STACY: Strength of Ties Automatic-Classifier over the Years

Michele A. Brandão[1], Pedro O. S. Vaz de Melo[2] and Mirella M. Moro[2]

[1] Instituto Federal de Minas Gerais (IFMG) – Ribeirão das Neves, MG – Brazil
[2] Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil
michele.brandao@ifmg.edu.br,{olmo,mirella}@dcc.ufmg.br

Abstract. With the evolution of Web technology and its worldwide use by regular people, there is now data about not only such people but also their relations. Database research has evolved as well to tackle the myriad of problems that arrive with such volumes of data. Here, we contribute to such a trend by proposing a new algorithm (STACY) to automatically classify tie strength (an intrinsic property of relationships) considering time. We show that each class has singular and different behavior, and analyze them over co-authorship networks. Also, STACY identifies strong relationships that persist more than the ones classified by a state of the art algorithm. Finally, we derive a computational model from STACY that is able to automatically identify relationships classes with low computational cost.

Categories and Subject Descriptors: Information Systems [**World Wide Web**]: Social Networks

Keywords: Social Networks, Tie Strength, Co-authorship Networks

1. INTRODUCTION

With the evolution of Web technology and its worldwide use by regular people (as opposed to its initial crowd composed of scientists only), we now have actual, physical data not only about such people but also about their relationships. For the first time, database researchers can gathered and analyze such data and help to understand the relationships as well. Such analyses are commonly performed over Social Networks (SN), which are complex structures that describe individuals and their relationships in any social context. Formally, they can be mapped to graphs where nodes (vertices) represent the individuals, and edges connect pairs of individuals who share a relationship. Then, properties can be extracted from the graph as well as metrics can be applied to nodes/edges to better understand the individuals' social behavior [Barabási 2016].

There are two ways of building a SN. One is to collect data from online social platforms, such as Facebook, Twitter or GitHub [Bigonha et al. 2012; Brandão and Moro 2017]. The other is to build it from data that can implicitly express relationships, such as a movie database (e.g., [Viana et al. 2016]) or a digital library (e.g., [Brandão and Moro 2015]). Here, we consider the second type and build *academic social networks*, for researchers and their academic relations. Nonetheless, our contributions are context-independent and could be easily applied to other networks.

Having the SN, the next step is to analyze it. Among possible properties, one central aspect of more complex analysis is *the strength of the ties*, as pairs of individuals have stronger or weaker connections depending on the degree of relationship. Such degree (or tie strength) may be defined from Granovetter's theory: ties are *weak* when they serve as bridges in the network by connecting users from different groups, and *strong* when they link individuals in the same group (community) [Granovetter 1973]. The analysis of tie strength has different purposes depending on the context.

In the academic context, studying the strength of co-authorship ties may reveal new insights about the dynamics of research collaborations, and any application based on co-authorship patterns may benefit. For instance, new strength metrics could help works on measuring research productivity [Chan et al. 2016], ranking researchers [Freire and Figueiredo 2011], name disambiguation [Levin and Heuser 2010], and recommending collaborations [Lopes et al. 2010]. Also, properly measuring co-authorship ties may help to identify which collaborations are more influent. Aggregating the time dimension to such analysis takes it to the next level, allowing to study team formation dynamics for example.

Measuring the strength of ties in a properly way is a relevant problem and this work tackles it. Indeed, the main goal of this article is to automatically classify such strength. Then, such goal can be divided in three research questions: *(i)* How to properly measure the strength of ties?; *(ii)* What are the characteristics of the ties according to their strength?; and *(iii)* How are the dynamism of the ties over time?. To characterize tie strength, we build up on an existing algorithm (RECAST – *Random rElationship ClASsifier sTrategy* by [Vaz de Melo et al. 2015]) that returns four classes of tie strength over mobile networks. Specifically, we double improve such algorithm by: speeding up its performance (with fast-RECAST) and perfecting its classification (with STACY). We also amplify the definition of tie strength, as we view it as the likelihood of its (re) appearance in the future. We estimate such likelihood by using three SN edge features related to tie strength (edge persistence, neighborhood overlap and co-authorship count). Overall, our new algorithm called *Strength of Ties Automatic-Classifier over the Years* works over *dynamic* networks and uses social network features to classify the strength of ties in eight classes (strong, bridge+, bridge, transient, periodic, bursty, weak and random). These algorithms are compared according to the behavior of the classes that they identify. Such analysis helps to validate the algorithms and their results. Note that this article is an extended version of [Brandão et al. 2017a].

After discussing related work in Section 2, we describe the temporal SN models and the original algorithm (RECAST) in Section 3. Then, our contributions are presented as follows. Section 4 presents a definition of tie strength, the new version of the algorithm RECAST (fast-RECAST), and our new algorithm (STACY) to measure the strength of ties in large dynamic SN. Section 5 evaluates STACY classes, details experimental results, and describes our new computational model derived from STACY to directly classify tie strength. Finally, Section 6 concludes this article.

2. RELATED WORK

The strength of relationships can be calculated by considering topological and/or semantic properties of the social networks [Castilho et al. 2017; Huang et al. 2018]. The topological properties capture the structural features of the graph that constitutes the social network [Zaki and Meira Jr 2014]. For example, Brandão and Moro [2015] and Wang et al. [2017] use neighborhood overlap (or Jaccard's coefficient) to measure the strength of relationships in networks, Levin and Heuser [2010] consider the number of paths of a specific size to calculate tie strength between authors in order to do name disambiguation, and Huang et al. [2018] use the number of interactions between two users to measure the strength of friendship. On the other hand, semantic properties capture non-structural features of nodes and edges in social networks. For example, Gilbert and Karahalios [2009] define relationships on Facebook by considering the history of interactions. Finally, combining both types of properties is also possible. For example, Zignani et al. [2016] consider graph topological and temporal properties to predict the strength of relationships.

An important semantic property is the *temporal aspect*. Even with so much research on SNs, the combination of the strength of relationships and temporal aspects has not yet been widely explored. For example, Shi et al. [2018] consider the duration of calls between contacts in order to understand calling behavior and its implications on marketing decisions. In addition, Karsai et al. [2014] use tie strength to characterize the impact of heterogeneous and time-varying interactions on rumor propagation. They consider the temporal evolution of the strength of the ties, but do not propose a new way of measuring this property using the temporal aspect. On the other hand, Kostakos [2009] and

Nicosia et al. [2013] propose a set of properties of graphs that consider the temporal aspect in its calculation. Such studies show that many of these properties need to be calculated differently from static networks. Hence, we propose a new algorithm that combine existing topological properties and the temporal aspect to measure the strength of ties. The main difference regarding the existing ones is that STACY identifies eight relationship classes, providing a finer granularity for a richer analysis.

Another problem is to define and distinguish strong/weak ties in temporal networks. For instance, Karsai et al. [2014] consider both amount and time of the interactions to define the tie strength, and Huang et al. [2018] consider as weak a relationship that decreases the number of interactions in the present when compared to the past; otherwise, the relationship is strong. Then, strong ties are repeated and frequent interactions among pairs of individuals, whereas weak ties occur only occasionally. Laurent et al. [2015] define strong ties as frequent interactions that connect nodes intra-communities and model the network structure locally; whereas weak ties are infrequent interactions situated inter-communities and maintain the network structure globally connected. Differently, Nicosia et al. [2013] define two nodes i and j as strongly connected if they are in a non symmetric relation (i is temporally connected to j, but not vice-versa); whereas they are weakly connected if in a symmetric relation (both i is temporally connected to j, and j is temporally connected to i).

In this work, we consider the concept of strong and weak ties for temporal SNs based on Karsai et al. [2014]'s idea, i.e., a strong tie persists over time, and a weak tie occurs sporadically. However, Karsai et al. [2014] characterized the strength of ties based on a *single* time window of the network. Here we experimentally verify if the time window is a factor for characterizing the strength of tie by analyzing the persistence and transformation of ties over time. We show that, in fact, the strength of ties is very sensitive to the time window used to compute it.

3. RELATED AND FUNDAMENTAL CONCEPTS

In this section, we first describe the temporal social networks models and then the original RECAST.

Temporal Social Networks Models. Instead of proposing a new model for temporal SN, we borrow the ideas from Vaz de Melo et al. [2015], who have modeled it for studying mobile networks. First, we associate a start time and a duration to each relationship. Then, a temporal social network is modeled as a graph $G_k(\mathcal{V}_k, \mathcal{E}_k)$ in which time is discretized into steps of duration δ, and k is the time step in which an encounter occurs. Here, given that our evaluation is over co-authorship networks, we consider a duration of $\delta = 1$ year, as this is the common granularity for publications (not month or day). Nonetheless, our definitions are general enough to allow a finer or a coarse granularity.

Given a graph $G = (\mathcal{V}, \mathcal{E})$, $\mathcal{V} = \{v_1, ..., v_n\}$ is the set of vertices, and $\mathcal{E} = \{e_1, ..., e_m\}$ is the set of edges that represent interactions between vertices. A time-varying representation of the co-authorship SNs can be defined by a temporal accumulation graph $G_t(\mathcal{V}_t, \mathcal{E}_t)$ in G that is the aggregation of interactions in each k discrete time steps until t. Thus, all vertices interact until the t-th time step for a given value of $\mathcal{V}_t$. All edges in the set $\mathcal{E}_t$ represent interactions between vertices (v_i, v_j) during each k time step until t. Since G_t accumulates all co-authorships from the datasets and evolves over time, such aggregate graph contains social and sporadic encounters (relationships). Also, according to [Vaz de Melo et al. 2015], a random version G_t^R of the temporal aggregated graph G_t is necessary to identify the social patterns. To do that, the random graph must have similar social network topological features as the G_t graph, namely the same number of nodes, edges, and the same empirical degree distribution. The only difference, then, is how the nodes are connected among themselves.

The Original RECAST. Following the model description, we overview its original implementation algorithm, RECAST [Vaz de Melo et al. 2015], which was applied in Dynamic Complex Wireless Networks (DCWN). One contribution of our work is to modify it to measure the strength of ties in large temporal SNs. We chose RECAST because it is the only one that assigns different classes to the tie strength in temporal networks. According to Vaz de Melo et al. [2015], any system is susceptible to random events and irrational decisions called semi-rational decisions. Nevertheless, conscious decisions still govern most of the interactions. Indeed, SNs that model real interactions have edges created from semi-rational decisions (i.e., such edges tend to be regular and repeat over time), whereas random networks have edges with the same probability of connecting any two nodes. Specifically, a random network G^R is built with the same number of nodes, edges and empirical degree distribution of its social counterpart G. RECAST considers such concept of social and random networks, and implements the described temporal network model by building both G_t and G_t^R. Two algorithms are necessary to generate G_t^R from G_t: RND and T-RND. Given a graph $G(\mathcal{V}, \mathcal{E})$, RND$(G)$ returns a random graph $G_t(\mathcal{V}^R, \mathcal{E}^R)$ with the same number of nodes, number of edges and degree distribution as G. Then, the only difference between G and G^R is the connection among nodes, which is the focus of our study. Therefore, RND assigns an edge between nodes i and j with probability $p_{i,j} = (d_i \times d_j)/\sum_{k=1}^{|V|} d_k$, in which the degree distribution is $D = (d_1, d_2, ..., d_n)$ of G with n nodes. The second algorithm T-RND is an extension of RND and generates random graphs for temporal networks G_t. Thus, T-RND$(G_1 \cup G_2 \cup ... \cup G_t)$ receives a set of consecutive event graphs G_t and returns a random temporal graph G_t^R. Such algorithm builds G_t^R by running RND in each event graph G_t and then accumulating it as $G_t^R = \texttt{RND}(G_1) \cup \texttt{RND}(G_2) \cup ... \cup \texttt{RND}(G_t)$.

RECAST considers two SN features to identify social relationships: **edge persistence** maps the regularity of relationships $per_t(i,j) = \frac{1}{t}\sum_{k=1}^{t}[(i,j) \in \mathcal{E}_k]([(i,j) \in \mathcal{E}_k]$, where $per_t(i,j)$ is 1 if there is an edge (i,j) in $\mathcal{E}_k$ at time k (0 otherwise) and complementary cumulative distribution function (CCDF) $\overline{F}_{per(i,j)}(x) = P[per_t(i,j) > x]$; and **topological overlap** (a.k.a. neighborhood overlap) represents the individuals similarity $to_t(i,j) = \frac{|k|(i,k) \in \mathcal{E}_t \cap k|(j,k) \in \mathcal{E}_t|}{|k|(i,k) \in \mathcal{E}_t \cup k|(j,k) \in \mathcal{E}_t|}$ and the CCDF $\overline{F}_{to(i,j)}(x) = P[to_t(i,j) > x]$. Furthermore, RECAST has a single parameter p_{rnd} to distinguish *social* (friends, bridges and acquaintances) from *random* values of the SN features. Thus, Vaz de Melo at al. [2015] identify the feature value $\overline{x}$ that represents a threshold, such that feature values greater than $\overline{x}$ happen with a probability lower than p_{rnd} in G_t^R. Also, for small values of p_{rnd}, feature values higher than $\overline{x}$ are very improbable to occur in a random network, happening mostly due to social relationships. Also, the parameter p_{rnd} can be interpreted as the expected classification error percentage.

4. MEASURING TIE STRENGTH

We now revisit the concept of tie strength (Section 4.1) and propose fast-RECAST, an extended RECAST with multiprocessing modules to classify ties (Section 4.2). Then, we propose *STACY*, which uses three properties (instead of two) to classify tie strength (Section 4.3). By presenting such algorithms, this section also helps to answer the research question "How to properly measure the strength of ties?".

4.1 Reviewing the Concept of the Strength of Ties

Given a temporal graph $G_k(\mathcal{V}_k, \mathcal{E}_k)$, where k is the time step in which a co-authorship occurs, a tie (i,j) is likely to be **strong** if it is present in G_k for most values of k. In vice-versa, the tie (i,j) is likely to be **weak** if it is present in G_k for just a few values of k. Simply put, strong ties are likely to persist over time, and weak ties probably occur sporadically. Another characteristic of a strong tie (i,j) is that probably i and j have many neighbors in common. As previously discussed, nodes that have many neighbors in common are more likely to persist over time.

Given these two features, we group ties into the four classes defined by fast-RECAST, namely *strong* (friends), *weak* (acquaintances), *bridges* and *random*. Each class gives a level of tie strength: *strong* are ties that persist over time and share many neighbors; *weak* do not persist over time, but share many neighbors; *bridges* persist over time but share at most a few neighbors; and *random* do not persist over time and share at most a few neighbors. Hence, using these four classes, we investigate if the strength of ties are likely to transform over time. With such analysis, we are able to go deeper

Algorithm 1 Multiprocessing RECAST (fast-RECAST): a parallelized code to classify edges of G_t as random or social – strong, weak or bridge.

Require: $p_{rnd} \geq 0$
1: **return** $class(i,j)\ \forall(i,j) \in U_t E_t$
2: Construct G_t^R and set $\mathbf{RND}(G_1), ..., \mathbf{RND}(G_t)$ using **T-RND** with **pool.map_async**
3: Get $\overline{F}_{to}(x)$ and $\overline{F}_{per}(x)$ from G_t^R using **pandas dataframe**
4: Get $\overline{x}_{to}|\overline{F}_{to}(\overline{x}_{to})$ and $\overline{x}_{per}|\overline{F}_{per}(\overline{x}_{per}) = p_{rnd}$ with **pool.apply_async**
5: **for all** edges$(i,j) \in E_t$ **do**
6:     **if** $per(i,j) > \overline{x}_{per}$ and $to(i,j) > \overline{x}_{to}$ **then**
7:         $class(i,j) \longleftarrow Strong$
8:     **else if** $per(i,j) > \overline{x}_{per}$ and $to(i,j) \leq \overline{x}_{to}$ **then**
9:         $class(i,j) \longleftarrow Bridges$
10:     **else if** $per(i,j) \leq \overline{x}_{per}$ and $to(i,j) > \overline{x}_{to}$ **then**
11:         $class(i,j) \longleftarrow Weak$
12:     **else**
13:         $class(i,j) \longleftarrow Random$

into temporal social networks and answer questions such as: are *strong ties* more likely to remain *strong* in the future? Are *weak ties* more likely to become *strong ties* or to become *random*?

4.2 Multiprocessing RECAST

The construction of G_t^R using `T-RND` increases the complexity of RECAST to $O(t \times (|\mathcal{V}_t| + |\mathcal{E}_t^R|))$. Then, we propose to apply a multiprocessing Pool module from Python (a module based on communicating processes for writing concurrent programs[1]) in such step of RECAST in order to reduce its complexity. We call this novelty, multiprocessing algorithm as fast-RECAST.

The idea is that more than one random event graph G_t^R is built at a time in a multi-core computer. Thus, the new computational cost is $O(\frac{t}{p} \times (|\mathcal{V}_t| + |\mathcal{E}_t^R|))$, where p is the number of processes. After building G_t^R, the complexity of the classification is $O(|E_t^R| \times |\mathcal{V}_t|)$, in which $O(|V_t|)$ is the cost of computing the two SN features of an edge. We also add a multiprocessing Pool module from Python to call the functions to compute the edge persistence and topological overlap from the aggregated graphs. Both features are computed in parallel and asynchronously.

Algorithm 1 summarizes the code for fast-RECAST[2] with multiprocessing Pool module (lines 2 and 4) and an optimization in the memory use by applying pandas dataframe from python to store the graphs before processing them (line 3). Also, to make it more general, we rename the social edges from *friends* to *strong* ties and *acquaintances* to *weak* ties. Finally, Brandão et al. [2017] present the performance improvement of fast-RECAST regarding RECAST.

4.3 STACY - Our New Algorithm

Now, we propose a new algorithm to automatically classify tie strength called as *STACY - Strength of Ties Automatic-Classifier over the Years*. STACY improves fast-RECAST by adding a significant property: the **edge weight** (in our example, the number of publications a pair of researchers co-authored in a given year). To create random graphs G_t^R with random edge weights, we use the same algorithm to distribute the edges degree proposed by [Miller and Hagberg 2011]. However, instead of assigning edge weight as 1 for all edges, we randomly sample a (co-authorship) count from the weighted temporal graph provided as input of STACY. Thus, a weight is assigned to an edge between nodes i

[1]Multiprocessing with python: `http://docs.python.org/2/library/multiprocessing.html`
[2]Source code available in `http://homepages.dcc.ufmg.br/~mirella/projs/apoena/datasets.html`

Table I: STACY relationship classes.

Class	Edge persistence	Neighborhood overlap	Co-authorship count
Class1 - strong	social	social	social
Class2 - bridge+	social	random	social
Class3 - transient	random	social	social
Class4 - periodic	social	social	random
Class5 - bursty	random	random	social
Class6 - bridge	social	random	random
Class7 - weak	random	social	random
Class8 - random	random	random	random

Algorithm 2 STACY: a parallelized code to classify weighted edges of G_t^W as eight different tie strength classes.

Input: Weighted temporal aggregated graph - G_t^W
Require: $p_{rnd} \geq 0$
1: **return** $class(i,j)\ \forall (i,j) \in U_t E_t$
2: Construct $G_t^{R,W}$ and set **RND**$(G_1^W), ...,$ **RND**(G_t^W) using **T-RND** with **pool.map_async**
3: Get $\overline{F}_{to}(x)$ and $\overline{F}_{per}(x)$ and $\overline{F}_{coAcount}(x)$ from $G_t^{R,W}$ using **pandas dataframe**
4: Get $\overline{x}_{to}|\overline{F}_{to}(\overline{x}_{to})$ and $\overline{x}_{per}|\overline{F}_{per}(\overline{x}_{per})$ and $\overline{x}_{coAcount}|\overline{F}_{coAcount}(\overline{x}_{coAcount}) = p_{rnd}$ with **pool.apply_async**
5: **for all** edges$(i,j) \in E_t$ **do**
6: ClassifyEdges(*per*, *to*, *coAcount*) *//Performed according to Table I*

and j with probability $p_{ij} = (w_i \times w_j)/\sum_{k=1}^{|V|} w_k$ for a weight distribution $D_w = (w_1, w_2, ..., w_n)$ of G with n nodes [Chung and Lu 2002; Miller and Hagberg 2011].

Our new algorithm classifies the edges in eight different classes: seven social and one random, as described in Table I. A social network property with value equal to "social" indicates an almost zero probability of this value be produced randomly. For instance, considering the property edge persistent, if it has a value high than one that has a high probability of being present in the random graph, then the edge with such edge persistence is classified as social. Likewise, a social network property value is denominated "random" if there is a high probability of this value be produced randomly. Note that *class1* defines the strongest ties since all properties are social, whereas *class8* represents a completely random relationship. Moreover, *class2* and *class6* denote bridges, i.e., edges that persist over time, but have a small number of common neighbors; *class2* represents bridges with a high count, and *class6* with a small one. Also, *class3* denotes a relationship that is strong (high neighborhood overlap and count), but only for a particular period of time; we call it *transient*. On the other hand, *class4* represents a *periodic* relationship since it persists over time and has a high number of common neighbors, but small count. Moreover, *class5* defines a relationship with high count, but does not persist and does not share many neighbors. This relationship tends to be isolated in the social network. Finally, *class7* represents a weak tie, because it does not persist over time and has small count.

As RECAST, the unique parameter of STACY is p_{rnd} (better explained in Section 3), which determines when a social network property value is social or random. In addition, we have analyzed the results for different values of p_{rnd}, and the smallest ones provide the best results, such as $p_{rnd} = 0$ or $p_{rnd} = 0.01$. Algorithm 2 presents how ties are classified in STACY. Note that STACY is also parallelized as fast-RECAST.

5. EXPERIMENTS AND RESULTS

We now analyze the dynamism of tie strength. We first present the datasets to build the social networks based on digital libraries from distinct areas of knowledge – Computer Science, Medicine and Physics

Table II: Datasets and their basic statistics and information.

Dataset	Number of nodes	Number of edges	Period
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

(Section 5.1). Then, we apply fast-RECAST and STACY in the full temporal co-authorship SNs to characterize their tie strength (Section 5.3). We also divide the SNs in two time windows to analyze the ties' dynamism over the years using two different strategies: link persistence and link transformation. Finally, we derive a computational model from STACY (Section 5.4).

5.1 Data Description

We consider three publication datasets: DBLP, PubMed and APS, as collected in September 2015, April 2016 and March 2016, respectively. DBLP is a digital library that stores Computer Science publications. We collect publications and divide them in two datasets: DBLP Inproceedings and DBLP Articles. Pubmed is a US national library of the Medicine National Institute of Health that comprises biomedical publications. We consider publications from the top-20 journals classified by h-index. APS (American Physical Society) is an organization for diffusing and advancing the knowledge of Physics. It provides a sample dataset with its journal publications.

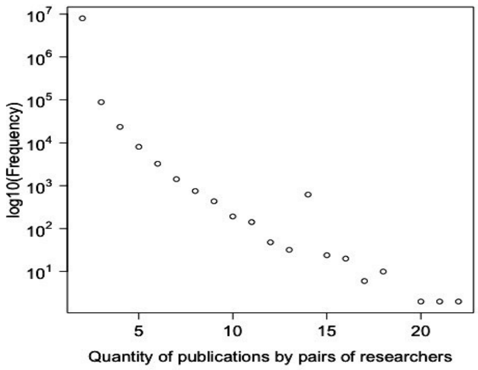
Considering these datasets, we build four co-authorship SNs whose main statistics are in Table II. Moreover, Figure 1 presents the distribution of pairs of researchers as counted yearly for each dataset. Note that the y-axis represents the frequency in log_{10}. For example, in Figure 1a, the number 5 in the x-axis and the corresponding number 10^4 in the y-axis indicate that the amount of 10^4 (in log_{10} scale) pairs of researchers have 5 publications in common considering all years in DBLP Articles dataset. The majority of co-authors have a small quantity of publications in a year, and PubMed has the largest number of co-authors in a single publication (a total of 140).

5.2 Characterizing STACY Classes

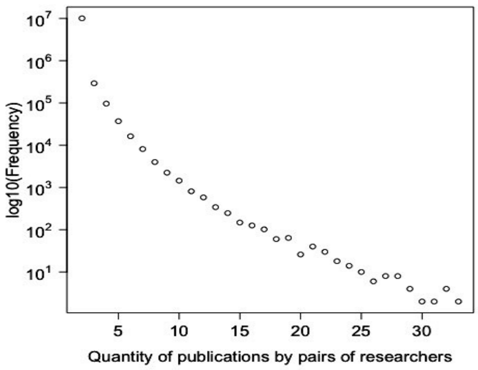
Before executing fast-RECAST and STACY, we have to set a value to the parameter p_{rnd}. Vaz de Melo et al. [2015] vary p_{rnd} through four orders of magnitude and observe that the number of edges per class keeps in the same magnitude. Therefore, such algorithm does not need an accurate definition of the parameter to consistently classify the edges. Here, we run fast-RECAST and STACY for $p_{rnd} = 0.01$ and $p_{rnd} = 0$; as we obtain similar results, we show only those for $p_{rnd} = 0$. In summary, when $p_{rnd} = 0$, a given value v of edge persistence (or topological overlap or co-authorship count) is considered *social* (or *not random*) when there are no edges in the random graphs with edge persistence (or topological overlap or co-authorship count) greater than or equal to v.

In this section, we characterize the eight classes of STACY according to the number of researchers' publications. Thus, Figure 2 presents (in box plots) the number of publications of pairs of researchers for each STACY class. In each box plot, the central rectangle spans the first to the third quartiles and shows the outliers of the distribution. Note that *strong*, *bridge+*, *transient* and *bursty* classes have pairs of researchers with more number of publications. This is trivial, because these classes have in common the value "social" to co-authorship count. However, an interesting result is that the *strong* class (value "social" for the three features) has more ties with high number of publications than the others in the four datasets. The second class that has ties with more publications is *transient*.

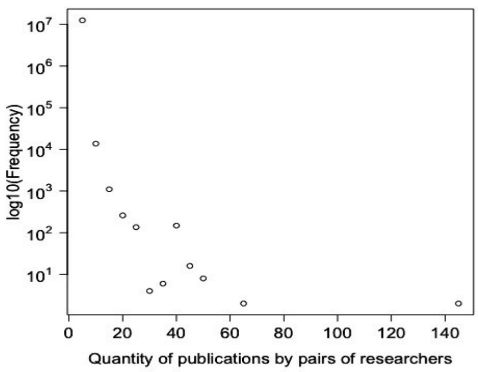
Then, Figure 3 shows the structure of APS co-authorship social networks in each STACY class. We do not show the topology for other networks due the lack of space. These visualizations allow to



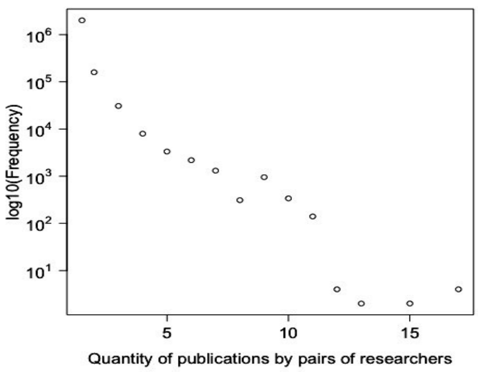
(a) DBLP Articles



(b) DBLP Inproceedings



(c) PubMed



(d) APS

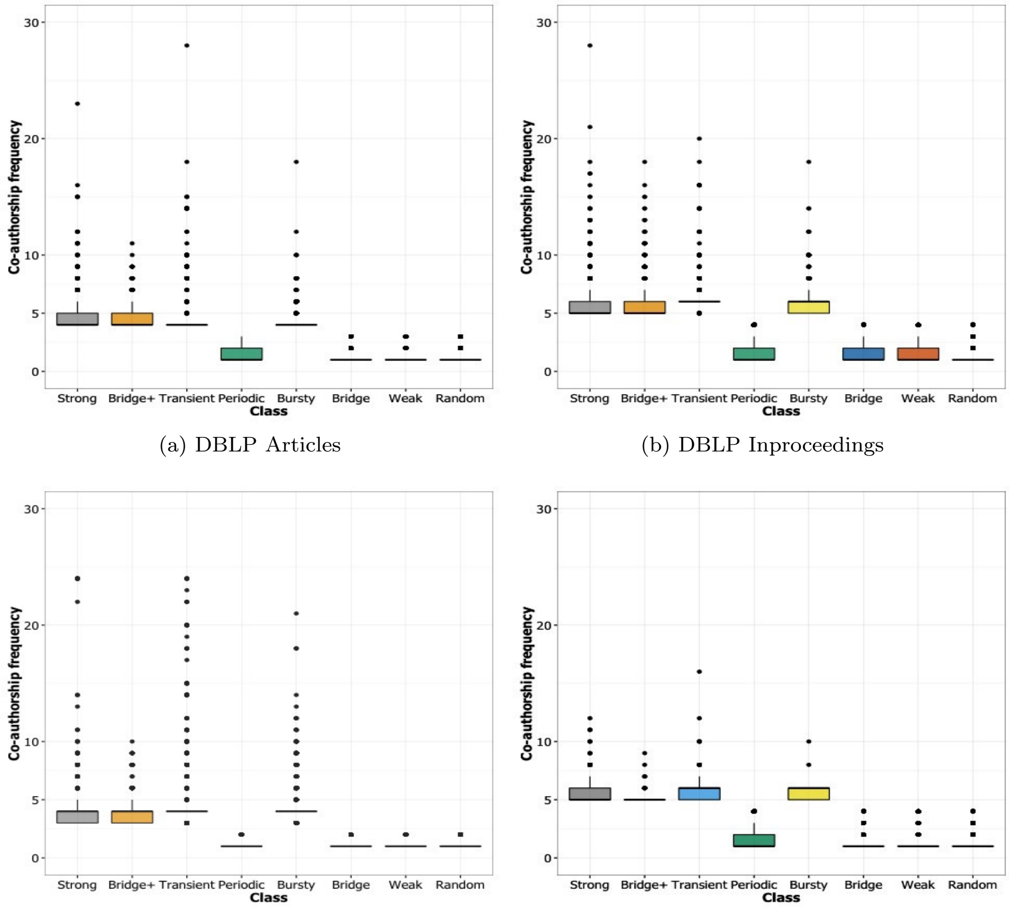
Fig. 1: Distribution of quantity of publications by pairs of researchers as counted yearly.

understand the networks' structure regarding their nodes and edges for each class. For generating the visualizations, we use *force directed layout* [Guerra-Gomez et al. 2016]. Also, Figure 4 presents the combination of the four classes with value "social" to the co-authorship count feature. Here, there are few relationships with significant amount of publications and their researchers are not well connected. Also, STACY allows to combine relationships with different properties in only one visualization.

5.3 Comparing fast-RECAST and STACY

RECAST was originally used to classify users' wireless interaction in mobile networks [Vaz de Melo et al. 2015]. The patterns and features of such networks are different from co-authorship social networks. Hence, our goal is to verify whether such algorithm identifies the kind of the relationships (social or random) between co-authors. We also do the same verification for STACY.

Classification Analyses. Here, we answer the research question "What are the characteristics of the ties according to their strength?". Figure 5 presents the classification of the co-authorships in each class generated by fast-RECAST for the four co-authorship SNs. In DBLP Articles, PubMed and APS, most co-authorships are classified as weak ties, i.e., edges with small (or *random*) topological overlap and edge persistence. Also, more co-authorships are classified as strong ties than as bridge ties. The exception is DBLP Inproceedings, in which most edges are assigned to the random class and more co-



(a) DBLP Articles

(b) DBLP Inproceedings

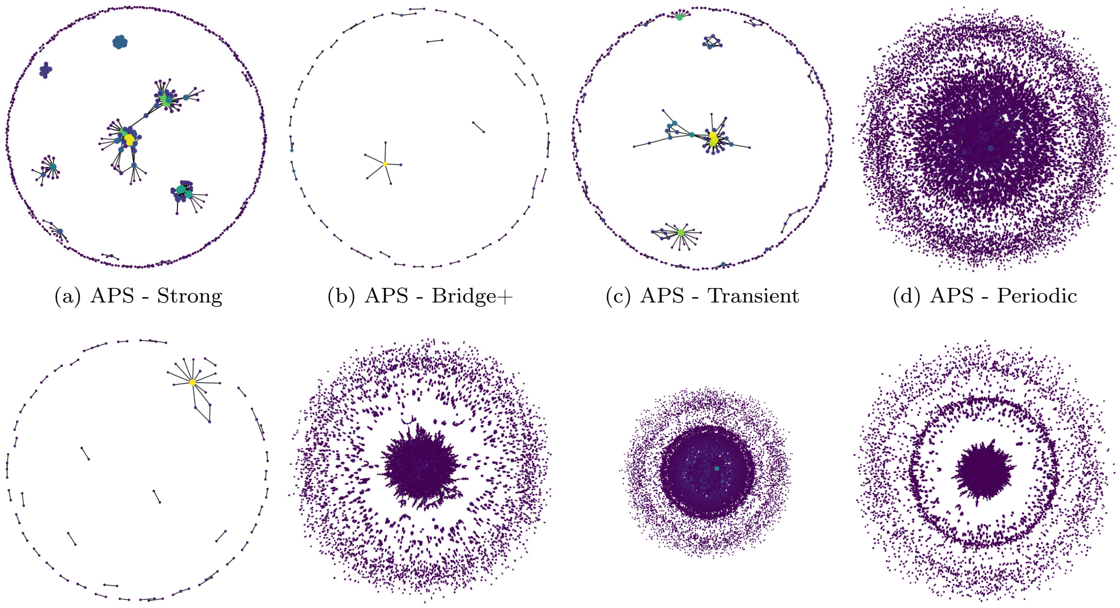
(c) PubMed

(d) APS

Fig. 2: The quantity of publications by pairs of researchers in each class detected by STACY.

authorships are classified as bridges than as strong ties. These results can be explained the datasets contain publications in journals; whereas DBLP Inproceedings has publications from conferences. As discussed in recent studies (such as [Montolio et al. 2013; Silva et al. 2014]), Computer Science has a very peculiar behavior when publishing in journals and conferences. Usually, conferences are for innovative ideas and journals for archival purposes. Hence, journal coauthor networks generally include authors who have already published together, then presenting stronger ties.

Moreover, Figure 6 shows how STACY classifies the co-authorship ties in eight different classes for each social network. As fast-RECAST, in all networks, most ties are classified as *class7* (weak) and *class8* (random). Also, many ties are classified as *class4* (periodic) and *class6* (bridge). The high quantity of ties in *class4* reveals that researchers tend to publish together with small frequency in a year with colleagues from the same community (e.g., team, department, laboratory, etc). Also, the large amount of ties in *class6* indicates that most bridges tend to have a small co-authorship count in each time. Note that less ties are classified as *class1*, *class2*, *class3* and *class5*. These four classes have in common the value "social" to the social network property co-authorship count (the other classes have a "random" value). This shows that co-authorship count is an important feature to measure tie strength since it helps to better differentiate the classes. These results are perceived in the four co-authorship social networks.



(a) APS - Strong (b) APS - Bridge+ (c) APS - Transient (d) APS - Periodic

(e) APS - Bursty (f) APS - Bridge (g) APS - Weak (h) APS - Random

Fig. 3: Social network for each relationship class from PubMed and APS dataset. The size of the nodes varies according to the number of publications of the researchers.

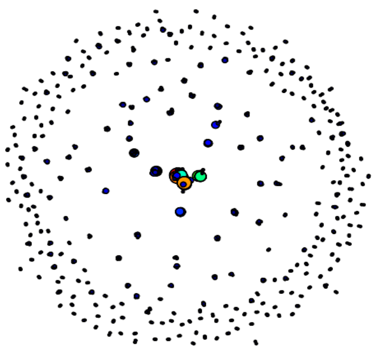
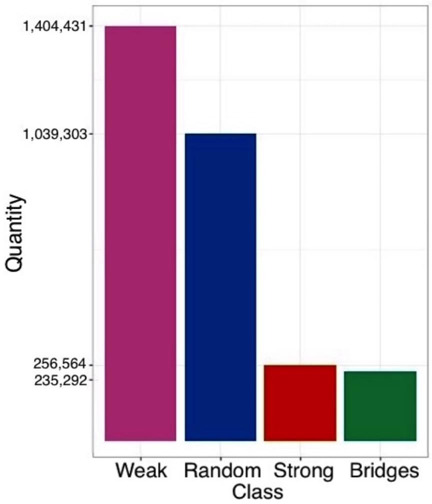


Fig. 4: APS - the merge of strong, bridge+, transient and bursty classes.

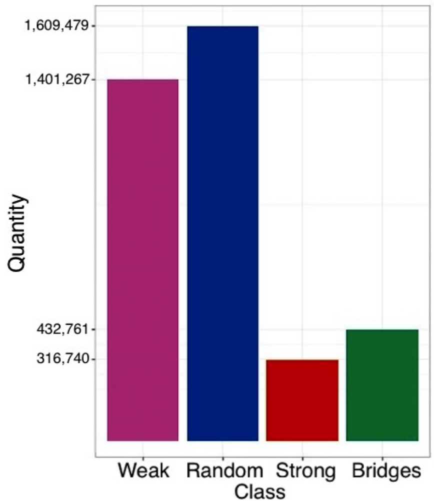
Table III: fast-RECAST: 80% represents the past (2000-2012 to DBLP, 2000-2013 to PubMed and 2000-2010 to APS) and 20% is the present (2013-2015 to DBLP, 2014-2016 to PubMed and 2011-2013 to APS).

Edge type	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)

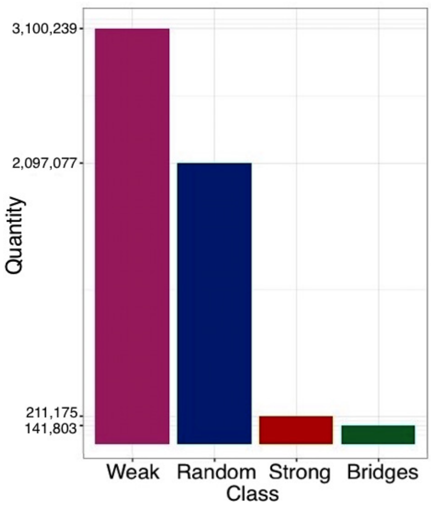
Link Persistence Analysis. Now, our goal is to investigate whether ties characterized with a given level of tie strength are likely to persist in the future. Then, to answer the research question



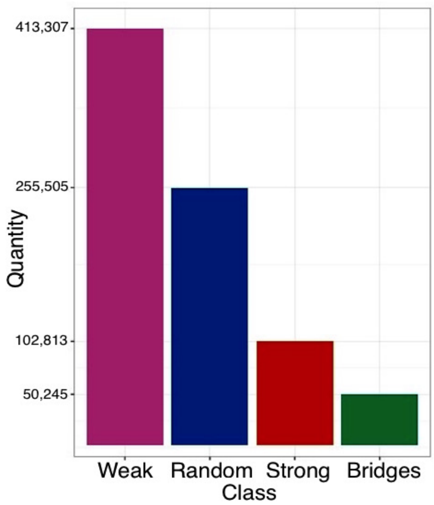
(a) DBLP Articles



(b) DBLP Inproceedings



(c) PubMed



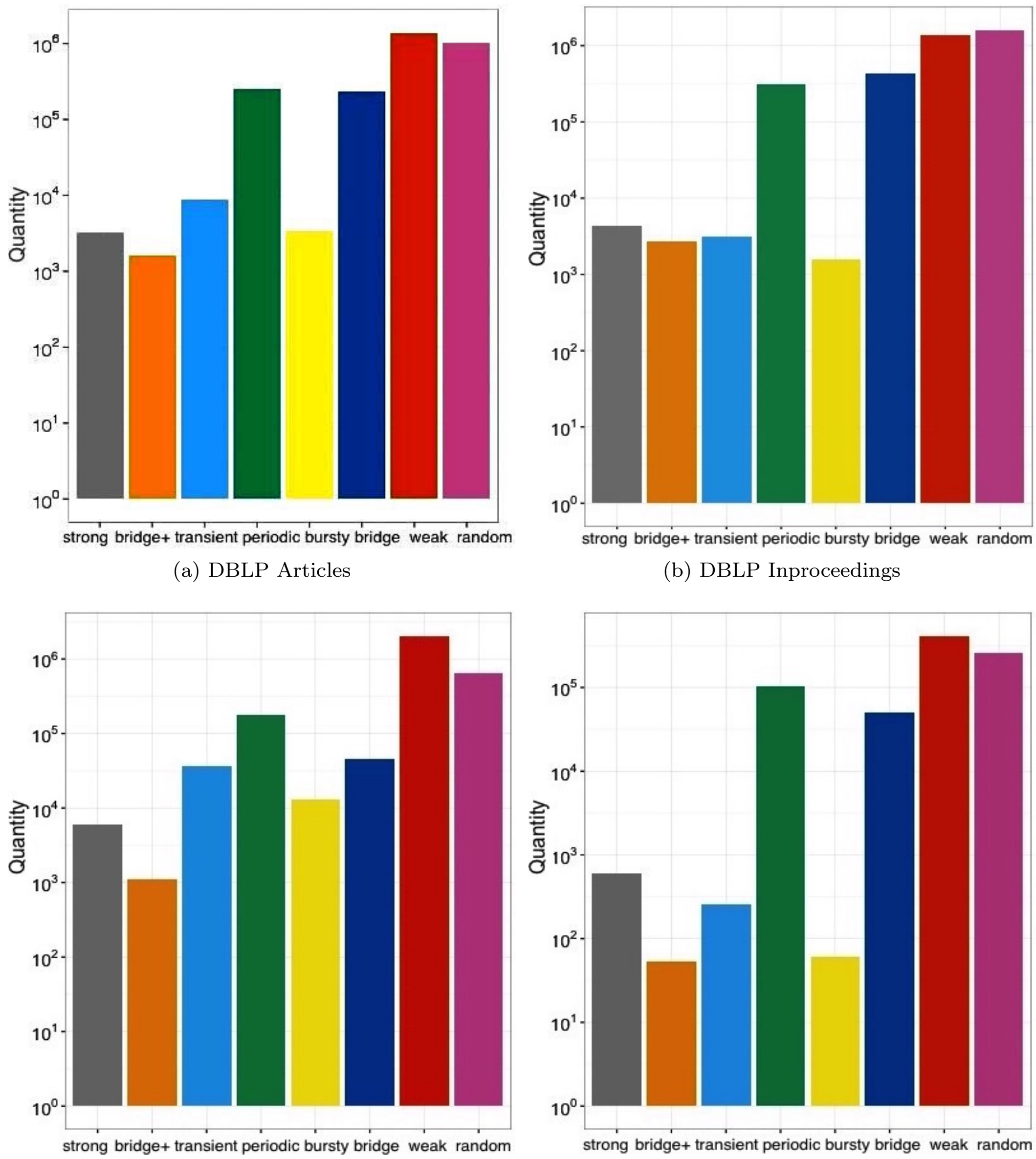
(d) APS

Fig. 5: Amount of pairs of authors in each class generated by fast-RECAST: weak, strong, bridge and random. Common behavior: the four SNs have a large number of weak and random ties.

"How are the dynamism of the ties over time?". In a social context, persistence is interpreted as the continuation of a relationship even with the progress of time, geographic distance, or occupational mobility [Adams 1967]. Here, we analyze co-authorship ties persistence over time.

To do so, we divide the networks into two time windows, which from now on we call *past* and *future*[3]. We apply *fast-RECAST* and STACY in the past and then, verify if the edges of each class (*strong*, *bridge*, *weak* and *random*) continue to be in that same class in the future. We split the social networks into two time windows and in two ways. First, we split the networks into a time window comprising 80% of the initial timestamp (*past*) and a time window comprising 20% of the final timestamp (*future*). Second, we divide the networks into time windows of 70% (*past*) and 30% (*future*). Tables III and IV present the results over the 80% and 20% partition for fast-RECAST and STACY, respectively.

[3]One may see the present as the timestamp between these two time windows



(a) DBLP Articles (b) DBLP Inproceedings

(c) PubMed (d) APS

Fig. 6: Amount of pairs of authors in each class generated by STACY: class1 to class8. Common behavior: most ties are in classes that co-authorship count has "random" value.

The values in the 80% column are the absolute number of edges from the 80% of the publications' years attributed to each class. The values in the 20% column are the number of edges from the past that are also in the future (proportions between parentheses). Answering the question "are strong ties more likely to remain strong in the future?" specified in Section 4.1, we observe that strong ties and bridges tend to persist over the years more than weak and random ties. Same conclusions apply to the 70%-30% split.

Considering fast-RECAST, we emphasize the differences in the results of the APS network in the 80%-20% and 70%-30% partitions. In the first partitioning, the proportion of strong and bridge ties

Table IV: STACY: 80% represents the past (2000-2012 to DBLP, 2000-2013 to PubMed and 2000-2010 to APS) and 20% is the present (2013-2015 to DBLP, 2014-2016 to PubMed and 2011-2013 to APS).

Edge type	DBLP Articles		DBLP Inproceedings		PubMed		APS	
	80%	**20%**	**80%**	**20%**	**80%**	**20%**	**80%**	**20%**
Class1	1,238	485 (0.39)	2,562	674 (0.26)	6,003	2,230 (0.37)	93	17 (0.18)
Class2	886	368 (0.41)	2,498	573 (0.23)	1,113	305 (0.27)	8	2 (0.25)
Class3	0	0	0	0	37,157	2,771 (0.07)	120	95 (0.79)
Class4	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)
Class5	0	0	0	0	12,862	1,372 (0.1)	4	3 (0.75)
Class6	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)
Class7	0	0	0	0	2,042,114	76,552 (0.04)	256,564	13,908 (0.05)
Class8	0	0	0	0	634,895	36,369 (0.05)	195,001	15,573 (0.08)

from the past to the present is very high, whereas in the second partitioning such proportion is lower. This result may indicate that the co-authorship social network from APS changes more through the years than the other networks. Another possibility is that physics researchers do not change very much the level of co-authorship with their collaborators over time, and this is a pattern of more recent researchers (note that 80% of data consider more recent co-authorships than 70%). We leave for future work further analyses of such possibilities.

Now, regarding STACY, we observe that strong ties tend to persist more than the others in DBLP Articles, DBLP Inproceedings and PubMed in both partitions. Also, STACY better classifies strong ties that persist over time than fast-RECAST: an increase of 0.18 for DBLP Articles, 0.12 for DBLP Inproceedings and for 0.16 PubMed in the 80%-20% partition. For the 70%-30% partition, growth is even better 0.22 for DBLP Articles, 0.16 for DBLP Inproceedings and for 0.22 PubMed. The exception is APS, in which most ties in *class3* (transient) and *class5* (bursty) tend to persist over time. This is an unexpected result since both classes have "random" value for edge persistence. Analyzing the main cause for this result, we note that co-authorships in such classes occur from 2009 to 2013, i.e., in the last years of the partitions (the 80% includes 2009 and 2010, and the 70% includes 2009). Thus, the edge persistence value is small, because the co-authorships occur in the years of the 30% (future). Additionally, no ties are classified as *class3*, *class5*, *class7* and *class8* in DBLP Articles and DBLP Inproceedings in both partitions. This reveals that in such networks transient, bursty, weak and random co-authorships are recent relationships, because they are found in the full version of these SNs (as shown by Figure 6). Also, weak and random ties are the ones that less persist over time in PubMed and APS.

Link Transformation Analysis. We now evaluate the amount of ties from a class in the past that continues in the same class (or changes) in the future. Here, we also complement the answer to the research question "How are the dynamism of the ties over time?". To avoid any kind of bias in the process of classifying the edges, here we divide the temporal co-authorship SNs into two time windows of 50% of the timestamp. We apply fast-RECAST and STACY in both parts and then we analyze the link transformation through the classes. Table V shows the results for fast-RECAST and Table VI for STACY. The values in each column represent the amount and the proportion (between parentheses) of ties from the past that persist or change class in the future. For instance, the first values 43,711 and 0.11 in Table V(a) are the number and the proportion, respectively, of *strong* links in the past that are still *strong* in the present. In the following, we answer the question "are weak ties more likely to become strong ties or to become random?" by considering the results of fast-RECAST and STACY.

Table V: Link transformation for fast-RECAST: (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)
(d)	**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)

Table VI: Link transformation for STACY: (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
Random	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)

Analyzing fast-RECAST results, surprisingly, we cannot see ties classified as *weak* and *random* for DBLP in Tables V(a) and V(b). This indicates that the features (edge persistence and topological overlap) of these SN have high (or *social*) values. Also, most past ties tend to disappear in the present, especially *bridges*. This may be explained by the nature of co-authorships, as researchers collaborate during a period towards a common goal and then, start to collaborate with others. This also reinforces the theory that weak ties are the ones that connect different communities [Granovetter 1973], which is the case of the bridge edges. For Tables V(c) and V(d), we observe similar behavior between PubMed and APS, and most ties tend to disappear, especially the bridges and random ties. Disregarding disappeared links, most strong and weak ties become weak or random. Surprisingly, the weak ties are those that remain the most in the same class, comparing to the others in both networks.

As for STACY, we also cannot see ties classified as weak (*class7*) and random (*class8*) in DBLP in Tables VI(a) and VI(b). Thus, co-authorship count in these SNs also has large (or *social*) values. Also, ties are not classified as transient (*class3*) in DBLP and APS (Table VI(d)), which reveals the

Table VII: Range of values per class in DBLP Articles.

Class	Range of values
Class1	$[0.27; 0.8]$
Class2	$[0.04; 0.12]$
Class3	$[0.15; 0.52]$
Class4	$[0.06; 0.2]$
Class5	$[0.005; 0.05]$
Class6	$[0.008; 0.03]$
Class7	$[0.04; 0.13]$
Class8	$[0.003; 0.05]$

Table VIII: Range of values per class in DBLP Inproc.

Class	Range of values
Class1	$[0.32; 0.9]$
Class2	$[0.06; 0.2]$
Class3	$[0.19; 0.6]$
Class4	$[0.06; 0.19]$
Class5	$[0.02; 0.08]$
Class6	$[0.008; 0.03]$
Class7	$[0.03; 0.13]$
Class8	$[0.003; 0.01]$

Table IX: Range of values per class in PubMed.

Class	Range of values
Class1	$[0.26; 0.67]$
Class2	$[0.08; 0.15]$
Class3	$[0.16; 0.5]$
Class4	$[0.08; 0.19]$
Class5	$[0.04; 0.07]$
Class6	$[0.02; 0.04]$
Class7	$[0.04; 0.15]$
Class8	$[0.009; 0.02]$

Table X: Range of values per class in APS.

Class	Range of values
Class1	$[0.66; 1.5]$
Class2	$[0.11; 0.26]$
Class3	$[0.29; 0.63]$
Class4	$[0.08; 0.25]$
Class5	$[0.015; 0.09]$
Class6	$[0.015; 0.04]$
Class7	$[0.04; 0.14]$
Class8	$[0.006; 0.018]$

absence of these co-authorships in earlier periods in such networks. Both DBLP networks do not have ties classified as bursty (*class5*), which indicates ties with high co-authorship count also share a large number of neighbors in this networks in the period covered by the 50% of data (this is also confirmed by the presence of ties in *class3*). Like fast-RECAST, most ties also tend to disappear when classified by STACY. The difference is that using STACY, we note that ties from different classes tend to change to *class4* (periodic) and *class6* (bridge) over time, specially, in DBLP and PubMed (Table VI).

5.4 Deriving a Computational Model from STACY – A New Direction

Here, our goal is to present a possible direction to extend our new algorithm. STACY classifies ties in eight types by combining topological overlap ($to_{(i,j)}$), edge persistence ($per_{(i,j)}$) and co-authorship count ($coAcount_{(i,j)}$). From this combination, we derive a computational model formally defined as $temporal\_tieness_{(i,j)} = per^{\alpha_1}_{(i,j)} \times to^{\alpha_2}_{(i,j)} \times coAcount^{\alpha_3}_{(i,j)}$, in which α_k (k is 1, 2 or 3) determines the weight that is given to each feature. In other words, when α_k is high for one metric and not for the other ones, it means that such metric is more import to define the strength than the others.

Considering $\alpha_1 = 1$, $\alpha_2 = 1$ and $\alpha_3 = 1$ by default, Tables VII to X present the range of values for *temporal_tieness* in each class. This new metric is calculated for each pair of researchers by using the values of the metrics computed by STACY. Note that we present these values only to show that there is a pattern in the classes through different datasets. To avoid extreme values [Brandão et al. 2014], we get the first and third quartiles of *temporal_tieness* in each class to define the beginning and end of the range of values. Note that *class1* (strong ties) has the largest values and *class8* has the smallest ones in the four co-authorship social networks. Also, *class3* has the second largest range of values in all networks. Unfortunately, there are still some overlaps between range of values between some classes, but it can be solved by better analyzing the values of the α parameter, which we leave for future work.

These results indicate that *temporal_tieness* has a pattern of values for each class in co-authorship social networks that have collaboration as an inherent characteristic. Although it is necessary to better define the range of values for some classes, *temporal_tieness* is able of directly identifying *strong*, *weak* and *random* ties. Thus, this new computational model can be used to measure tie strength in SN without using STACY, which has more computational cost. This is so, because if there is a pattern in the range of values to determine the classes, then we are able to analyze the resulting value of the combination of the metrics and identify which class better represent the strength of the relationship.

6. CONCLUSION

In this article, we characterized the strength of ties in temporal networks by measuring the persistence and the transformation of ties over time. To do so, we built four temporal co-authorship SNs considering three real datasets. We also proposed fast-RECAST, a parallel and faster version of an existing algorithm (RECAST) that classifies edges into four classes of relationship. Moreover, we proposed STACY, a parallel and fast algorithm that classifies the ties into eight different classes. We characterize each class according to the number of publications of the researchers. Regarding the results, the link persistence analysis reveals that strong ties and bridges tend to persist over the years more than weak and random ties. Overall, this supports our initial hypothesis that strong ties persist more than the others. Furthermore, STACY was able of finding strong ties that persist more than those found by fast-RECAST. The results of fast-RECAST also show a different pattern for co-authorship social network from APS when the data is divided in 80% and 20%. In this experimental setting, the proportion of strong and bridge ties from the past to the present is very high compared to other social networks. The link transformation analysis by using fast-RECAST and STACY revealed that most ties tend to disappear over time. This may occur due to the co-authorships nature, e.g., researchers tend to publish with students during a period and when the students graduate, they finalize the process of publishing together. Finally, by using STACY, we defined a new computational model called temporal_tieness and a range of values for each class. Thus, tie strength can be computed with low computational cost when compared to fast-RECAST and STACY.

As future work, we plan to apply STACY to other types of social networks, for example, in GitHub that reveals developer interactions during the software development process. We also want to add properties to STACY to differentiate recent relationships from the old while measuring tie strength. Finally, we plan to improve the experimental evaluation in order to better show the advantages of our new algorithm.

Acknowledgments. The authors would like to thank the funding agencies CAPES, CNPq and FAPEMIG, as well as the project FAPEMIG-PRONEX-MASWeb.

REFERENCES

Adams, B. N. Interaction theory and the social network. *Sociometry*, 1967.

Barabási, A.-L. *Network science.* Cambridge university press, 2016.

Bigonha, C., Cardoso, T. N. C., Moro, M. M., Gonçalves, M. A., and Almeida, V. A. F. Sentiment-based influence detection on twitter. *Journal of the Brazilian Computer Society* 18 (3): 169–183, 2012.

Brandão, M. A., de Melo, P. O. V., and Moro, M. M. Stacy: Um novo algoritmo para automaticamente classificar a força dos relacionamentos ao longo dos anos. In *Procs. of SimpÃşsio Brasileiro de Bancos de Dados (SBBD)*. Uberlândia, Brazil, pp. 136–147, 2017a.

Brandão, M. A., de Melo, P. O. V., and Moro, M. M. Tie strength persistence and transformation. In *Procs. of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web (AMW)*. Montevideo, Uruguay, pp. 1–4, 2017b.

Brandão, M. A. and Moro, M. M. Analyzing the strength of co-authorship ties with neighborhood overlap. In *Procs. of the International Conference on Database and Expert Systems Applications (DEXA)*. Valencia, Spain, pp. 527–542, 2015.

Brandão, M. A. and Moro, M. M. Social professional networks: A survey and taxonomy. *Computer Communications* vol. 100, pp. 20 – 31, 2017.

Brandão, M. A., Moro, M. M., and Almeida, J. M. Experimental evaluation of academic collaboration recommendation using factorial design. *Journal of Information and Data Management* 5 (1): 52, 2014.

Castilho, D., de Melo, P. O. V., and Benevenuto, F. The strength of the work ties. *Information Sciences* vol. 375, pp. 155–170, 2017.

Chan, H. F., Önder, A. S., and Torgler, B. The first cut is the deepest: repeated interactions of coauthorship and academic productivity in nobel laureate teams. *Scientometrics* 106 (2): 509–524, 2016.

Chung, F. and Lu, L. Connected components in random graphs with given expected degree sequences. *Annals of combinatorics* 6 (2): 125–145, 2002.

Freire, V. P. and Figueiredo, D. R. Ranking in collaboration networks using a group based metric. *Journal of the Brazilian Computer Society* 17 (41): 255–266, 2011.

Gilbert, E. and Karahalios, K. Predicting tie strength with social media. In *Procs. of Conference on Human Factors in Computing Systems (SIGCHI)*. Boston, USA, pp. 211–220, 2009.

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

Guerra-Gomez, J. A., Wilson, A., Liu, J., Davies, D., Jarvis, P., and Bier, E. Network explorer: Design, implementation, and real world deployment of a large network visualization tool. In *Procs. of International Working Conference on Advanced Visual Interfaces (AVI)*. Bari, Italy, pp. 108–111, 2016.

Huang, H., Dong, Y., Tang, J., Yang, H., Chawla, N. V., and Fu, X. Will triadic closure strengthen ties in social networks? *ACM Trans. Knowl. Discov. Data* 12 (3): 30:1–30:25, 2018.

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

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

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

Levin, F. H. and Heuser, C. A. Evaluating the use of social networks in author name disambiguation in digital libraries. *Journal of Information and Data Management* 1 (2): 183, 2010.

Lopes, G. R., Moro, M. M., Wives, L. K., and de Oliveira, J. P. M. Collaboration Recommendation on Academic Social Networks. In *ER Workshops*. pp. 190–199, 2010.

Miller, J. C. and Hagberg, A. Efficient generation of networks with given expected degrees. In *Procs. of International Workshop on Algorithms and Models for the Web-Graph (WAW)*. Barcelona, Spain, pp. 115–126, 2011.

Montolio, S. L., Dominguez-Sal, D., and Larriba-Pey, J. L. Research endogamy as an indicator of conference quality. *ACM SIGMOD Record* 42 (2): 11–16, 2013.

Nicosia et al., V. Graph metrics for temporal networks. In *Temporal Networks*, P. Holme and J. Saramäki (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 15–40, 2013.

Shi, M., Yang, B., and Chiang, J. Dyad calling behavior: Asymmetric power and tie strength dynamics. *Journal of Interactive Marketing* vol. 42, pp. 63–79, 2018.

Silva, T. H. P., Moro, M. M., Silva, A. P. C., Meira, Jr., W., and Laender, A. H. F. Community-based endogamy as an influence indicator. In *Procs. of ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*. London, United Kingdom, pp. 67–76, 2014.

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

Viana, W., da Silva, A. P. C., and Moro, M. M. Pick the Right Team and Make a Blockbuster: A Social Analysis Through Movie History. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. Pisa, Italy, pp. 1108–1114, 2016.

Wang, X., Hoi, S. C., Ester, M., Bu, J., and Chen, C. Learning personalized preference of strong and weak ties for social recommendation. In *Procs. of the 26th International Conference on World Wide Web (WWW)*. Perth, Australia, pp. 1601–1610, 2017.

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

Zignani, M., Gaito, S., and Rossi, G. P. Predicting the link strength of newborn links. In *Procs. of International Conference on World Wide Web (WWW)*. Montreal, Canada, pp. 147–148, 2016.