Tie Strength in Co-authorship Social Networks: Analyses, Metrics and a New Computational Model

Michele A. Brandão
Instituto Federal de Minas Gerais
Ribeirão das Neves, MG, Brazil
michele.brandao@ifmg.edu.br

Mirella M. Moro (advisor)
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
mirella@dcc.ufmg.br

ABSTRACT

The study of social ties has lead to build rigorous models that reveal the evolution of social networks and their dynamism. A property related to social ties is the strength of ties, which has been largely explored in different contexts, such as information diffusion, analyses of patterns in communication logs and evaluation of scientific researchers productivity. Specially, analyzing tie strength allows investigating how distinct relationships play different roles and identifying impact at micro-macro levels in the network. We present and propose different ways to measure the strength of co-authorship ties in non-temporal and temporal real academic social networks. Specially, tie strength can be measured by topological and semantic properties, as well as their combination. Finally, this thesis reveals different concepts that define tie strength and properties that influence it, along with metrics, algorithms and a classification for distinct relationships.

KEYWORDS

social network, tie strength, temporal networks

1 INTRODUCTION

This article presents a summary of the results obtained in the Thesis/dissertation defended in the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais on April 20, 2017. The work was developed by the 1st author, over a period of 49 months, under the guidance of the latter author.

Social networks (SN) are complex structures that describe individuals in any social context. Theoretically, SN can be mapped to graphs in which nodes represent individuals and edges connect nodes according to their relationships [9]. Initial studies of SNs have emphasized the importance of properly measuring the *strength of social ties* to understand social behaviors. Also, the study of social ties enables to build rigorous models that reveal the evolution of social networks and the dynamics of information exchange. More recently, analyzing tie strength has allowed to investigate the different roles of relationships including ranking for influence detection, identify impact at micro-macro levels in the network, its influence in patterns of communications and team formation. Furthermore, measuring the strength of ties allows to identify different classes [12], understand their persistence and transformation [13], and evaluate clustering algorithms quality [10].

In: I Concurso de Teses e Dissertações (CTD 2019), Rio de Janeiro, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2019.
ISSN 2596-1683

Despite the importance of analyzing tie strength, there are not many studies on evaluating how to measure it in *scientific collaboration networks* (also called *co-authorship networks*). In such networks, nodes are researchers and there is an edge between those pairs that have co-authored at least one publication. Studying the strength of co-authorship ties may reveal how its behavior relate to research, and any application based on co-authorship patterns may benefit. For instance, new strength-related metrics could help existing works on ranking researchers and their graduate programs. Furthermore, properly measuring the strength of co-authorship ties may help to identify which collaborations are more influent to each researcher. For example, if a researcher A collaborates with other researchers B and C, the strength of ties reveals which one is more important to A, then allowing different studies, such as team formation analyses. Also, researchers that form mostly weak (or strong) ties in the network may indicate different collaboration patterns. For example, a researcher who has many collaborators through single papers, i.e., that person has collaborated only once with many people.

Overall, this thesis advances tie strength theory by providing distinct analyses and a new computational model. Specifically, the main contributions of this thesis are: *(1)* a new general taxonomy to social networks that classifies related work according to their main goal; *(2)* an analysis of how nine topological properties affect the strength of co-authorship ties when measured by neighborhood overlap; *(3)* a nominal scale to neighborhood overlap for classifying a tie as weak or strong; *(4)* a new metric to measure the strength of ties in non-temporal SNs called *tieness*, along with its nominal scale; *(5)* a new algorithm called STACY (Strength of Ties Automatic-Classifier over the Years) that automatically classifies the strength of ties in temporal networks, along with a computational model named *temporal_tieness*; *(6)* a set of eight tie strength classes identified by STACY; *(7)* an analysis of how tie strength is defined over time; and *(8)* real applications of tie strength metrics and algorithms.

The results of this thesis appear on eleven publications: [9], [12] best paper honorable mention, [4], [6], [11], [10], [13], [7] best paper honorable mention, [8], [5] and [14]. This thesis has also contributed with a vast set of datasets[1] and a new visualization tool, called CNARe[2]. Problems defined in the scopus of this thesis helped to shape other works as follows: six undergrad research scholarships (Gabriel Oliveira, Gabriela Brant, Guilherme de Sousa, Matheus Diniz, Nivaldo Teixeira e Thiago Prado), five undergrad final projects (Gabriela Brant, Gabriela Duarte, Guilherme de Sousa, Mariana de Oliveira e Natércia Batista), and three master thesis (Jeancarlo Leão, Mariana de Oliveira e Natércia Batista). Also, an

[1] Datasets: http://www.dcc.ufmg.br/~mirella/projs/apoena
[2] CNARe: http://www.dcc.ufmg.br/~mirella/Tools/CNARe

extension of this thesis was published in WebMedia 2018 and win honorable mention [18]. Finally, this thesis has technical and scientific impact by providing new useful concepts, metrics, algorithms and model. Also, it has economic and political impact[3] by offering new ways to model user behavior, and social impact by helping users to obtain relevant knowledge.

2 SOCIAL PROFESSIONAL NETWORK: A SURVEY AND TAXONOMY

The Web has introduced different and new ways in which professionals can easily share their work, publish content, find job opportunities, interact with other professionals, and so on. Besides general purpose social networks, such as Facebook and Twitter, there are online social professional networks (SPN) whose focus is on those activities. Indeed, there are currently more than 20 websites for social professional purposes. Furthermore, online social networks as a type of communication networks enable straightforward information access. Finally, with such a big volume of data available, researchers have used the data from those sites to study SPN characteristics and discover behavioral patterns.

Nonetheless, there are many challenges on SNs [15]: collecting the data, inferring social process from the data, keeping individual privacy, choosing the best technique to select the data, among others. The social professional networks have an additional challenge that is modeling user emotion. For instance, it is hard to differ if a professional behavior is based on emotional reasons or not. Therefore, with an SPN survey, we help to identify possible existing solutions for these challenges by categorizing existing work according to the social professional network type, goal and stage of development. Note that co-authorship social network is a type of social professional network. Thus, this study helps to situate tie strength research in the state of the art.

Overall, we define a general taxonomy considering issues and tasks as a first-level classification [9]. Issues are problems *within* social networks regarding their maintenance and usage, whereas tasks are problems whose solutions benefit from *using* SN data. The new taxonomy is based on four classes, being two for issues and two for tasks respectively: We propose a taxonomy based on the tasks and issues of social networks. By analyzing the publications on the area, we have identified two main tasks (analysis and application) and two main issues (data acquisition and preparation, and data storage), as illustrated in Figure 1 and detailed next. *(1)* data acquisition and preparation: the focus is obtaining the data from social networks or other sources; *(2)* data storage: SNs require storing and accessing their data; *(3)* analysis: the goal is to examine nodes and their interactions in SN towards a specific goal; and *(4)* application: the goal is to use social network to develop methods, features and programs to benefit users. Therefore, in the survey of this work, we detail studies on recommendation and ranking strategies in social professional network context. Both applications have motivated competitions such as Netflix Prize, CAMRA, the Yahoo! Music KDD Cup 2011 and Kaggle's competitions.
Solutions applied to the two issues (data acquisition/preparation and data storage) may be adapted to the SN context as they are *not* exclusive for social professional networks. The same cannot be said about the two tasks (analysis and application), which we study specifically in the context of social professional networks.

[3]Stanford business: https://www.gsb.stanford.edu/insights/studying-social-networks-developing-worlds-five-key-insights

3 TIE STRENGTH OVER NON-TEMPORAL CO-AUTHORSHIP SOCIAL NETWORKS

In non-temporal SNs, neighborhood overlap and absolute frequency of interaction (a.k.a. edge weight or co-authorship frequency – $coAfrequency$) have been largely used to measure the strength of ties. Indeed, we have initially measured the strength of ties by using neighborhood overlap (NO) and contrasting it with co-authorship frequency (note: the neighborhood overlap metric of an edge connecting researchers v_i and v_j is given by an equation in the thesis Section 4.2.1). Furthermore, the absolute frequency of interaction is the total number of publications that a pair of researchers v_i and v_j have together. The combination of these two metrics allowed to define a nominal scale to NO.

Afterwards, we analyze how nine topological properties impact on the strength of ties [8]. Initially, we study the correlation between each property and neighborhood overlap. Then, our analysis takes one step forward and considers a regression model to quantify how the combination of each property influences neighborhood overlap. The results showed that different properties influence such metric in a linear or non-linear way and revealed which properties should be combined with NO. In addition, since we are measuring the strength of ties, we verified if Granovetter's theory governs co-authorship SN when such strength is measured by neighborhood overlap. Our results were positive to such theory. Therefore, our evaluations indicate that neighborhood overlap can be used to measure the strength of ties [5].

However, by empirically analyzing the results, we identified four main problems with using solely neighborhood overlap and co-authorship frequency to measure tie strength [7, 11]: *(Case 1)* when a pair of collaborators does not have any common neighbor, neighborhood overlap is zero; *(Case 2)* when determining if two collaborators are from the same community (or not), co-authorship frequency fails, as it considers only the absolute frequency of interaction; *(Case 3)* when there is little collaboration between a pair of collaborators and a plenty of common neighbors, neighborhood overlap and co-authorship frequency present opposite results; and *(Case 4)* when the results are extreme values, neighborhood overlap may not represent the reality. Hence, we proposed a new metric entitled *tieness* that combines a modified neighborhood overlap with co-authorship frequency [11] (also defined by Equation 5.1 in the thesis Section 5.3).

The main advantage of tieness is to allow in a single way and low computational cost, the combination of a topological property from SNs with a semantic one. The topological property is always a modification in neighborhood overlap, but the semantic property can be any one that represents the strength of the relationship. For instance, the number of shared projects in a software development social network.

Figure 1: Main social networks topics: the tasks refer to using social networks to solve problems, and the issues address problems related to managing social networks.

4 TIE STRENGTH OVER TEMPORAL CO-AUTHORSHIP SOCIAL NETWORKS

Changing the context to *temporal* SNs, this thesis proposes two algorithms to measure tie strength: fast-RECAST [6] and STACY [12]. Both algorithms classify the ties by comparing the values of SNs features with values from random networks. In summary, fast-RECAST identifies four relationships classes (strong, weak, bridge and random), whereas STACY classifies the ties in eight different classes (strong, bridge+, bridge, transient, periodic, bursty, weak and random). As STACY recognizes more tie strength classes, it allows to identify more types of relationships. Thus, such a new algorithm is able to automatically finding distinct kinds of co-authorships. For example, Figure 2 shows the SN built with the edges of each class. This reveals how the classes allow the network to be filtered in order to perform, for instance, link prediction, recommendations and community detection. Furthermore, from STACY, we are able to derive a computational model called *temporal_tieness* that can classify tie strength with low computational cost (as detailed by Equation 6.3 in the thesis Section 6.3.4).

There are different concepts related to tie strength in non-temporal social networks. However, few studies have addressed the strength of ties in temporal networks. In this work, we considered that a strong tie characterizes interactions that are likely to appear in the future, whereas a weak tie occurs sporadically. Our results confirm such claim, since strong ties persist more than weak ones.

We also investigate tie strength dynamism over time through analyzing tie persistence and transformation in different classes by applying fast-RECAST and STACY [6, 13]. Surprisingly, most ties tend to perish 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. Moreover, the link persistence analysis reveals that strong ties and bridges tend to persist over the years more than weak and random ties. Also, STACY reveals that more persistent bridges have "social" value to co-authorship frequency. Likewise, it is able to finding strong ties that persist more than those found by fast-RECAST. All these results show that STACY is able of automatically finding different kinds of relationships in temporal co-authorship social networks.

5 TIE STRENGTH APPLICATION

In this research, tie strength metrics/algorithms are also applied in two contexts.

Clustering algorithm analyses and evaluation. We analyze the strength of ties intra and inter communities by using neighborhood overlap and co-authorship frequency in different clustering algorithms [10, 20]. Then, we investigate how the removal of random edges can improve the results of clustering algorithms [16].

Social network visualization. After classifying the co-authorships by using neighborhood overlap and fast-RECAST, we plot the visualization of the social networks with the relationship classes [4, 14]. Thus, this thesis also contributed to the generation of a new tool called CNARe.

6 CONCLUDING REMARKS

This work addressed distinct aspects related to the strength of co-authorship ties. Specially, we did analyses, formulated metrics and developed a new computational model. Also, we presented a survey and taxonomy to SPNs, an analysis of topological properties related

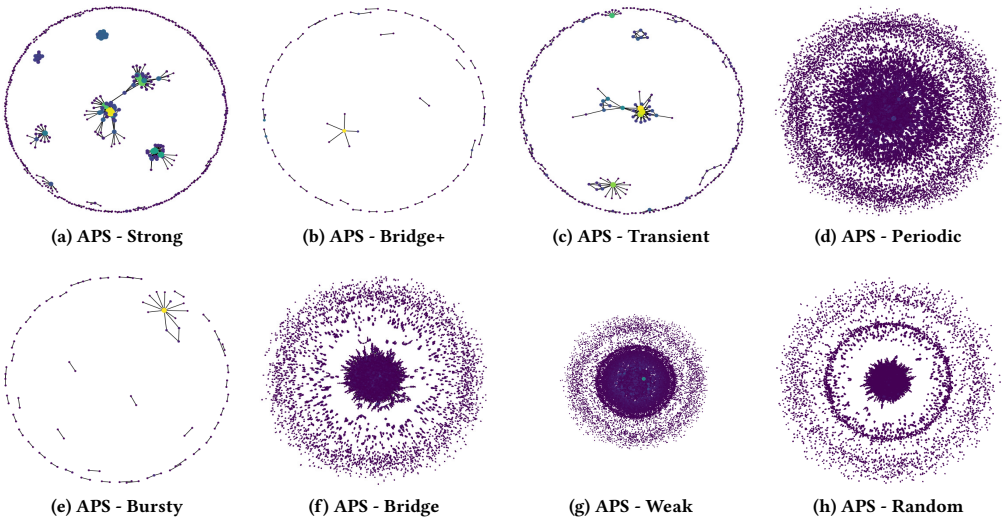

Figure 2: SN for each relationship class from APS (American Physical Society) dataset. The size of the nodes varies according to the number of publications of the researchers.

to tie strength, tieness, STACY and temporal_tieness. Due to the lack of space, we have only summarized the contributions and experiments conducted in our work. In terms of measuring co-authorship tie strength, we have used different real datasets, analyzed the impact of different properties in metrics commonly applied to measure tie strength and compared the results from existing algorithms and the proposed ones. Also, we have shown that the results obtained by tieness and STACY bring new information about the strength of the relationships and in an automatic and generic way. Details and discussions may be found in the publications described in Section 1 and the thesis itself.

Finally, directions for future work include: expanding the study to other collaboration SNs (with results already published for software development networks); using qualitative research to evaluate tie strength; evaluating tie strength methods by comparing with synthetic data; clustering analyses and evaluation; exploring parameters in temporal_tieness; and adding other SN features to STACY. We are currently exploring such ideas, and have recently published results and ongoing work in: [1–3, 16, 17, 19, 20].

ACKNOWLEDGMENTS

Thesis funded by a scholarship from CAPES, Brazil.

REFERENCES

[1] Natércia A Batista, Gabriela B Alves, André L Gonzaga, and Michele A Brandão. 2017. GitSED: Um Conjunto de Dados com Informaç oes Sociais baseado no GitHub. In *SBBD*. Uberlândia, Brazil, 224–233.
[2] Natércia A Batista, Michele A Brandão, Ana Paula C da Silva, and Mirella M Moro. 2017. Aspectos Temporais para Medir a Força da Colaboraç ao no GitHub. In *SBBD*. Uberlândia, Brazil, 234–239.
[3] Natércia A Batista, Michele A Brandão, Gabriela B Alves, Ana Paula Couto da Silva, and Mirella M Moro. 2017. Collaboration strength metrics and analyses on GitHub. In *Procs. of the International Conference on Web Intelligence (WI)*. Leipzig, Germany, 170–178.
[4] Michele A. Brandão, Matheus A. Diniz, Guilherme A. de Sousa, and Mirella M. Moro. 2017. Visualizing Co-Authorship Social Networks and Collaboration Recommendations With CNARe. In *Graph Theoretic Approaches for Analyzing Large-Scale Social Networks*, Natarajan Meghanathan (Ed.). IGI Global, 173–188.
[5] M. A. Brandão and M. M. Moro. 2015. Neighborhood Overlap: Can This Metric Be Used to Characterize the Strength of Co-authorship Ties?. In *ACM Student Research Competition & Grace Hopper Celebration*.
[6] Michele A. Brandão, Pedro O. S. Vaz de Melo, and Mirella M. Moro. 2017. Tie Strength Dynamics over Temporal Co-authorship Social Networks. In *Procs. of the International Conference on Web Intelligence (WI)*. Leipzig, Germany, 306–313.
[7] Michele A. Brandão, Matheus A. Diniz, and Mirella M. Moro. 2016. Using Topological Properties to Measure the Strength of Co-authorship Ties. In *BraSNAM*. Rio de Janeiro, 199–210.
[8] M. A. Brandão and M. M. Moro. 2015. Analyzing the Strength of Co-authorship Ties with Neighborhood Overlap. In *DEXA*. Valencia, Spain, 527–542.
[9] Michele A. Brandão and Mirella M. Moro. 2017. Social professional networks: A survey and taxonomy. *Computer Communications* 100 (2017), 20 – 31.
[10] Michele A Brandão and Mirella M Moro. 2017. Strength of Co-authorship Ties in Clusters: a Comparative Analysis. In *AMW*. Montevideo, Uruguay.
[11] Michele A Brandão and Mirella M Moro. 2017. The strength of co-authorship ties through different topological properties. *Journal of the Brazilian Computer Society* 23, 1 (2017), 5.
[12] M. A. Brandão, Pedro O. S. Vaz de Melo, and M. M. Moro. 2017. STACY: Um Novo Algoritmo para Automaticamente Classificar a Força dos Relacionamentos ao Longo dos Anos. In *SBBD*. Uberlândia, 136–147.
[13] M. A. Brandão, P. O. S. Vaz de Melo, and M. M. Moro. 2017. Tie Strength Persistence and Transformation. In *AMW*. Montevideo, Uruguay.
[14] Guilherme A. de Sousa, Matheus A. Diniz, Michele A. Brandao, and Mirella M. Moro. 2015. CNARe: Co-authorship social network analysis and recommendations. In *RecSys*. Vienna, Austria, 329–330.
[15] J. M. Kleinberg. 2007. Challenges in mining social network data: Processes, privacy, and paradoxes. In *SIGKDD*. San Jose, USA, 4–5.
[16] Jeancarlo C. Leão, Michele A. Brandão, Pedro O. Vaz de Melo, and Alberto H. F. Laender. 2017. Classificação de Relações Sociais para Melhorar a Detecção de Comunidades. In *BraSNAM*. São Paulo, Brazil, 647–657.
[17] Jeancarlo C. Leão, Michele A. Brandão, Pedro O. Vaz de Melo, and Alberto H. F. Laender. 2017. Mineração de Perfis Sociais em Redes Temporais. In *SBBD*. Uberlândia, Brazil, 264–269.
[18] Gabriel P. Oliveira, Natércia A. Batista, Michele A. Brandão, and Mirella M. Moro. 2018. Tie Strength in GitHub Heterogeneous Networks. In *Procs. of Brazilian Symposium on Multimedia and the Web (WebMedia '18)*. Salvador, BA, Brazil, 363–370.
[19] Talita S. Orfanó, Michele A. Brandão, Larissa E. Maia, and Mirella M. Moro. 2017. Análise da Formação e Evolução de Times de Desenvolvimento no Hibernate-ORM. In *SBBD*. Uberlândia, Brazil, 216–221.
[20] Mariana O Silva, Michele A Brandão, and Mirella M Moro. 2017. A Força dos Relacionamentos pode Medir a Qualidade de Comunidades?. In *SBBD*. Uberlândia, Brazil, 204–209.