Multimodal graph-based analysis over the DBLP repository: critical discoveries and hypotheses

Gabriel P Gimenes, Hugo Gualdron,
Jose F Rodrigues Jr
University of Sao Paulo
Av Trab Sao-carlense, 400
Sao Carlos, SP, Brazil - 13566-590
{ggimenes,gualdron,junio}@icmc.usp.br

Mario Gazziro
Fed. University of Santo Andre
Av dos Estados, 5001
Santo Andre-SP-Brazil - 09210-580
mario.gazziro@ufabc.edu.br

ABSTRACT

The use of graph theory for analyzing network-like data has gained central importance with the rise of the Web 2.0. However, many graph-based techniques are not well-disseminated and neither explored at their full potential, what might depend on a complimentary approach achieved with the combination of multiple techniques. This paper describes the systematic use of graph-based techniques of different types (multimodal) combining the resultant analytical insights around a common domain, the Digital Bibliography & Library Project (DBLP). To do so, we introduce an analytical ensemble based on statistical (degree, and weakly-connected components distribution), topological (average clustering coefficient, and effective diameter evolution), algorithmic (link prediction/machine learning), and algebraic techniques to inspect non-evident features of DBLP at the same time that we interpret the heterogeneous discoveries found along the work. As a result, we have put together a set of techniques demonstrating over DBLP what we call multimodal analysis, an innovative process of information understanding that demands a wide technical knowledge and a deep understanding of the data domain. We expect that our methodology and our findings will foster other multimodal analyses and also that they will bring light over the Computer Science research.

Categories and Subject Descriptors

G.2.2 [**Discrete Mathematics**]: [Digital Libraries] Graph Theory

General Terms

Network analysis, graph analysis, DBLP

SAC 2015 April 13 - 17 2015, Salamanca, Spain
ACM Copyright 2015 ACM 978-1-4503-3196-8/15/04 $15.00.
http://dx.doi.org/10.1145/2695664.2695699

1. INTRODUCTION

The properties and evolution of real-world networks are relevant topics nowadays, when network and mobile technologies are matured and disseminated. Network-like data arise, along the time dimension, from multiple domains in the order of hundreds of thousands of entities (nodes) and millions of relationships (edges). In respect to this matter, we study such networks by characterizing a particular type of data, that of the Computer Science literature - a network of co-authoring, co-edition of publications, and co-publication in periodicals. Scientific collaboration answers for a broad scope of interest; not only authors and editors, but the funding agencies and the society demand knowledge about how scientists behave concerning their collective production. As so, the analysis of Digital Bibliography & Library Project (DBLP) [1], one of the world's major Computer Science literature repositories, from a network perspective can bring insights about the academic field and its future development.

Discovering non-evident facts about DBLP is not a trivial task, therefore we introduce what we call multimodal analysis, an ensemble of analytical techniques each with a different characteristic. We rely on statistical (degree, and weakly-connected components distribution), topological (average clustering coefficient, and effective diameter evolution), algorithmic (link prediction/machine learning), and algebraic techniques to inspect non-evident features of DBLP. First, we calculate statistical distributions over one snapshot of DBLP, what leads to a panorama of the main characteristics of its underlying net. Second, we draw further time-related topological measurements to present how DBLP evolves along the time. Third, algorithmic calculations translate metrics into meaningful probabilistic raisings. Lastly, simple counting and algebraic analysis reveal intuitive, although not evident, aspects. We considered relationships co-authoring, co-edition, and co-publication in order to produce conclusive observations of how the Computer Science community has behaved along the years.

Our contributions refer to the use of a wide set of measurements in light of three different relationships, in static and in dynamic fashion, generating conclusions over a dataset of public interest. In order to explore the content that lies within DBLP, we employ a broad range of graph-based measurements, that is, Social Network Analysis (SNA). From our computations, we combine our findings to achieve a deep

[1] http://www.informatik.uni-trier.de/ ley/db/

perspective of the practices of the Computer Science community. Moreover, we present our methodology as a generic model that for the analysis of similar networks - our work comes as a systematic analytical process to be reproduced by academic peers and by practitioners.

2. RELATED WORK

There are plenty of studies that use Social Network Analysis (SNA) to transform network data into knowledge. Osiek *et al.*[17] try to answer whether attending conferences tend to increase scientific collaboration. To do so, the authors assume that having papers in the same conference correspond to a chance of conference-induced collaboration. With simple counting, they drew their conclusions by tracking the first common conference of each pair of authors and the first paper they wrote together. Conclusively, only 4.61% of the pairs of authors satisfied their supposition. Z. Huang *el al.*[9] used the Clique Percolation Method [15] to monitor the presence and the size of semantic communities over DBLP; they identified giant and small communities, each one with peculiarities about content, size and evolution. According to the notion of centrality, Leydesdorff [14] uses measures degree, betweenness, and closeness to evaluate the interdisciplinarity that is found in journals - although inconclusive, the author brings light to the problem.

Bollen *et al.*[3] collected online requests for electronic publications (clickstream) from Thomson Scientific, Elsevier, JSTOR, Ingenta, University of Texas, and California State University. Then, with metrics PageRank and betweenness they built a science map based on data from a broader audience updated in real time. J. Huang *et al.* [8] investigates a fragment of the Computer Science CiteSeer Digital Library [2]; the authors performed a three-level analysis: network level, community level, and individual level. Their conclusions compare Database and Artificial Intelligence communities and introduce a Stochastic Poisson model to predict future collaboration behavior.

These previous works aim at characterizing the properties of nodes alone or, at last, the global properties of the structure by means of single metrics. In this work, we analyze the DBLP data by drawing the statistical distribution of several of its properties, and by drawing metrics that consider the time dimension.

In a different line, Leskovec *et al.* [11] present an extensive work on collecting metrics from a time evolving graph. In this work, the authors discuss the dynamics of viral marketing based on a large set of metrics, and on a recommendation propagation model. More recently, Benevenuto *et al.* [2] collected and analyzed time-evolving clickstream data from a large social network and, through statistical measures, deduced many aspects of its behavior. Finally, Huffaker *et al.*[10] describe an interesting analysis of interaction patterns on a virtual world environment; they do so by means of multiple measures such as shortest path, group similarity, clustering coefficient, and largest connected component. In conclusion, the authors describe the role of collective structures in determining the conduct of its members.

Recently, Aiello *et al.* [1] described how friends that have similar profiles (homophily) tend to get interconnected. In their study, the authors consider the groups to which the users belong, and the annotations (tags) of the users, among other features. With these features, the authors calculate the similarity between users, proposing a similarity threshold to state whether two users are to define a connection, or not. Regardless of its significative results, this study extrapolates the topological information of the network; it relies on information that, often, is not available or is not well-defined. This same limitation is faced by Brandao *et al.* [4] and Lim *et al.* [16].

By considering static and dynamic analytical approaches, as those presented in this section, we propose a multi-faceted analysis of DBLP – section 3.1. In our work, we draw conclusions from different points of view, static and dynamic, and from diverse complementary metrics. As so, we introduce necessary concepts in section 3, describe their application in section 4, and draw our conclusions in section 5.

3. MATERIAL AND METHODS

3.1 Digital Bibliography & Library Project (DBLP)

We used DBLP, one of the largest Computer Science bibliographic repositories available and, now, part of the ACM SIGMOD Anthology project [3]. DBLP includes journals and conference proceedings since year 1936 - the main fields of its data entries are title, publication, authors, year of publication, page numbers, and editors – among others. In our investigation, we used records dated between 1970 and 2011 with the goal of characterizing the DBLP research community globally and along its evolution. Table 1 lists the cardinality of each entity extracted from DBLP used in our analysis.

Table 1: Number of entities involved in our analysis.

Entity	Number
Authors	*1.060.221*
Articles	*1.801.576*
Events	*14.654*
Publications	*4.262*

DBLP is available as an XML file that demands specific software for parsing its semi-structured data. It is a reference collection whose data quality is an important worry of its custodians; nevertheless, it presents some minor problems like name ambiguity and lack of data standardization. These problems prevented us from thoroughly using its content; not, however, causing prejudice to our analysis, as less than 5% of the data could not be used.

In order to use DBLP benefiting from the facilities of a Database Management Systems, we parsed it migrating its data to a rigid relational structure, from what we derived relationships. Figure 1 shows the data modeling, which has four many-to-many relationships: author-authorship-article, author-publishes-in-publication, author-attends-conference,

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

[3] http://www.sigmod.org/publications/sigmod-record/anthology

and author-edits-conference; another point is that conference is a weak entity of publication, what means that every conference must have a correspondent publication (journal, proceedings, or book).

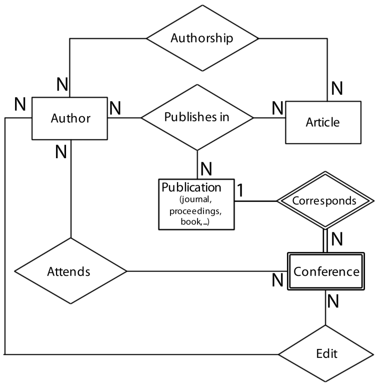


Figure 1: Entity-relationship modeling of our data.

Defining co-relationships from DBPL

From the original data of DBLP, we are interested in analyzing author co-relationships as, for example, given many-to-many relationship "Author (A) publishes (R) Article (B)", the object of our study is the intrinsic co-authoring relationship. As a formal example, the "publishes" relationship can be represented in relational notation as $A = \{pk_a, name\}, B = \{pk_b, title\}$ and $R = \{pk_a, pk_b\}$. In this case, we are interested in the authors that published papers together; that is, we need a relation given, in relational algebra, by $\textit{co-authorship} \leftarrow \Pi(A.pk_a, R'.pk_a)((A \bowtie_{A.pk_a=R.pk_a} R) \bowtie_{A.pk_a \neq R'.pk_a \wedge R.pk_b \neq R'.pk_b} \rho_R(R'))$. The result is relation $\textit{co-authorship} = \{author, author', count\}$ used in our analysis. Furthermore, we created relations *co-participation* for authors who had papers at the same conferences, *co-publication* for those who had papers in the same journal, and *co-edition* for those who appear as editors of the same event or journal. Table 2 summarizes the datasets:

Table 2: Relations extracted from DBLP and used in our analyses.

Relation	Description
Co-authorship	*Authors who published papers together.*
Co-participation	*Authors who had papers in the same conference.*
Co-publication	*Authors who had papers in the same journal.*
Co-edition	*Authors who appeared as editors of the same event or journal.*

We created these relations in two versions: one having the number of times the relationship occurred, the other including the year when the relationship first took place. Respectively, they correspond to static weighted graphs and to graphs that evolve along time, both undirected for our experiments.

3.2 Methods

We make use of a number of social network metrics and techniques to inspect the characteristics of DBLP in complementary fashion. We apply these techniques either for the entire static network, or for consecutive annual snapshots of it. We present the results as summarizing plots and tables along the spectrum of each variable. We make use of the following methods:

- weakly-connected components (WCC) distribution: a WCC is an undirected subgraph in which every node has a path to every other node; counting the sizes of the connected components indicates how integrated are the research sub communities of the network;
- average clustering coefficient (ACC): global tendency of nodes to form clusters, or communities within the network – in a network of authors, it refers to the property of transitivity (presence of triangles), or, how likely the co-authors of my co-authors will become my co-authors; the coefficient ranges from 0 to 1, higher values indicating higher tendency of clustering;
- degree distribution (densification): the counting of the number of nodes with each given degree answers for how intense authors interact one with each other revealing important aspects of the evolution of the network;
- effective diameter evolution: the length of the 90th percentile path between any pair of nodes considered over time – states whether it happens and how intense is the small world phenomenon;
- predictability: refers to algorithmically calculating the probability that given existent vertices of a network will define new associations; it is based on the assumption that the past and the present behavior of the net can indicate what may happen in the future.

We apply these measures considering the relationships listed in Table 2.

4. MULTIMODAL ANALYSIS OF DBLP

We coin the term *multimodal analysis*, in the data mining context, referring to it as any ensemble of techniques from at least three distinct categories of data analysis. In this work, we instantiate this modality of knowledge discovery considering statistical, topological, algorithmic, and algebraic techniques; as explained in the following sections.

4.1 Weakly-connected components distribution (WCC)

Initially, we have verified interesting facts about WCC distribution over DBLP. Figure 2 depicts the case for co-authorship; in the figure we see that only 13% of the nodes form small components with up to 30 nodes; meanwhile, a giant component formed by nearly 87% of the nodes ($\sim 10^6$ authors) defines a huge network in which researchers of the entire world share scientific expertise. Also interesting, the smaller components account for over 44.000 co-authorship sub-networks ($\sim$ 120.000 authors); a fact not so easy to explain, but that, probably, corresponds to the eventual researchers that got involved with the academy only for the sake of obtaining a degree, without further research activity. Another possible explanation, comes from the white papers divulged by the industry; these papers aim at divulgating new techniques or processes without a strict scientific contextualization.

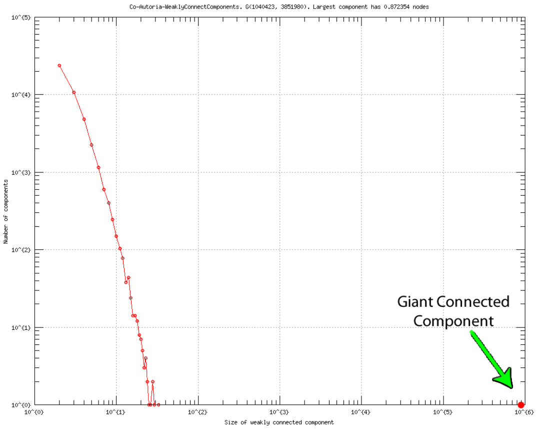


Figure 2: Co-authorship weakly-connected components (WCC) distribution.

4.2 Average Clustering Coefficient (ACC)

An expected property of co-authoring graphs is the presence of significant values of ACC. Through the co-authorship Node degree × ACC plot – Figure 3, we verified that this property is prominent for DBLP. High values (close to 1) of ACC are observed only for nodes with degree up to around 10, with values decreasing along the sequence so that the following power law is observed: $ACC \propto degree^{-1.06}$. It means that nodes with degree up to 10 tend to have their connections highly interconnected, hence, they all together tend to form clusters (sets of nodes highly interconnected). This tendency decreases along with the increase of the degree.

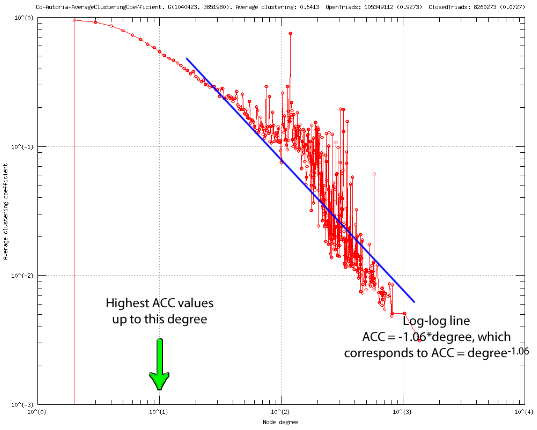


Figure 3: Co-authorship Average Clustering Coefficient (ACC) against node degree.

We explain the ACC behavior of the co-authorship because, supposedly, authors tend to collaborate to co-authors of their co-authors, forming triangles in the network. Another explanation comes from the fact that hierarchically organized graphs tend to present progressively decreasing ACC's [18]. In addition, older authors (advisors) tend to have bigger and bigger collaboration networks and degree, what means that they are less likely to be part of one well-defined and highly interconnected cluster; rather, they tend to be connected to multiple other sub graphs that tend to be sparse as their sizes increase.

4.3 Densification

From the degree-distribution plot – Figure 4, one can see that, along time, DBLP's degree distribution is obeying to a power law with exponent $\gamma = -1.36$ (approximately). According to Leskovec *et al.* [12], this fact indicates that as more nodes enter DBLP, more edges appear following to the exponential relation:

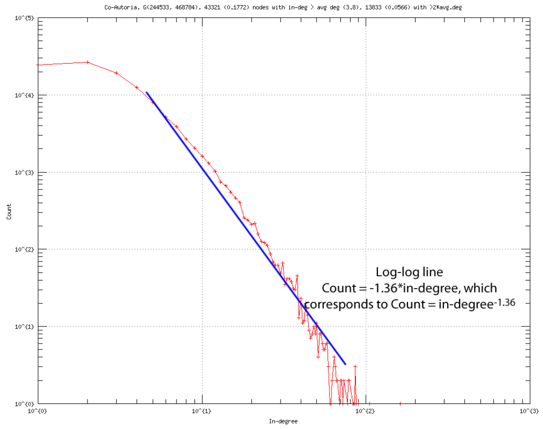
$$e(t) \propto n(t)^a \qquad (1)$$

where $e(t)$ is the number of edges and $n(t)$ is the number of nodes, and a is a specific exponent dictated by the slope $\gamma = -1.36$ of the degree distribution.

More specifically, according to Leskovec, $a = 2/\gamma$ for the case in which $1 \leq \gamma \leq 2$, or $a = 2/1.36 = 1.47$ for DBLP. This process is called densification, which answers for the intensity of according to which new edges appear in the network.

One can observe similar power law densification in other environments, as the web, where links correspond to new edges. However, one might wonder why the number of edges grows *exponentially* in an environment where new edges are not as cheap as in the web; but that, rather, depend on the publication of lengthy elaborated papers. We presume that two facts help to explain this tendency. First, master and Ph.D. titles were, originally, certifications of knowledge and experience - granted on an *ad hoc* basis; in the last decades, though, they became regular courses with well-defined time schedules and expected production. Consequently, a demand for "where to publish", rather than "what to publish", was created. This fact has led to a scientific literature that is prolix and that presents varying degrees of quality – more does not necessarily means better. Second, private and public science-funding agencies have demanded results in the form of publications as a condition for keeping up with their financing. Hence, researchers are pressed for numbers, be it good or bad – a straight consequence of this fact is the increasing number of authors per paper; in some cases, the so-called "academic collaboration" does not always corresponds to intellectual guidance and labor, but to co-financing and personal exchange as a means to increase one's production.

DBLP obeys to a stable power-law distribution. Although the network is huge and mostly connected, new nodes cannot alter the main properties of the entire structure. One might think that a massive introduction of new nodes and edges could do so; however, as big as it can be, new generations of nodes and edges are still small if compared to the mega structure of DBLP. The equilibrium presented by DBLP is



Log-log line
Count = -1.36*in-degree, which
corresponds to Count = in-degree$^{-1.36}$

Figure 4: In-degree distribution of the co-authorship activity in DBLP – the higher the in-degree (the co-authorship), the smaller the number of authors according to power law $Count \propto indegree^{-1.36}$.

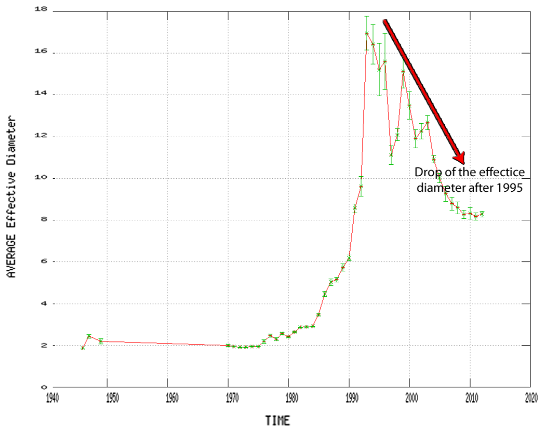
observed in other systems as well [13] – as the respiratory system, automobile networks, and other social networks, being an instance of a well-defined natural phenomenon. As observed in the seminal work of Faloutsos [6], the exponent of its correspondent power-law distribution personifies this equilibrium.

4.4 Diameter

By inspecting the effective diameter evolution of the co-edition network, Figure 5, it is possible to see that the effective diameter starts to shrink after a certain point in time – around the year 1995. This fact suggests that, before this time there were new publication vehicles – with new editors appearing as well – showing up in the research community until a pick, when the same editors started to edit/co-edit for existing vehicles. It also suggests that after then, the number of new edges entering the network was much higher than the number of new nodes, what initiated a densification period. A possible explanation is that the committees of editors tend to have the same members that alternate between a limited set of possible committees, year after year. As so, the distance in between any two editors tends to decrease along the time. Possibly, this is the case because editing publications is a task that demands higher experience and expertise, characteristics of a limited set of researchers; moreover, as the number of editors is quite limited, the activity of edition is a matter of dispute in the community, what poses additional obstacles for newcomers.

4.5 Co-authoring predictability

Also relevant is to evaluate how predictable DBLP is. We do this by means of link recommendation techniques together with measures of accuracy. Here we verify this property by extracting metrics from the co-authoring network of DBLP; specifically, for each author we extract Number of common authors (a), Jaccard's coefficient (b), Preferential attachment (c), Adamic-Adar coefficient (d), Resource allocation index (f), and Local path (g) – please check the work of Gimenes *et al.* [7] for details. These metrics are then used with supervised machine learning classifiers [19] J48, Naïve



Drop of the effectice
diameter after 1995

Figure 5: Co-edition effective diameter evolution.

Bayes, Multilayer Perceptron, Bagging, and Random Forest, all of them available in the Weka framework, developed by the University of Waikato [5]. That is, we give some existing (past) co-authorings from 1995 to 2005 to the algorithms and they are expected to predict ("future") new co-authorings from 2006 to 2007 – of course, the algorithms are not given the answer, we use it to measure accuracy based on 10-fold cross-validation.

We performed link prediction for different profiles of authors as indicated by their degree; we considered authors with at least $d \geq 1$ existing co-authorings, with at least $d \geq 2$ existing co-authorings, and so on until at least $d \geq 8$ existing co-authorings. This extra parameter was set as a way to compare the predictability of the less ($d \geq 1$) and of the more ($d \geq 8$) active authors – pondering whether their profile is a relevant factor.

The accuracy of this methodology indicates how predictable DBLP is. We use measures Precision, Recall, F-Measure, and Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC). The higher the values the more predictable DBLP is because the metrics will indicate that we were successfully able to foresee new co-authorings. Table 3 presents the results for each minimum degree d (line sets) and for each classifier (line) across each metric (columns). In the table we can see the highest values in bold; the best values (Random Forest classifier) are all above 0.8 (maximum is 1.0 – 100%), ranging from 0.867 to 0.977. Despite some classifiers having performed not well (NB and MLP), the others indicate a quite reasonable predictability, especially for minimum degrees above 1.

The predictability of DBLP indicates a network in which researchers do not interact intensely with fields not related to their own (further in the graph), what translates to low multidisciplinarity; it also indicates an interaction pattern in which authors tend to collaborate with the same authors recurrently or with co-authors of their co-authors (in triangle fashion). The fact that less active authors ($d \geq 1$) are harder to predict possibly corresponds to casual researchers that abandon the academy after getting their degree, and that do not aim at interacting with no other authors at all.

d	Classifier	PREC	REC	F-MEAS	AUC
1	J48	0.723	0.706	0.7	0.764
	NB	0.741	0.585	0.505	0.626
	MLP	0.562	0.555	0.541	0.593
	Bagging	0.809	0.8	0.798	0.887
	RF	**0.877**	**0.868**	**0.867**	**0.939**
2	J48	0.787	0.759	0.753	0.817
	NB	0.777	0.598	0.52	0.648
	MLP	0.628	0.618	0.61	0.639
	Bagging	0.84	0.83	0.829	0.913
	RF	**0.914**	**0.903**	**0.902**	**0.977**
4	J48	0.852	0.845	0.844	0.87
	NB	0.773	0.585	0.499	0.704
	MLP	0.715	0.714	0.713	0.735
	Bagging	0.846	0.841	0.841	0.925
	RF	**0.917**	**0.913**	**0.912**	**0.974**
6	J48	0.827	0.771	0.761	0.79
	NB	0.778	0.601	0.526	0.727
	MLP	0.695	0.679	0.672	0.74
	Bagging	0.844	0.83	0.828	0.913
	RF	**0.897**	**0.888**	**0.887**	**0.972**
8	J48	0.861	0.839	0.836	0.867
	NB	0.786	0.626	0.566	0.741
	MLP	0.725	0.719	0.717	0.785
	Bagging	0.883	0.866	0.865	0.94
	RF	**0.914**	**0.908**	**0.907**	**0.971**

Table 3: Link prediction accuracy (Precision, Recall, F-Measure, and Area Under Curve) of five supervised machine-learning classifiers over DBLP considering years 1995 through 2005 for training, and years 2006 through 2007 for testing. The tests were performed for author profiles of degree $d \geq 1$, $d \geq 2$, $d \geq 4$, $d \geq 6$, and $d \geq 8$.

4.6 Counting and algebraic analysis

Lastly, we perform counting, the most common and useful kind of analysis over DBLP. To do so, we consider the network as a bipartite graph with sets authors and articles in which the edges are time-stamped (year). Over this graph, we count two things for each author: *accomplishment* the number of distinct years when she/he published at least one article; and *silence* the biggest number of consecutive years without publishing. The first is an indicator of how active an author is, the bigger the *accomplishment* the more works were accomplished and the more productive is his career. The second indicates how constant the author is, the bigger the *silence*, the less regular is her/his activity. Both metrics range from 0 to 50 years (nearly the longest academic career).

Figure 6 shows the histogram for *accomplishment* and *silence*. There one can see that the majority of DBLP's authors is low-active with accomplishments between 1 to 4 years. Meanwhile, the histogram of Silence shows that a great share of DBLP is quite constant in what concerns their production regularity, with most authors having published something in the last year (silence 0) or in the years before (silence 1); however, a considerable share of authors has silence periods between 2 and 20 years. Above 20 years we can safely consider that either the author has abandoned the academy or has passed away, in contrast to new ($silence \leq 5$) active students.

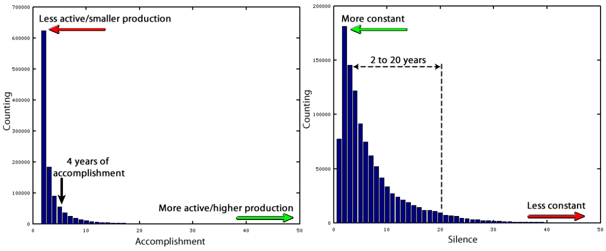


Figure 6: Counting (histogram) of metrics *Accomplishment* and *Silence* (both 0 to 50 years) for all the authors of DBLP.

We consider that these two metrics can provide an interesting characterization the overall profile of DBLP as they can combine to a single metric that translates to both productivity and constancy. In order to combine the two metrics we considered that they have inverted semantics: for one "more" is desirable, while for the other "less" is desirable, as depicted by the green and red arrows in Figure 6. Algebraically speaking, one is proportional and the other is inversely proportional, by combining both of them we got the metric *Sao Paulo's Importance*[4] as expressed by Equation 2. In this equation, the logarithm of the accomplishment stands for the magnitude of the metric, and the square root of the silence penalizes the importance of a given author.

$$SP's\ Importance = \frac{1}{\sqrt{silence + 1}} * log(Accomplishment) \tag{2}$$

Sao Paulo's Importance, or simply *Importance*, provides a number that translates how important a given author is; in the context of this work, importance refers to the insertion of edges in the network (publication and co-authoring), rather than to the relevance of articles. Having this metric in mind, we created a plot that figures the panorama of DBLP with respect to the hole of its authors. In Figure 7(a), it is possible to see the raw curve of metric *Sao Paulo's Importance*; from the figure, it is possible to understand the behavior of the metric, which favors low silence and high accomplishment – reddish regions (high importance) as highlighted with a circle. In Figure 7(b), we present the counting (histogram) of authors per pair of Silence and Accomplishment – we use the same color mapping of figure (a), with high importance expressed in reddish colors.

This figure expresses how the *Sao Paulo's Importance* is instantiated in DBLP; just a few authors (reddish region indicated by arrow) have high importance, with the great majority presenting low importance. It is interesting to see that there is a great share of authors with low silence ($\leq$ 5 years) and low accomplishment ($\leq$ 5 years) – this specific region defines a pick at the left-hand lower corner of the plot. Possibly, these authors refer to students that are still doing their PhD course, or that have recently finished it. This finding reveals how Computer Science is dependent on

[4] in reference to the state of Sao Paulo in Brazil, which hosted this research.

casual researchers, and also how competitive it is, since just a few authors are able to migrate to the more important region of the plot.

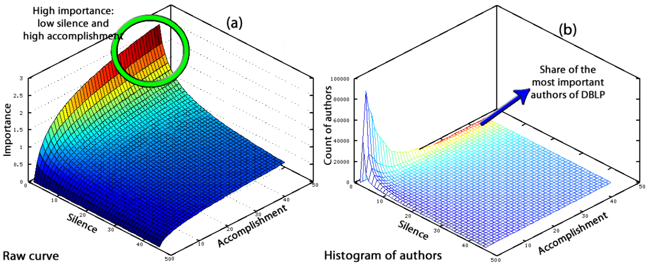


Figure 7: Plot of metric *Sao Paulo's Importance*. (a) Raw curve of Equation 2. (b) Counting (3D histogram) of authors in relation to the possible values of metric *Sao Paulo's Importance*.

5. CONCLUSIONS

We introduced a multimodal analytical process defined as an ensemble of statistical (degree, and weakly-connected components distribution), topological (average clustering coefficient, and effective diameter evolution), algorithmic (link prediction/machine learning), and algebraic techniques to reveal non-evident features of network-like data, including networks of co-authoring, recommendation, computer routing, social interaction, protein interaction, to name a few. We demonstrated our process over the DBLP repository of Computer Science publications pointing out critical discoveries about its behavior from multiple perspectives.

Our methodology introduces an innovative course of action based on techniques that, although apart, can be used in complementary fashion. This kind of analytical approach is challenging due to its demand for heterogeneous technical knowledge and due to the diversity of the outputs to interpret. Nevertheless, it demonstrated a relevant potential in the form of ample interpretations of DBLP. These interpretations, in turn, can bring light to the research activity, possibly assisting in the decision making of funding agencies and academic personnel.

Acknowledgments

This work received support from Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (560104/2010-3), and Fundacao de Amparo a Pesquisa do Estado de Sao Paulo (2011/13724-1, 2013/03906-0, 2013/10026-7).

6. REFERENCES

[1] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. Friendship prediction and homophily in social media. *ACM Trans. Web*, 6(2):9:1–9:33, 2012.

[2] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user navigation and interactions in online social networks. *Information Sciences*, 195:1 – 24, 2012.

[3] J. Bollen, H. Van de Sompel, A. Hagberg, L. Bettencourt, R. Chute, M. A. Rodriguez, and L. Balakireva. Clickstream data yields high-resolution maps of science. *PLoS ONE*, 4(3):e4803, 03 2009.

[4] M. A. Brandao, M. M. Moro, G. R. Lopes, and J. P. M. Oliveira. Using link semantics to recommend collaborations in academic social networks. In *WWW*, pages 833–840, 2013.

[5] J. Breslin and S. Decker. The future of social networks on the internet: The need for semantics. *IEEE Internet Computing*, 11(6):86–90, 2007.

[6] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

[7] G. Gimenes, H. Gualdron, T. R. Raddo, and J. F. R. Jr. Supervised-learning link recommendation in the dblp co-authoring network. In *IEEE PerCom Works. on Social and Community Intelligence*, pages 563–569, 2014.

[8] J. Huang, Z. Zhuang, J. Li, and C. L. Giles. Collaboration over time: characterizing and modeling network evolution. In *Web search and web data mining*, pages 107–116, 2008.

[9] Z. Huang, Y. Yan, Y. Qiu, and S. Qiao. Exploring emergent semantic communities from dblp bibliography database. In *ASONAM*, pages 219–224, 2009.

[10] D. A. Huffaker, C. Teng, M. P. Simmons, L. Gong, and L. A. Adamic. Group membership and diffusion in virtual worlds. In *Social Computing*, pages 331–338, 2011.

[11] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. on the Web*, 1(1), 2007.

[12] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1), Mar. 2007.

[13] J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD*, pages 177–187, 2005.

[14] L. Leydesdorff. Betweenness centralityŤ as an indicator of the ŞinterdisciplinarityŤ of scientific journals. *Information Science and Technology*, 58:1303–1309, 2006.

[15] M. Li, J. Wang, and J. Chen. A graph-theoretic method for mining overlapping functional modules in protein interaction networks. In *Bioinformatics research and applications*, pages 208–219, 2008.

[16] E. Lim, D. Correa, D. Lo, M. Finegold, and F. Zhu. Reviving dormant ties in an online social network experiment. In *Weblogs and Social Media*, pages 361–369. AAAI Press, 2013.

[17] B. A. Osiek, G. Xexeo, A. S. Vivacqua, and J. M. de Souza. Does conference participation lead to increased collaboration? a quantitative investigation. *IEEE Computer Supported Cooperative Work in Design*, pages 642–647, 2009.

[18] E. Ravasz and A. Barabasi. Hierarchical organization in complex networks. *Phys Rev E*, 67(2):1–7, 2002.

[19] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining - Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2011.