticamente Classificar a Força dos Relacionamentos ao Longo dos Anos

Michele A. Brandão, Pedro O. S. Vaz de Melo, Mirella M. Moro

[1]Universidade Federal de Minas Gerais – Belo Horizonte – MG

{micheleabrandao,olmo,mirella}@dcc.ufmg.br

Abstract. *Understanding the relationships between people in a social network and measuring their strength over time are interesting problems with distinct applications. Here, we propose a new algorithm (STACY) to automatically classify tie strength in eight different classes by considering the temporal aspect. Our results show: such classes represent different behaviors, and STACY identifies strong relationships that persist more than the ones classified by a state of the art algorithm.*

Resumo. *Entender relacionamentos entre pessoas em uma rede social e medir suas forças ao longo do tempo são problemas interessantes com várias aplicações. Aqui, propomos um novo algoritmo (STACY) para automaticamente classificar a força dos relacionamentos em oito diferentes classes considerando o aspecto temporal. Os resultados mostram que tais classes representam e distinguem relacionamentos, e STACY é capaz de identificar relacionamentos fortes que persistem mais do que os classificados por um algoritmo do estado-da-arte.*

1. Introdução

Tempo é um fator fundamental na caracterização da natureza e força dos relacionamentos. Por exemplo, conhecidos podem se tornar amigos (e vice-versa). Relações que variam ao longo do tempo podem ser modeladas como uma rede social temporal, ou grafo temporal, em que cada nó é uma pessoa e há uma aresta entre dois nós em um dado tempo se eles compartilharam qualquer relação naquele tempo. Abordagens que focam em grafos agregados estáticos (não-temporais) dão o mesmo grau de importância para todas as interações anteriores. No entanto, os relacionamentos mais recentes são geralmente mais representativos da classe da relação [Gilbert and Karahalios 2009]. Se em grafos estáticos esses aspectos temporais são agregados e, portanto, escondidos, em grafos temporais eles estão naturalmente presentes, servindo como um modelo apropriado para redes dinâmicas.

No entanto, a computação de propriedades temporais de redes sociais é um desafio. Por exemplo, o coeficiente de agrupamento de uma rede no tempo t_1 não é necessariamente o mesmo no tempo t_2, pois as interações podem aparecer ou desaparecer ao longo do tempo. Além disso, a precisa ordenação temporal das arestas influencia essencialmente a noção de adjacência dos nós e acessibilidade em tais redes [Nicosia et al. 2013]. Portanto, conceitos e métricas projetados e aplicados à análise de redes estáticas devem ser adaptados e estendidos a redes que variam em função do tempo. A força do relacionamento (também chamada força dos laços) é um desses conceitos, definido originalmente como uma fusão do tempo da relação, da força emocional, da intimidade e dos serviços recíprocos que representam um relacionamento entre pessoas [Granovetter 1973].

Neste artigo, propomos um novo algoritmo intitulado *STACY - Strength of Ties Automatic-Classifier over the Years*. *STACY* usa recursos de redes sociais para classificar a força dos relacionamentos em oito classes (*strong*, *bridge+*, *bridge*, *transient*, *periodic*, *bursty*, *weak* e *random*). Ele baseia-se em um algoritmo existente – RECAST (*Random rElationship ClASsifer sTrategy*) [Vaz de Melo et al. 2015], que foi aplicado para medir a força de relacionamentos em redes móveis, e não redes sociais de co-autoria que são o nosso foco. Aqui, também melhoramos o desempenho de tal algoritmo (agora denominado *fast-RECAST*) e comparamos os resultados com os gerados pelo *STACY*.

Nesta pesquisa, vemos a força de um relacionamento como a probabilidade de sua (re) aparência no futuro. Estimamos essa probabilidade usando três características das arestas das redes sociais relacionadas à força do relacionamento (persistência das arestas, sobreposição de vizinhos e frequência de co-autoria). Ademais, contrastamos os resultados estimando essa probabilidade com persistência das arestas e sobreposição de vizinhos (ambas são consideradas em fast-RECAST). Essas propriedades captam a regularidade da interação e a similaridade entre os indivíduos envolvidos nessa interação. Nosso principal objetivo é verificar se as atuais definições da força dos relacionamentos mantêm-se para redes sociais temporais. Para isso, analisamos o dinamismo da força dos relacionamentos observando a persistência e a transformação ao longo do tempo.

A seguir, a Seção 2 apresenta os trabalhos relacionados, e a Seção 3 descreve conceitos e o RECAST original. A Seção 4 detalha a força dos relacionamentos, o RECAST com multiprocessamento e novo algoritmo *STACY*. A Seção 5 analisa persistência e transformação dos relacionamentos, e a Seção 6 conclui este trabalho.

2. Trabalhos Relacionados

A força dos relacionamentos pode ser calculada considerando propriedades topológicas e/ou semânticas na rede social [Alves et al. 2016, Castilho et al. 2017]. As propriedades topológicas capturam características estruturais do grafo que constitui a rede social [Zaki and Meira Jr 2014]. Por exemplo, Brandão e Moro [2015] usam sobreposição de vizinhos para medir a força dos relacionamentos em redes de co-autoria. Por outro lado, as propriedades semânticas captam características não estruturais de nós e arestas em redes sociais. Por exemplo, Gilbert e Karahalios [2009] definem relacionamentos no Facebook considerando o histórico de interações. Por fim, também é possível combinar ambos os tipos de propriedades. Por exemplo, Zignani et al. [2016] usam propriedades de topológicas do grafo e temporais (semânticas) para prever a força dos relacionamentos.

Uma importante propriedade semântica é o aspecto temporal. Mesmo com tanta pesquisa em redes sociais, a combinação de força dos relacionamentos e aspectos temporais ainda não foi amplamente explorada. Por exemplo, Karsai et al. [2014] usa força de empate para caracterizar o impacto de interações heterogêneas e variáveis no tempo sobre a propagação de rumores. Eles consideram a evolução temporal da força dos laços, mas não propõem uma nova maneira de medir essa propriedade incluindo o tempo. Por outro lado, Kostakos [2009] e Nicosia et al. [2013] propõem um conjunto de propriedades de grafos que consideram o aspecto temporal em seu cálculo. Tais estudos mostram que muitas dessas propriedades precisam ser calculadas de forma diferente das redes estáticas.

Um problema relacionado é como definir relacionamentos fortes e fracos em redes temporais. Por exemplo, Laurent et al. [2015] definem laços fortes como interações

frequentes que conectam os nós intra-comunidades e modelam a estrutura da rede localmente, ao passo que os laços fracos são interações infrequentes situadas entre comunidades e mantêm a estrutura de rede globalmente conectada. Karsai et al. [2014] consideram tanto a quantidade de interações como o tempo das interações para definir a força dos relacionamentos. Então, os laços fortes representam interações frequentes ao longo do tempo entre pares de indivíduos, enquanto laços fracos ocorrem apenas ocasionalmente. Por outro lado, Nicosia et al. [2013] definem dois nós i e j como fortemente conectados se estiverem em uma relação não simétrica (i está temporariamente conectado a j, mas não vice-versa), enquanto que estão fracamente conectados se estão em uma relação simétrica (ambos i está temporariamente conectado a j, e j está temporariamente conectado a i).

Neste artigo, consideramos o conceito de laços fortes e fracos para redes sociais temporais baseados na ideia de Karsai et al. [2014], ou seja, um laço forte persiste ao longo do tempo, e um laço fraco ocorre esporadicamente. No entanto, Karsai et al. [2014] caracterizam a força dos relacionamentos com base em uma única janela temporal. Já nós verificamos experimentalmente se a janela temporal é um fator para caracterizar a força dos relacionamentos, analisando a persistência e a transformação dos relacionamentos ao longo do tempo com o nosso novo algoritmo *STACY*. Então mostramos que, de fato, a força dos relacionamentos é muito sensível à janela de tempo usada para calculá-la.

3. Conceitos Fundamentais

Nesta seção, descrevemos modelos para redes sociais temporais e o RECAST original.

Modelos para Redes Sociais Temporais. Inicialmente, definimos formalmente um modelo para redes sociais temporais. Em vez de propor um novo modelo, utilizamos a ideia de Vaz de Melo et al. [2015], que consiste em modelar uma rede temporal para estudar redes móveis. Seguindo tal ideia, associamos uma hora de início e uma duração a cada co-autoria. Em seguida, uma *rede social de co-autoria temporal* é modelada como um grafo $G_k(\mathcal{V}_k, \mathcal{E}_k)$ em que o tempo é discretizado em etapas de duração Δ, a duração $\delta = 1\ Ano$, e k é o período de tempo em que ocorre uma co-autoria. O conjunto de nós $\mathcal{V}_k$ é formado por todos os nós da rede durante o k-th tempo, e o conjunto de arestas $\mathcal{E}_k$ é composto de co-autorias durante o mesmo período de tempo. Assim, há uma aresta em $\mathcal{E}_k$ entre dois nós i e j com $i, j \in \mathcal{V}_k$, se i e j são co-autores de uma publicação durante o tempo k. Uma representação da variação temporal em redes de co-autoria pode ser definida por um grafo temporal acumulado $G_t(\mathcal{V}_t, \mathcal{E}_t)$, onde $G_t = G_1 \cup G_2 \cup ... \cup G_t$. Então, $\mathcal{V}_t$ e $\mathcal{E}_t$ são o conjunto de todos os nós e arestas nas redes, respectivamente, no período de tempo 0 a t. Como G_t acumula todas as co-autoria dos conjuntos de dados e evolui ao longo do tempo, tal grafo agregado contém relacionamentos sociais e aleatórios. Também de acordo com Vaz de Melo et al. [2015], é necessária uma versão aleatória G_t^R do grafo temporal acumulado G_t para analisar os padrões dessa rede. O grafo aleatório deve ter características topológicas da rede social semelhantes ao grafo G_t (número de nós, arestas e distribuição dos graus), e os nós estão conectados de forma diferente de G_t.

O RECAST Original. Seguindo a descrição do modelo, resumimos a implementação original do algoritmo, chamado RECAST [Vaz de Melo et al. 2015]. Uma contribuição deste trabalho é modificá-lo para medir a força de laços em grandes redes sociais temporais. Escolhemos RECAST por ser o único que atribui classes diferentes à força dos relacionamentos em redes temporais. Esse algoritmo diferencia interações aleatórias das

sociais (*amigos* - chamados como fortes, *pontes* e *conhecidos* - chamados de fracos). RECAST implementa o modelo descrito anteriormente construindo tanto G_t como G_t^R. Dois algoritmos são necessários para gerar G_t^R de G_t: `RND` e `T-RND` [Vaz de Melo et al. 2015]. Dado um grafo $G(\mathcal{V}, \mathcal{E})$, $\texttt{RND}(G)$ retorna um grafo aleatório $G_t(\mathcal{V}^R, \mathcal{E}^R)$ com o mesmo número de nós, arestas e distribuição de graus que G. A única diferença entre G e G^R é a conexão entre os nós, que é o foco de nosso estudo. Portanto, `RND` atribui uma aresta entre nós i e j com probabilidade $p_{i,j} = (d_i \times d_j)/\sum_{k=1}|V|d_k$, em que a distribuição de graus é $D = (d_1, d_2, ..., d_n)$ de G com n nós. O segundo algoritmo `T-RND` é uma extensão de `RND` e gera grafos aleatórios para redes temporais G_t. Assim, `T-RND` $(G_1 \cup G_2 \cup ... \cup G_t)$ recebe um conjunto de grafos de eventos consecutivos G_t e retorna um grafo temporal aleatório G_t^R. Esse algoritmo constrói G_t^R executando `RND` em cada grafo de evento G_t e depois acumulando-o como $G_t^R = \texttt{RND}(G_1) \cup \texttt{RND}(G_2) \cup ... \cup \texttt{RND}(G_t)$. Em resumo, tanto `RND` quanto `T-RND` reproduzem aleatoriamente o número total de co-autores com autores (no nosso contexto) distintos que cada pessoa tem em um instante de tempo.

4. Algoritmos para Medir a Força dos Relacionamentos

Nesta seção, revisitamos o conceito da força dos relacionamentos (Seção 4.1). Também propomos fast-RECAST (Seção 4.2) e *STACY* (Seção 4.3).

4.1. Revisitando o Conceito da Força dos Relacionamentos

Dado um grafo temporal $G_k(\mathcal{V}_k, \mathcal{E}_k)$, onde k é o intervalo de tempo em que ocorre uma co-autoria, um relacionamento (i, j) é provável que seja forte se estiver presente em G_k para a maioria dos valores de k. Já o relacionamento (i, j) é provável que seja fraco se estiver presente em G_k para alguns valores de k. Em outras palavras, é provável que os relacionamentos fortes persistam ao longo do tempo, e os fracos provavelmente ocorrem esporadicamente. Outra característica de um relacionamento forte (i, j) é que provavelmente i e j têm muitos vizinhos em comum. Como discutido anteriormente, os nós que têm muitos vizinhos em comum são mais propensos a persistirem ao longo do tempo.

Dadas estas duas características, fast-RECAST agrupa os relacionamentos em quatro classes: *strong* (amigos), *weak* (conhecidos), *bridges* e *random*. Cada classe define um nível de força dos laços: *strong* são relacionamentos que persistem ao longo do tempo e compartilham muitos vizinhos em comum; *weak* não persistem ao longo do tempo, mas compartilham muitos vizinhos em comum; *bridges* persistem ao longo do tempo, mas não possuem muitos vizinhos em comum; e *random* não persistem ao longo do tempo e compartilham poucos vizinhos em comum. Utilizando essas quatro classes de relacionamentos, investigamos se a força dos laços se transformam ao longo do tempo. Com essa análise, somos capazes de aprofundar em redes sociais temporais e responder a perguntas como: relacionamentos *strong* são mais propensos a permanecer *strong* no futuro? Ou relacionamentos *weak* são mais propensos a tornar-se *strong* ou *random*? Além disso, considerando uma terceira propriedade, frequência de co-autoria, um relacionamento forte (i, j) é provável que i e j tenham uma alta frequência de co-autoria. *STACY* usa essas três características para classificar relacionamentos em oito classes (Seção 4.3).

4.2. RECAST com Multiprocessamento

A construção de G_t^R por meio de `T-RND` aumenta a complexidade de RECAST para $O(t \times (|\mathcal{V}_t| + |\mathcal{E}_T^R|))$. Dessa forma, aplicamos um módulo Pool de multiprocessamento do Python

Algoritmo 1 fast-RECAST: um código paralelizado para classificar arestas de G_t em aleatórias e sociais – strong, weak ou bridge.

Require: $p_{rnd} \geq 0$
1: **return** $class(i,j)\ \forall (i,j) \in U_t E_t$
2: Construir G_t^R e o conjunto $\texttt{RND}(G_1), ..., \texttt{RND}(G_t)$ utilizando `T-RND` com **pool.map_async**
3: Obter $\overline{F}_{to}(x)$ e $\overline{F}_{per}(x)$ de G_t^R utilizando **pandas dataframe**
4: Obter $\overline{x}_{to} | \overline{F}_{to}(\overline{x}_{to})$ e $\overline{x}_{per} | \overline{F}_{per}(\overline{x}_{per}) = p_{rnd}$ com **pool.apply_async**
5: **for all** arestas$(i,j) \in E_t$ **do**
6: **if** $per(i,j) > \overline{x}_{per}$ e $to(i,j) > \overline{x}_{to}$ **then**
7: $class(i,j) \longleftarrow Strong$
8: **else if** $per(i,j) > \overline{x}_{per}$ e $to(i,j) \leq \overline{x}_{to}$ **then**
9: $class(i,j) \longleftarrow Bridges$
10: **else if** $per(i,j) \leq \overline{x}_{per}$ e $to(i,j) > \overline{x}_{to}$ **then**
11: $class(i,j) \longleftarrow Weak$
12: **else**
13: $class(i,j) \longleftarrow Random$

(módulo baseado em processos de comunicação para escrever programas concorrentes[1]) em tal etapa do RECAST, a fim de reduzir a sua complexidade. Chamamos essa nova versão do algoritmo com multiprocessamento de fast-RECAST. A ideia é que mais de um grafo aleatório G_t^R é construído de cada vez em um computador multi-core. Assim, o novo custo computacional é $O(\frac{t}{p} \times (|\mathcal{V}_t| + |\mathcal{E}_t^R|))$, onde p é o número de processos. Após construir G_t^R, a complexidade da classificação é $O(|E_t^R| \times |\mathcal{V}_t|)$, onde $O(|V_t|)$ é o custo de computar as duas características de uma aresta na rede social. Também adicionamos um módulo Pool de multiprocessamento do Python para chamar as funções para calcular a persistência das arestas e a sobreposição de vizinhos dos grafos agregados. Ambos os recursos são computados em paralelo e de forma assíncrona.

O Algoritmo 1 resume o código de fast-RECAST[2] com multiprocessamento (linhas 2 e 4) e uma otimização na utilização da memória, aplicando Pandas dataframe do python para armazenar os grafos antes de processá-los (linha 3). Por ser redes de co-autoria, renomeamos as arestas de *friend* para *strong* e *acquaintance* para *weak*.

4.3. STACY - *Strength of Ties Automatic-Classifier over the Years*

Agora propomos um algoritmo para classificar automaticamente a força dos relacionados chamado *STACY - Strength of Ties Automatic-Classifier over the Years*, uma melhoria do fast-RECAST que considera o peso das arestas (frequência de co-autoria) diferente de 1. Para distribuir a frequência de co-autoria no grafo aleatório G_t^R, usamos o mesmo algoritmo para distribuir o grau das arestas proposto por Miller and Hagberg [2011]. A diferença é que atribuímos um peso às arestas entre i e j com probabilidade $p_{ij} = (w_i \times w_j) / \sum_{k=1}^{|V|} w_k$ para uma distribuição de pesos $D_w = (w_1, w_2, ..., w_n)$ de G com N nós.

Estendendo a descrição de G_t na Seção 3, definimos um grafo temporal acumulado com peso $G_t^W = G_1^W \cup G_2^W \cup ... \cup G_t^W$. $\mathcal{V}_t$ e $\mathcal{E}_t$ são o conjunto de todos os nós e arestas com peso na rede social, respectivamente, no período de tempo 0 a t. Como G_t^W acumula todas as co-autorias dos conjuntos de dados e evolui ao longo do tempo, o grafo agregado possui relacionamentos sociais e aleatórios. Além disso, consideramos uma versão com

[1]Python e multiprocessamento: `docs.python.org/2/library/multiprocessing.html`
[2]Código fonte disponível em *`github.com/lab-csx-ufmg/RECAST`*

Tabela 1. Classes de relacionamentos do STACY.

Classe	Persistência das arestas	Sobreposição de vizinhos	Frequência de co-autoria
1 - strong	social	social	social
2 - bridge+	social	random	social
3 - transient	random	social	social
4 - periodic	social	social	random
5 - bursty	random	random	social
6 - bridge	social	random	random
7 - weak	random	social	random
8 - random	random	random	random

Algoritmo 2 STACY: um código paralelizado para classificar arestas com pesos de G_t^W em oito classes diferentes.

Entrada: Grafo temporal agregado com peso - G_t^W

Require: $p_{rnd} \geq 0$

1: **return** $class(i,j)\ \forall (i,j) \in U_t E_t$
2: Construir $G_t^{R,W}$ e obter $\texttt{RND}(G_1^W), ..., \texttt{RND}(G_t^W)$ usando `T-RND` com **pool.map_async**
3: Obter $\overline{F}_{to}(x)$ e $\overline{F}_{per}(x)$ e $\overline{F}_{coAfrequency}(x)$ de $G_t^{R,W}$ usando **pandas dataframe**
4: Obter $\overline{x}_{to}|\overline{F}_{to}(\overline{x}_{to})$ e $\overline{x}_{per}|\overline{F}_{per}(\overline{x}_{per})$ e $\overline{x}_{coAfrequency}|\overline{F}_{coAfrequency}(\overline{x}_{coAfrequency})$ = p_{rnd} com **pool.apply_async**
5: **for all** arestas$(i,j) \in E_t$ **do**
6: ClassificaAresta(*per*, *to*, *coAfrequency*) *//Realizado de acordo com Tabela 1*

peso $G_t^{R,W}$ do grafo temporal agregado G_t^W, que é necessário para analisar os padrões da rede. O grafo aleatório deve ter características topológicas similares ao grafo G_T^W (número de nós, arestas, e distribuição do grau), os nós são conectados de modo diferente a partir de G_T^W e o peso (frequência de co-autoria) são distribuídos aleatoriamente através das arestas. É importante ressaltar que a frequência de co-autoria de cada aresta em G_t^W é a soma de todas frequências de co-autoria em cada período de tempo.

STACY classifica as arestas em oito classes diferentes: sete sociais e uma aleatória. Essas classes de relacionamento são descritas na Tabela 1. Uma propriedade da rede com valor igual a "social" indica uma probabilidade quase zero do valor ser produzido de forma aleatória. Por outro lado, um valor da propriedade da rede é denominada "random" se há uma alta probabilidade do valor ser produzido de forma aleatória. Note que *strong* define relacionamentos fortes já que todos os valores das propriedades são "social", enquanto que *random* representa um relacionamento completamente aleatório. Além disso, as classes *bridge+* e *bridge* denotam pontes, isto é, persistem ao longo do tempo, mas têm um pequeno número de vizinhos comuns. A classe *bridge+* representa pontes com uma alta frequência de co-autoria e *bridge* com uma pequena. Ademais, a classe *transient* denota uma relação que acontece com alta intensidade (alta frequência de co-autoria e alta sobreposição de vizinhos), mas apenas em um momento específico. Já *periodic* representa um relacionamento que persiste ao longo do tempo e têm um elevado número de vizinhos comuns, mas pequena frequência de co-autoria (por exemplo, uma co-autoria entre colegas de um mesmo departamento que acontece uma vez no ano). *bursty* define uma relação com alta frequência de co-autoria, mas não persiste e não compartilha muitos vizinhos. Esta relação tende a ser isolada na rede. Por fim, *weak* representa relacionamentos fracos, pois não persistem ao longo do tempo e a frequência de co-autoria é pequena.

Como o RECAST, o único parâmetro do *STACY* é o p_{rnd} (Seção 3), que determina

Tabela 2. Informações principais sobre os conjuntos de dados.

Conjunto de dados	№ de nós	№ de arestas	Período
DBLP Articles	837,583	2,935,590	2000 to 2015
DBLP Inproceedings	945,297	3,760,247	2000 to 2015
PubMed	443,784	5,550,294	2000 to 2016
APS	180,718	821,870	2000 to 2013

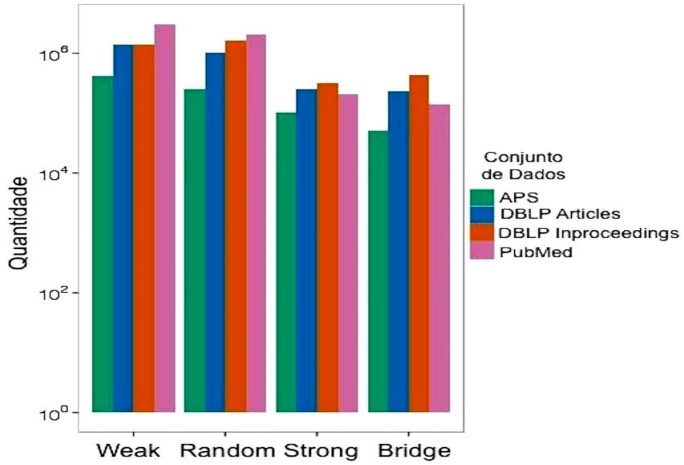


Figura 1. Pares de autores por classe gerada pelo fast-RECAST

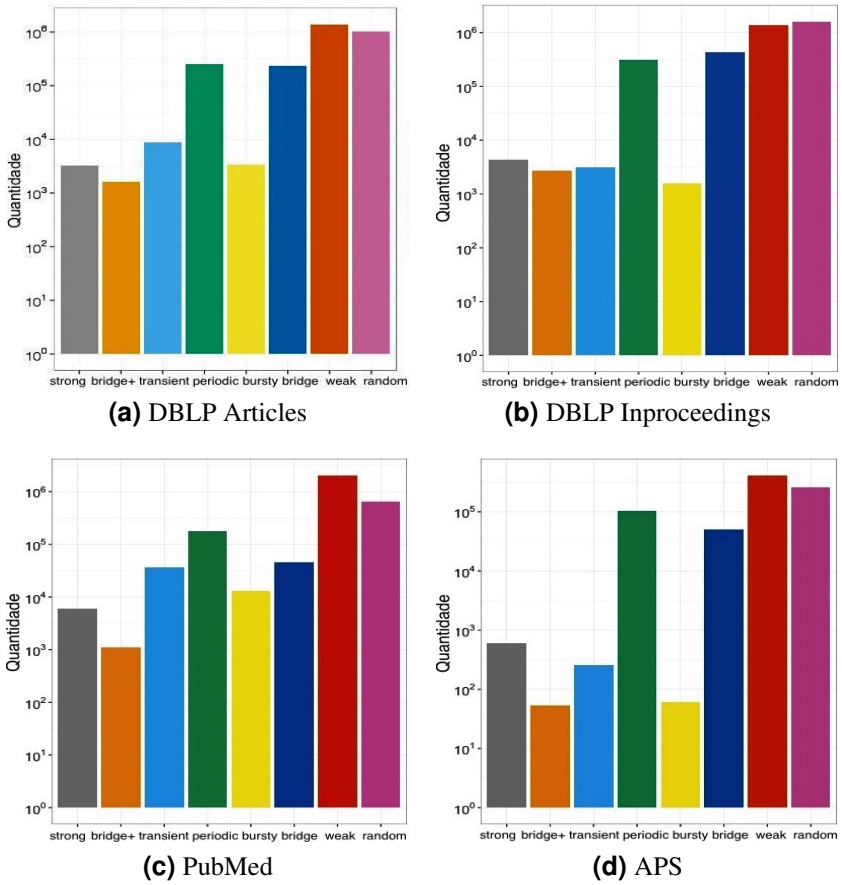
quando o valor de propriedade da rede é social ou aleatória. O Algoritmo 2 apresenta a classificação do *STACY*, paralelizada como o fast-RECAST.

5. Experimentos e Resultados

Utilizamos três conjuntos de dados para avaliar *STACY* e fast-RECAST: DBLP[3], PubMed[4] e APS[5]. A partir deles, construímos quatro redes sociais temporais cujas principais estatísticas estão na Tabela 2. Note que DBLP Inproceedings e DBLP Articles possuem a maior quantidade de nós (pesquisadores), mas PubMed tem a maior quantidade de arestas. A seguir, são detalhados três conjuntos de experimentos: aplicamos fast-RECAST e *STACY* nas redes para calcular a força dos laços; e dividimos as redes em duas janelas temporais para analisar a persistência e a transformação dos relacionamentos.

Fast-RECAST versus *STACY*. RECAST foi usado para classificar interações em redes móveis (Seção 3), mas os padrões sociais e suas características são diferentes das de redes de co-autoria. Neste trabalho, verificamos se o algoritmo é capaz de identificar os diferentes tipos de relacionamentos entre os co-autores. Também fizemos a mesma verificação para o *STACY*. Inicialmente, precisamos configurar o valor do parâmetro para p_{rnd}. Vaz de Melo et al. [2015] variou p_{rnd} em quatro ordens de magnitude e observou que o número de arestas por classe mantém-se na mesma magnitude. Portanto, o algoritmo não precisa de uma definição muito precisa do parâmetro para classificar as arestas de forma consistente. Executamos fast-RECAST e *STACY* para $p_{rnd} = 0.01$ e $p_{rnd} = 0$; como obtivemos resultados similares, apresentamos resultados apenas para $p_{rnd} = 0$.

[3]DBLP: `http://dblp.uni-trier.de`
[4]PubMed: `http://www.ncbi.nlm.nih.gov`
[5]APS: `http://www.aps.org`



(a) DBLP Articles **(b)** DBLP Inproceedings

(c) PubMed **(d)** APS

Figura 2. Pares de autores por classe gerada pelo *STACY*.

A Figura 1 apresenta a classificação das co-autorias em cada classe gerada por fast-RECAST para as quatro redes sociais. Em DBLP Articles, PubMed e APS, a maioria das co-autorias são classificadas como relacionamentos fracos, isto é, arestas com baixo (ou *aleatório*) sobreposição de vizinhos e persistência das arestas. Em tais redes, a maioria das co-autorias são mais classificadas como relacionamentos fortes (e não pontes). A exceção é DBLP Inproceedings, em que a maioria das arestas são mais atribuídas à classe aleatória e mais co-autorias são pontes ao invés de laços fracos. Uma possível explicação é que DBLP Articles, PubMed e APS possuem publicações em periódicos, enquanto que DBLP Inproceedings apenas em conferências. Conforme estudos recentes (e.g., [Lima et al. 2013]), Computação possui um comportamento incomum, no qual conferências são para publicação de ideias inovadoras e periódicos para fins de arquivamento. Desse modo, as redes de co-autoras de periódicos incluem autores que já publicaram juntos, apresentando assim relacionamentos mais fortes.

A Figura 2 tem a classificação da *STACY* para cada rede de coautoria. Similar ao fast-RECAST, a maioria dos relacionamentos são classificados como *weak* e *random*. Além disso, muitos relacionamentos são classificados como *periodic* e *bridge*. A grande quantidade de *periodic* revela que os pesquisadores da mesma comunidade (por exemplo, equipe, departamento, etc) tendem a publicar juntos com pequena frequência em um ano. Igualmente, a grande quantidade de *bridge* indica que a maioria das pontes tendem a ter uma pequena frequência de co-autoria em cada ano. Note que menos relacionamentos

Tabela 3. fast-RECAST versus *STACY* em divisão 80/20%

fast-RECAST	DBLP Articles		DBLP Inproceedings		PubMed		APS	
	80%	**20%**	**80%**	**20%**	**80%**	**20%**	**80%**	**20%**
Strong	75,128	16,083 (0.21)	136,159	19,608 (0.14)	91,143	19,555 (0.21)	45,020	30,046 (0.67)
Bridge	133,071	28,090 (0.21)	368,177	55,327 (0.15)	50,903	11,239 (0.22)	50,464	31,767 (0.63)
Weak	767,143	28,683 (0.04)	750,837	16,244 (0.02)	1,790,986	67,752 (0.04)	201,978	102,108 (0.51)
Random	931,796	76,298 (0.08)	1,340,167	69,661 (0.05)	1,021,710	63,986 (0.06)	249,711	128,479 (0.51)
	DBLP Articles		**DBLP Inproceedings**		**PubMed**		**APS**	
STACY	**80%**	**20%**	**80%**	**20%**	**80%**	**20%**	**80%**	**20%**
Strong	1,238	485 (0.39)	2,562	674 (0.26)	6,003	2,230 (0.37)	93	17 (0.18)
Bridge+	886	368 (0.41)	2,498	573 (0.23)	1,113	305 (0.27)	8	2 (0.25)
Transient	0	0	0	0	37,157	2,771 (0.07)	120	95 (0.79)
Periodic	1,070,400	64,249 (0.06)	1,149,339	53,445 (0.05)	175,179	34,215 (0.2)	58,663	12,122 (0.21)
Bursty	0	0	0	0	12,862	1,372 (0.1)	4	3 (0.75)
Bridge	834,614	84,052 (0.1)	1,440,941	106,148 (0.07)	45,419	8,718 (0.19)	36,720	6,840 (0.19)
Weak	0	0	0	0	2,042,114	76,552 (0.04)	256,564	13,908 (0.05)
Random	0	0	0	0	634,895	36,369 (0.05)	195,001	15,573 (0.08)

80% representa o passado (2000-2012 para DBLP articles e inproceedings, 2000-2013 PubMed e 2000-2010 APS) e 20% é o presente (2013-2015 para DBLP articles e inproceedings, 2014-2016 PubMed e 2011-2013 APS).

são classificados como *strong*, *bridge+*, *transient* e *bursty*. Essas quatro classes têm em comum o valor "social" para a propriedade frequência de co-autoria na rede social (as outras quatro classes têm um valor "random" para esta propriedade). Tal resultado mostra que a frequência de co-autoria é uma característica importante para medir a força dos relacionamentos, uma vez que ajuda a diferenciar as classes.

Análise da Persistência dos Relacionamentos. O objetivo aqui é investigar se os relacionamentos caracterizados com um determinado nível de força são prováveis de persistir. Para tal: (*i*) dividimos as redes em duas janelas temporais, que chamamos de *passado* e *presente*[6]; (*ii*) aplicamos fast-RECAST e *STACY* no passado; e (*iii*) verificamos se as arestas em cada classe mantêm-se na mesma classe no presente. As redes sociais foram divididas em duas janelas temporais de duas formas: uma janela temporal com 80% do timestamp inicial (*passado*) e outra com 20% do timestamp final (*presente*); e a mesma divisão para 70% (*passado*) e 30% (*presente*). Com resultados similares, a Tabela 3 apresenta apenas para a partição 80% e 20%. Os valores na coluna 80% são o número absoluto de arestas dos 80% dos anos das publicações atribuídos a cada classe. Os valores na coluna 20% são o número de arestas do passado que também estão no presente (proporções entre parênteses). Observamos que relacionamentos fortes e pontes tendem a persistir ao longo dos anos mais do que laços fracos e aleatórios.

Considerando os resultados do fast-RECAST, enfatizamos as diferenças nos resultados da rede APS nas partições 80% - 20% (e 70% - 30%). Na primeira divisão, a proporção de relacionamentos fortes e pontes do passado para o presente é muito alta, enquanto que na segunda divisão essa proporção é menor. Esse resultado pode indicar que a rede social de co-autoria da APS muda mais ao longo dos anos do que as outras redes. Outra possibilidade é que os pesquisadores da Física não alteram muito o nível de co-autoria com seus colaboradores ao longo do tempo, e esse é um padrão de pesquisadores mais recentes (observe que 80% dos dados consideram co-autoria mais recentes do que 70%). Deixamos tais análises para etapas futuras do trabalho.

[6]Pode-se ver o presente como o *timestamp* das janelas temporais.

Tabela 4. fast-RECAST e transformação dos relacionamentos: (a) DBLP Articles, (b) DBLP Inproceedings, (c) PubMed, (d) APS

(a)	Strong	Bridge	Weak	Random	Disappear
Strong	43,711 (0.11)	27,134 (0.07)	0	0	312,765 (0.82)
Bridge	14,650 (0.04)	13,874 (0.035)	0	0	361,041 (0.925)
Weak	0	0	0	0	0
Random	0	0	0	0	0
(b)	**Strong**	**Bridge**	**Weak**	**Random**	**Disappear**
Strong	34,761 (0.08)	26,411 (0.06)	0	0	351,935 (0.86)
Bridge	13,601 (0.02)	16,298 (0.024)	0	0	659,608 (0.96)
Weak	0	0	0	0	0
Random	0	0	0	0	0
(c)	**Strong**	**Bridge**	**Weak**	**Random**	**Disappear**
Strong	349 (0.02)	387 (0.02)	3,267 (0.16)	2,664 (0.13)	17,044 (0.67)
Bridge	66 (0.01)	97 (0.01)	659 (0.07)	667 (0.07)	8,643 (0.84)
Weak	10,532 (0.02)	10,425 (0.02)	94,800 (0.18)	73,039 (0.13)	346,559 (0.65)
Random	1,476 (0.01)	1,792 (0.01)	13,105 (0.06)	11,941 (0.05)	195,803 (0.87)
	Strong	**Bridge**	**Weak**	**Random**	**Disappear**
Strong	836 (0.03)	571 (0.02)	2,219 (0.09)	1,691 (0.06)	19,625 (0.8)
Bridge	450 (0.02)	421 (0.02)	918 (0.04)	910 (0.04)	19,173 (0.88)
Weak	4,013 (0.03)	2,071 (0.02)	14,185 (0.11)	7,154 (0.06)	99,844 (0.78)
Random	1,561 (0.013)	1,158 (0.01)	4,072 (0.03)	3,625 (0.03)	107,452 (0.92)

Em relação ao *STACY*, os relacionamentos fortes tendem a persistir mais que os outros em DBLP Articles, DBLP Inproceedings e PubMed nas partições 80% - 20% (e 70% - 30%). *STACY* é capaz de classificar melhor (do que o fast-RECAST) os relacionamentos fortes que persistem ao longo do tempo: um aumento de 0,18 para DBLP Articles, 0,12 para DBLP Inproceedings e 0,16 para PubMed na partição 80% - 20% (para 70% - 30%, o crescimento é ainda maior: 0,22 para DBLP Articles, 0,16 para DBLP Inproceedings e 0,22 para PubMed). A exceção é APS, em que a maioria dos relacionamentos em *transient* e *bursty* tendem a persistir ao longo do tempo. Este é um resultado inesperado, pois ambas as classes têm um valor "random" para a propriedade persistência das arestas. Analisando melhor a causa, observamos que as co-autorias em tais classes ocorrem de 2009 a 2013, ou seja, nos últimos anos das partições (os 80% incluem 2009 e 2010, e os 70% incluem 2009). Assim, o valor de persistência das arestas é pequeno, porque as co-autorias ocorrem nos anos incluídos no presente. Além disso, nenhuma aresta é classificada como *transient*, *bursty*, *weak* e *random* em DBLP Articles e DBLP Inproceedings em ambas as partições. Isso revela que, em tais redes, as co-autoria transitórias, *bursty*, fracas e aleatórias são relações recentes, porque elas são encontradas na versão completa dessas redes sociais (como mostra a Figura 2). Ademais, laços fracos e aleatórios são os que menos persistem ao longo do tempo na PubMed e APS.

Análise da Transformação das Classes dos Relacionamentos. Agora, avaliamos a quantidade de relacionamentos de uma classe no passado que continua na mesma classe (ou muda) no presente. Para análise mais uniforme, dividimos as redes em duas janelas temporais de 50% do timestamp, aplicamos fast-RECAST e *STACY* em ambas, e analisamos a transformação dos relacionamentos por meio das classes. A Tabela 4 mostra os resultados para fast-RECAST e a Tabela 5 para *STACY*. Os valores em cada coluna representam a quantidade e a proporção (entre parênteses) de laços do passado que persistem ou mudam de classe no presente (e.g., valores 43,711 e 0,11 na Tabela 4 são o número e a proporção de pares *strong* no passado que permanecem *strong* no presente).

Analisando os resultados de fast-RECAST, não há relacionamentos classificados como *weak* e *random* em DBLP Articles e DBLP Inproceedings na Tabela 4 partes (a) e (b). Isso indica que as propriedades (persistência das arestas e sobreposição de vizinhos) dessas redes têm valores altos (ou *social*). Além disso, a maioria dos relacionamentos do passado tendem a desaparecer no presente, especialmente *bridges*. Esse resultado pode

Tabela 5. *STACY* e transformação dos relacionamentos: (a) DBLP Articles, (b) DBLP Inproceedings, (c) PubMed, (d) APS

(a)	Strong	Bridge+	Transient	Periodic	Bursty	Bridge	Weak	Random	Disappear
Strong	0	1 (0.002)	0	54 (0.09)	0	19 (0.03)	0	0	549 (0.88)
Bridge+	0	0	0	8 (0.03)	0	9 (0.03)	0	0	238 (0.93)
Transient	0	0	0	0	0	0	0	0	0
Periodic	58 (1e-04)	7 (1.39e-05)	0	59,823 (0.12)	0	19,568 (0.04)	0	0	423,247 (0.84)
Bursty	0	0	0	0	0	0	0	0	0
Bridge	24 (8.9e-05)	4 (1.5e-05)	0	13,465 (0.05)	0	6,329 (0.02)	0	0	249,772 (0.92)
Weak	0	0	0	0	0	0	0	0	0
Strong	0	0	0	0	0	0	0	0	0
(b)	**Strong**	**Bridge+**	**Transient**	**Periodic**	**Bursty**	**Bridge**	**Weak**	**Random**	**Disappear**
Strong	0	0	0	28 (0.06)	0	21 (0.05)	0	0	387 (0.88)
Bridge+	0	0	0	21 (0.03)	0	7 (0.01)	0	0	596 (0.96)
Transient	0	0	0	0	0	0	0	0	0
Periodic	28 (6.79e-05)	5 (1.2e-05)	0	44,665 (0.1)	0	16,425 (0.04)	0	0	351,548 (0.85)
Bursty	0	0	0	0	0	0	0	0	0
Bridge	26 (3.8e-05)	6 (8.7e-06)	0	19,148 (0.03)	0	10,691 (0.02)	0	0	659,012 (0.95)
Weak	0	0	0	0	0	0	0	0	0
Random	0	0	0	0	0	0	0	0	0
(c)	**Strong**	**Bridge+**	**Transient**	**Periodic**	**Bursty**	**Bridge**	**Weak**	**Random**	**Disappear**
Transient	0	0	0	91 (0.14)	0	74 (0.12)	0	0	478 (0.74)
Bridge+	0	0	0	4 (0.05)	0	3 (0.03)	0	0	75 (0.91)
Transient	0	0	0	344 (0.19)	0	106 (0.06)	0	0	1348 (0.74)
Periodic	0	1 (4.1e-05)	0	4,780 (0.2)	0	2,440 (0.1)	0	0	17,192 (0.7)
Bursty	0	0	0	27 (0.05)	0	18 (0.03)	0	0	494 (0.9)
Bridge	0	0	0	473 (0.09)	0	290 (0.05)	0	0	4,675 (0.86)
Weak	35 (5.7e-05)	7 (1.1e-05)	0	137,563 (0.22)	0	62,939 (0.1)	0	0	416,963 (0.67)
Random	1 (7.2e-06)	0	0	10,216 (0.07)	0	5,854 (0.04)	0	0	123,557 (0.88)
(d)	**Strong**	**Bridge+**	**Transient**	**Periodic**	**Bursty**	**Bridge**	**Weak**	**Random**	**Disappear**
Strong	0	0	0	0	0	2 (0.3)	0	0	5 (0.7)
Bridge+	0	0	0	0	0	0	0	0	3 (1.0)
Transient	0	0	0	0	0	0	0	0	0
Periodic	0	0	0	836 (0.03)	0	569 (0.02)	2,219 (0.09)	1,691 (0.07)	19,620 (0.8)
Bursty	0	0	0	0	0	0	1 (1.0)	0	0
Bridge	0	0	0	450 (0.02)	0	421 (0.02)	918 (0.04)	910 (0.04)	19,170 (0.9)
Weak	11 (1e-04)	2 (1e-05)	0	4,002 (0.03)	0	2,069 (0.02)	14,185 (0.11)	7,154 (0.06)	99,844 (0.8)
Random	4 (3e-05)	2 (1e-05)	0	1,557 (0.01)	1 (8e-06)	1,156 (0.01)	4,071 (0.03)	3,624 (0.03)	107,452 (0.9)

ser explicado pela natureza das co-autorias, pois os pesquisadores colaboram durante um período por um objetivo comum e, em seguida, começam a colaborar com outros. Na Tabela 4, as partes (c) e (d) mostram comportamento semelhante para PubMed e APS, e a maioria dos relacionamentos tendem a desaparecer, especialmente pontes e aleatórios. Desconsiderando os relacionamentos que desaparecem, relacionamentos mais fortes e fracos tornam-se fracos ou aleatórios. Surpreendentemente, relacionamentos fracos são os que mais se mantêm na mesma classe, comparando com os outros em ambas as redes.

Para *STACY*, também não há relacionamentos *weak* e *random* em DBLP Articles e DBLP Inproceedings na Tabela 5 partes (a) e (b). Assim, a frequência de co-autoria dessas redes também possuem um valor alto (ou *social*). Os relacionamentos também não são classificados como transitórios em DBLP Articles, DBLP Inproceedings e APS (Tabela 5 parte (d)), o que revela a ausência dessas co-autorias em períodos anteriores nessas redes. Ademais, DBLP Articles e DBLP Inproceedings não têm relacionamentos *bursty*, o que indica que os relacionamentos com alta frequência de co-autoria também compartilham um grande número de vizinhos nessas redes no período coberto pelos 50% de dados (isso também é confirmado pela presença de relacionamentos em *transient*). Como fast-RECAST, a maioria dos relacionamentos também tende a desaparecer quando classificados por *STACY*. A diferença é que usando *STACY*, observamos que os relacionamentos de diferentes classes tendem a mudar para *periodic* e *bridge* ao longo do tempo, especialmente, em DBLP Articles, DBLP Inproceedings e PubMed (Tabela 5 parte (c)).

6. Conclusões

Neste artigo, propusemos o *STACY*, um algoritmo paralelo e rápido que classifica os relacionamentos de coautoria em oito classes. Também detalhamos o fast-RECAST, uma

versão paralela e mais rápida do RECAST que gera quatro classes. Tais algoritmos são úteis por serem capazes de automaticamente detectar classes de relacionamentos em redes sociais, além de diferenciar relacionamentos sociais dos aleatórios. Essas classes podem auxiliar a entender o processo de difusão de informação, melhorar algoritmos que detectam comunidades e recomendam pessoas. De fato, ao agrupar as arestas em classes utilizando redes sociais reais, analisamos o dinamismo da força dos relacionamentos ao longo do tempo. A análise da persistência revela que relacionamentos fortes e pontes tendem a persistir ao longo dos anos mais que outros tipos, confirmando nossa hipótese de que os laços fortes persistem mais. Além disso, *STACY* foi capaz de encontrar laços fortes que persistem mais que os encontrados por fast-RECAST. A análise de transformação dos relacionamentos revelou que a maioria dos laços tendem a desaparecer ao longo do tempo. Isso pode ocorrer devido à natureza das co-autorias, por exemplo, os pesquisadores tendem a publicar com alunos durante um período e quando os alunos se formam, o processo de publicação em conjunto é finalizado. Como trabalhos futuros, planejamos aplicar *STACY* em outros tipos de redes sociais, por exemplo, no GitHub que revela interações entre desenvolvedores durante o processo de desenvolvimento de softwares (estendendo [Alves et al. 2016]). Também pretendemos adicionar propriedades ao *STACY* para diferenciar relacionamentos recentes dos antigos enquanto mede a força.

Agradecimentos. Trabalho parcialmente financiado por CAPES, CNPq e FAPEMIG.

Referências

Alves, G. B., Brandão, M. A., Santana, D. M., da Silva, A. P. C., and Moro, M. M. (2016). The strength of social coding collaboration on github. In *SBBD*, pages 247–252, Salvador, Brazil.

Brandão, M. A. and Moro, M. M. (2015). Analyzing the strength of co-authorship ties with neighborhood overlap. In *DEXA*, pages 527–542, Linz, Austria.

Castilho, D., Vaz de Melo, P. O. S., and Benevenuto, F. (2017). The strength of the work ties. *Information Sciences*, 375:155–170.

Gilbert, E. and Karahalios, K. (2009). Predicting tie strength with social media. In *SIGCHI*, pages 211–220, Boston, USA.

Granovetter, M. S. (1973). The strength of weak ties. *American Joumal af Sociology*, 78(6):1360–1380.

Karsai, M., Perra, N., and Vespignani, A. (2014). Time varying networks and the weakness of strong ties. *Scientific reports*, 4.

Kostakos, V. (2009). Temporal graphs. *Physica A: Statistical Mechanics and its Applications*, 388(6):1007–1023.

Laurent, G., Saramäki, J., and Karsai, M. (2015). From calls to communities: a model for time-varying social networks. *The European Physical Journal B*, 88(11):1–10.

Lima, H., Silva, T. H., Moro, M. M., Santos, R. L., Meira Jr, W., and Laender, A. H. (2013). Aggregating productivity indices for ranking researchers across multiple areas. In *JCDL*, pages 97–106, NYC, USA.

Miller, J. C. and Hagberg, A. (2011). Efficient generation of networks with given expected degrees. In *WAW*, pages 115–126, Atlanta, USA.

Nicosia, V. et al. (2013). *Temporal Networks*, chapter Graph Metrics for Temporal Networks, pages 15–40. Springer Berlin Heidelberg, Berlin, Heidelberg.

Vaz de Melo, P. O. S. et al. (2015). Recast: Telling apart social and random relationships in dynamic networks. *Performance Evaluation*, 87:19–36.

Zaki, M. J. and Meira Jr, W. (2014). *Data mining and analysis: Fundamental concepts and algorithms*. Cambridge University Press.

Zignani, M., Gaito, S., and Rossi, G. P. (2016). Predicting the link strength of newborn links. In *WWW*, pages 147–148, Montreal, Canada.