WEB CRAWLER FOR SOCIAL NETWORK USER DATA PREDICTION USING SOFT COMPUTING METHODS

José L. V. Sobrinho[1], Gélson da Cruz Júnior[2] and Cássio Dener Noronha Vinhal[3]

[1,2,3]Faculty of Electrical and Computing Engineering, Federal University of Goiás, Brazil

ABSTRACT

This paper addresses how elementary data from a public user profile in Instagram can be scraped and loaded into a database without any consent. Furthermore, discusses how soft computing methods such neural networks can be used to determine the popularity of a user's post. Conclusively, raises questions about user's privacy and how tools like this can be used for better or for worse.

KEYWORDS

Social Network, Instagram, Business Intelligence, Soft Computing, Neural Network, User Privacy, ETL, Database, Node.js, Data Analysis

1. INTRODUCTION

In the past few years, the world has been experiencing a remarkable advance in the field of data technologies. As hardware and software limitations are being overcome, storage and analysis of huge datasets are not a problem anymore. Meanwhile, seizing the peak of the information age, companies are massively investing in analysis tools that help to improve the understanding of people and its behavior.

Social networks are an inexhaustible source of data about its users. They provide insights into likes and dislikes, affinities and many other aspects that maybe only who genuinely know these people can identify. This matter is raising serious concern about privacy and promoting an important debate about the efforts of companies to secure user information.

Players like Facebook and Instagram have been implementing mechanisms to protect and extend the privacy of its customers. They are constantly restricting access to the data and also decreasing the available amount for authorized developers. This access limitation is not only motivated by the pressure to preserve user's privacy but also because this information is a priceless asset for these companies.
Even though with all the efforts, public profiles inside social networks can be scanned by a third party without any consent. In most cases, it's not necessary to have access to the API or even credentials to the network itself. Taking advantage of this breach, we can make possible to determine many aspects of a specific user.

This task is generally accomplished by crawlers, mechanisms that scrapes internet pages looking for raw data. Once the data is clustered, an ETL (extract, transform and load) process is executed to make sure that meaningful information about the user is correctly stored for further analysis. In Instagram, for example, it's possible to extract data like full name, posts, and more.

DOI: 10.5121/ijcsit.2019.11207

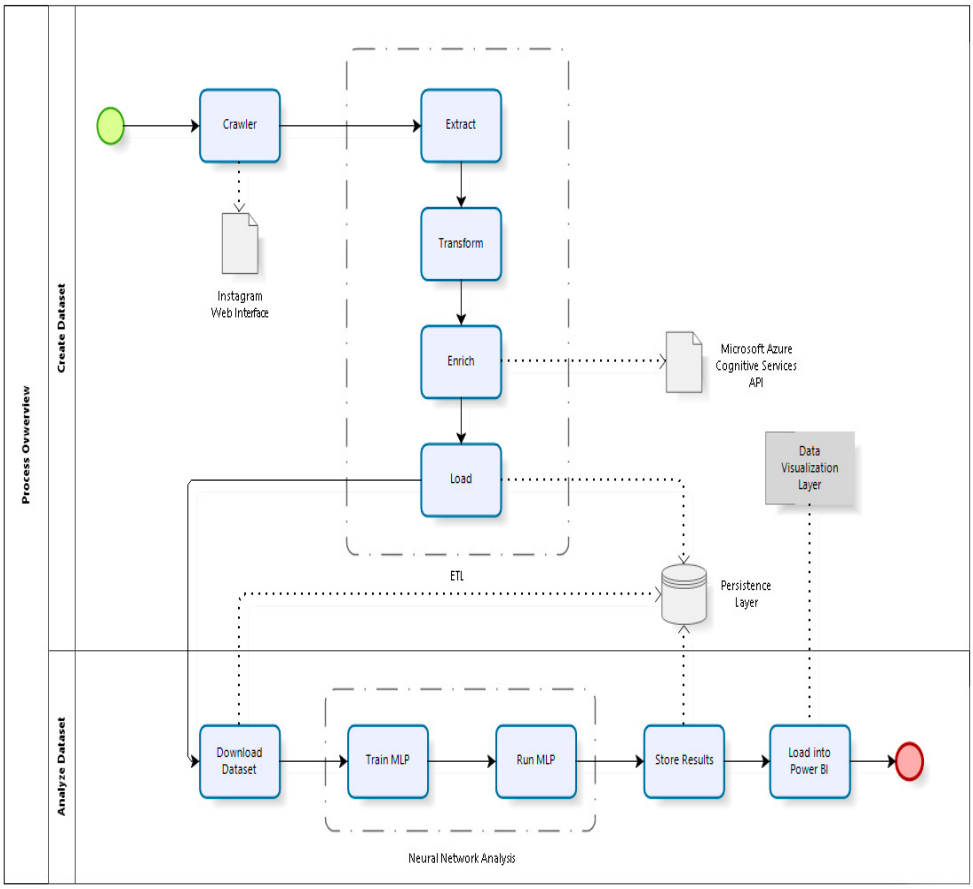

Figure 1. Four major steps process overview

In the next sections, this article introduces a process where data from a public user Instagram profile is extracted, transformed, enriched, and loaded into a database, providing basic information about a user. Moreover, the paper addresses how soft computing methods can be implemented to determine, for example, the popularity (in likes) of a user's post. Results are shown to prove how effective this approach is and to raise questions about user's privacy.

2. THE CRAWLER

Web crawlers are typically defined as "a system for the bulk downloading of web pages" as in [1]. Commonly referred to as a robot or a spider, they are used for a diversity of purposes. In fact, they are one of the core pieces of web search engines. They are also used for web archiving where huge sets of web pages are systematically downloaded and stored for posterity. Additionally, they are used for data mining, where web pages are examined and data analytics is performed on them.

In this project, the crawler is the first piece of a four major steps process (Fig. 1). It's written in JavaScript and executed in Node.js ®, a JavaScript runtime built on Google Chrome's V8 JavaScript engine. Considering its web browser background, the Node.js ® is an indeed appropriated environment for running tasks like this.

The crawler algorithm essentially works sending an HTTP request (GET method) to a specific public user profile URL on the social network (e.g., htttps://instagram.com/username). Since the Instagram uses the Graph API, an equivalent approach of the GraphQL, remnant JSON (JavaScript Object Notation) data can be found on the page source code. When a successful response arrives, the algorithm then scraps the page looking for the user information.

GraphQL is a query language for APIs that provides a runtime for fulfilling queries with data sources. It's a project maintained by Facebook which delivers a complete and understandable description of the data in the API, giving clients the power to ask for exactly what they need and nothing more. Additionally, as explained in [2], makes it easier to evolve APIs over time and enables powerful developer tools. The Instagram Graph API is a standard REST (representational state transfer) API, alike to the GraphQL, that supports basic CRUD (create, read, update, and delete) operations, as described in [3]. The standard way to get data from Instagram, or any other application, is to use its API, but that's not the point of this project. Even not being a user it's possible to scrape the API data from a crawled page.

The compiled raw data is forwarded to the ETL where it will be arranged in two different entities. The first of them is the User, which contains information like full name, biography and profile image. The second is the Media, that stores comments, likes, hashtags, people who interacted, and more from the user latest posts.

3. THE ETL

The SAS Institute, a market leader company in advanced and predictive analytics, states in [4] that "ETL is a type of data integration that refers to the three steps (extract, transform, load) used to blend data from multiple sources." Throughout the process, data is extracted from a source system, transformed into a format that can be analyzed, and loaded into other system. It's frequently employed to build data warehouses in many applications. In other words, the acronym ETL refers to a process where data from disparate sources can be programmatically clustered to analyze and discover insights. In this project, a likewise JavaScript algorithm extracts and transform the raw data from the crawler into valid JSON objects. This step of the process works in two major stages as follows. The ETL is a recurrent procedure; ergo it's important to clarify that the data load process is incremental.

3.1. EXTRACT AND TRANSFORM

As discussed in the crawler section, data is fundamentally arranged in two different entities, and its attributes are further detailed right next. Nevertheless, this is an elementary representation, since in the database information is allocated over many different collections.

- User – user ID, username, full name, biography, profile picture URL, external URL (if applicable), a private profile flag, followed by count, follows count, and media count.
- Media – media ID, caption, caption hashtags, comments count, likes count, a is video flag, media URL, type of media, users tagged, users who commented, location, and post date.

Right after this point, the data is enriched using the Microsoft Azure Cognitive Services API, a cloud-based service that provides cutting-edge image analysis algorithms. This step has two main functions: to detect and recognize faces with attributes on user profile images and to classify user media on a predefined set of categories. According to the specification in [5], the API detects up to 64 human faces with high precision on related attributes such as gender, age, hair color, and more. Microsoft makes available a free tier of this service, where an application can consume up

to 20 transactions per minute and 30,000 transactions per month with no cost, what is more than enough for the established purpose. There are many other lines to achieve this same goal (e.g., opencvn4nodejs library), but considering the project commercial approach, priority was given to an enterprise level solution. Objectively, this step is meant to be performed just before the load into the database. In other words, as soon as the data is ready to be stored, a script is executed (also in Node.js ®) to identify the user in its profile image
.

3.2. LOAD INTO DATABASE

This project uses MongoDB, an open source cross-platform document-oriented database and by far one of the most popular NoSQL databases worldwide. Essentially, there are four different collections to store the user information. At the end of this step, data from a specific Instagram profile is normalized, stored and ready to be analyzed.

Even as simple as it seems, a lot of work has been done to reach this mark. The available data is suitable for many applications, such as tracking an account growth over the time or simply identifying which posts get higher engagement (likes and comments) based on hashtags.
The application, when starting, establishes the connection with the database in a very simple process, using the library mongoose.js, to guarantee the correct functioning of the ETL. Widely used in projects involving Node.js and MongoDB, the library is based on schemas that model application data, natively offering a system of type conversion, validation, query creation, and hooks to business logic.

4. NEURAL NETWORK ANALYSIS

This project implements the brain.js library, that brings together a set of different implementations of artificial neural networks, written in JavaScript and designed for server-side applications. This step of the process implements a multilayer perceptron network based on the hypothesis that the popularity of a user's post can be estimated based on the popularity of the publications of the people who interacted. Although it sounds a bit confusing, the goal is to train the network with the media of the people who interacted with the main profile (the one being analyzed), and once the learning process is complete, the network estimates the popularity values of the posts.

The neural module connects to the database and loads the information stored by the ETL. As soon as this task is completed, the first step to be performed is the selection of which records will compose the training set. It is important to evaluate within the universe of the users who interacted with the main profile which of them have amounts of followers compatible with the one of the user being analyzed, avoiding the network to be trained with profiles that present a very discrepant audience.

As important as determining which profiles are good choices for composing the training set is determining which media are also good references. As known, in Instagram it’s possible to share images and videos, which in terms of engagement, present copiously different numbers. Photos have a strong appeal to likes, which is exactly the number that the MLP should predict, while videos have a strong appeal to views. Thus, it seems to make no sense to use the same network, with the same parameters and the same training set to predict the engagement of these two different types of posts. In this step, the algorithm only selects images to compose the training set.

Next, the data for the training of the network are prepared and standardized considering the minimum and maximum values of each entry, ensuring that these are in the range of 0 to 1. The

input data standardization is important so that the activation functions of the MLP do not operate in saturation intervals.

The data acquisition process builds a very rich base of information. Nevertheless, it's important to choose wisely which attributes to use as input to the MLP, ensuring that it will produce efficient results that are close to the real numbers. The learning process is adjusted to perform until one of the following criteria is met: reach the maximum number of 100,000 iterations or reach the error mark less than or equal to 0.5%. It was also determined a learning rate of 30%.

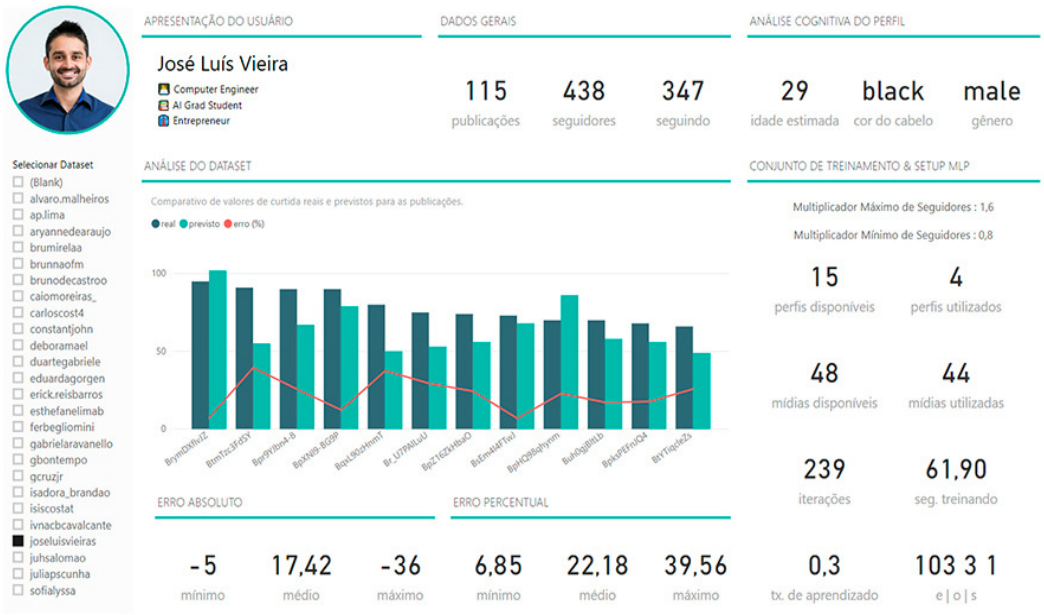

Figure 2. Analysis of the Author's Instagram Profile

The neural network has 103 neurons in the input layer and, since the goal is to determine the post popularity, it has only a single neuron in the output layer. Regarding the number of hidden layers (and their neurons), it was specified that for each execution of the neural analysis, different network configurations must be tested, and the one with the lowest error will be selected as ideal. There is no consensus in the academic community on how to determine the architecture of these hidden layers, despite the large number of scientific productions and theories that permeate this subject. In this project, the structure was defined after testing three different assumptions.

- The number of neurons in the hidden layer must be in the range between the size of the input layer and the size of the output layer.
- The number of neurons in the hidden layer should be 2/3 the size of the input layer, plus the size of the output layer.
- The number of hidden neurons must be less than twice the size of the input layer.

The strategy to use a single hidden layer and to vary its number of neurons was adopted after the first tests indicated that the only perceived impact in increasing the amount of these layers was negative, because it also increased the execution time and did not produce better outputs. By the end of this step, the algorithm saves the results found in the same database, as well as all the criteria used to obtain them.

5. DATA VISUALIZATION

On the peak of the whole process, there's the final step, where the collected data has to be presented in a meaningful way for those who will use it (Fig. 2). There are several possible ways to display the information, and since it's a sole matter of design, the project sticks with another free tier Microsoft solution, the Microsoft Power BI. It is a business analytics solution that lets users visualize data and share insights across its organization, or embed them in an app or website.

Microsoft proudly showcases in [6] that Power BI allows to connect to hundreds of data sources and bring data to life with live dashboards and reports. In fact, "for the 11th consecutive year, Microsoft has been positioned as a leader in Gartner's 2018 Magic Quadrant for Analytics and Business Intelligence Platforms," as pointed out in [7].

As important as any other step of the process so far, this is the simplest part of the mechanism. Essentially, it aggregates the information from all the searched users in a user-friendly dashboard, removing all the complexity away. Actually, dashboards are highly customizable and as easy to use as an excel spreadsheet. The Power BI connects to the MongoDB instance and from that point on it creates a workspace where different reports, which group different graphs and indicators, can be built and stored.

There are data loading processes of the type ELT (extract, load and transform), where the due transformations are carried out in the destination base, and that is what happens after reading the information by the Power BI. From that point on, data stored in three collections of documents in MongoDB are broken down into 17 different tables to ensure that proper relationships between entities are established and that indicators can be presented to the end user.

6. RESULTS

Twenty-five different Instagram users were selected for analysis. From these, 2,522 interactions were identified, allowing the tracer to collect information from 606 different profiles and, consequently, 6,522 posts. The mechanism concluded that among the users of the analysis, 13 are women, 10 are men and 2 could not be identified (manually verified as female), with ages estimated to be between 22 and 43 years old (Fig. 3).

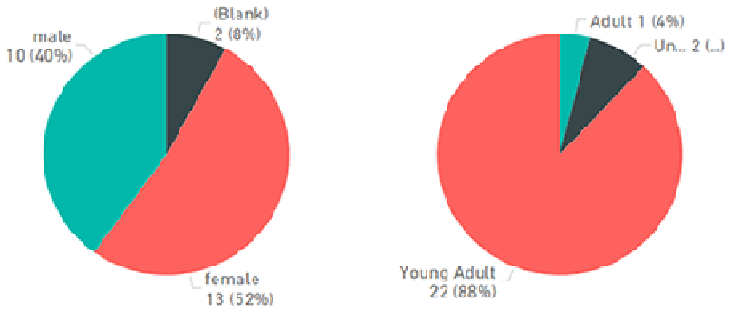

Figure 2. Dataset gender and age distribution

The results found are encouraging, since the mean error of 26% in a sample of 25 users (and their respective 291 publications) categorically corroborates the established hypothesis. In this way,

the popularity of a user in a social network, more precisely Instagram, can be determined by analyzing the popularity of their cycle of friendships. There are certain profiles in which the applied method does not work with the efficiency of the others, however, this is due more to certain isolated characteristics from the posts of these users than to aspects of the interactions or the profile itself.

Results shown in Table 1 demonstrates the minimum, average, and maximum errors (absolute and percentual) of the overall dataset. Meanwhile, results shown in Table 2 reveals the same errors for each one of the 25 analyzed users.

Table 1. Overall dataset error.

Min. Error Abs	Avg. Error Abs	Max. Error Abs.	Min. Error (%)	Avg. Error (%)	Max. Error (%)
0.00	42.60	503.00	0.00	26.05	282.76

Table 2. Results overview.

Username	Min. Error (%)	Avg. Error (%)	Max. Error (%)
alvaro.malheiros	0.92	24.68	50
ap.lima	4.17	24.93	53.09
aryannedearaujo	0.84	20.26	55.17
brumirelaa	0.62	31.59	121.88
brunnaofm	2.78	16.82	36.24
brunodecastroo	0.49	32.25	130.07
caiomoreiras_	1.32	44.32	282.76
carloscost4	0.82	50.37	102.17
constantjohn	1.28	23	83.87
deboramael	1.54	18.67	36.27
duartegabriele	2	39.59	128.21
eduardagorgen	2.22	36.31	196.6
erick.reisbarros	4.17	31.59	82.18
esthefanelimab	2.47	21.72	53.45
ferbegliomini	0.81	24.9	55.29
gabrielaravanello	0.6	14.26	35.87
gbontempo	0	31.33	65.96
gcruzjr	3.53	16.07	35.14
isadora_brandao	1.67	22.46	53.24
isiscostat	1.38	21.22	43.15
ivnacbcavalcante	4.47	20.79	39.55
joseluisvieiras	6.85	22.18	39.56
juhsalomao	2.12	17.54	40.82
juliapscunha	5.36	30.07	60.61
sofialyssa	0	14.32	47.78

Besides all the technical apparatus where this project relies on, one of the main discussion points of this paper is to raise questions about how tools like this can be used for better or for worse. At this point, one can say that public profiles on Instagram can reveal a large amount of information about its user. Nevertheless, private profiles can also be crawled, and a smaller portion of data can be retrieved. This rationale extends to other social networks, like Facebook and Twitter for instance.

Companies can take advantage of tools like this in very different ways. For example, a beauty products company pretends to release a new shampoo for blonde women. They have a pretty well-defined target (gender, hair color and maybe age) and need to set strategies to reach out to its audience. A system like this could easily find segment users for a marketing campaign. Data from users can generate a great deal of money. Companies can strategically use this information to show the right advertisement for the right client.

Another way to illustrate what can be done is to analyze profiles of users who tagged a specific location on their Instagram profiles. That could provide a comprehensive insight into the people who frequent the site. Socially speaking, a town's poor neighborhood place would have a completely different audience than a place on a rich side of a town. Tools like this are typically used for marketing and social applications, but it's no wrong to say that they're quite intrusive.

7. CONCLUSIONS

Although users agree with specific rules for using social networks, still impresses how it's possible to manipulate data from someone without any consent. The truth is that if someone doesn't want to be exposed, then he should stay out of the Internet and other social networks.

Even if the last affirmation may sound exaggerated, one who carefully analyzes the Terms of Use of Instagram, as seen in [8], will notice in its Data Policy the following statement: "public information can be seen by anyone, on or off our products, including if they don't have an account." This excerpt not only corroborates the defended point of view but also makes clear that public information "includes your Instagram username; any information you share with a public audience; information in your public profile on Facebook; and content you share on a Facebook Page, public Instagram account or any other public forum, such as Facebook Marketplace," as can be perceived in [9].

Instagram makes crystal clear that "public information can also be seen, accessed, reshared or downloaded through third-party services such as search engines, APIs, and offline media," what in other words means that they're probably aware of crawlers and any other automated systems that may compromise their user's privacy. Finally, this paper endorses the rationale surrounding good computer ethics and how negative intrusive technology is nowadays, questioning how "technologies are moving from knowing what we are doing (and where) to knowing who we are."

8. RELATED WORK

The article Instagram Popularity Prediction via Neural Networks and Regression Analysis, produced by students at Princeton University, is one of the closest to the approach proposed in this paper. It establishes a question as to which of the factors analyzed, the content of the image or the social metadata, has a dominant predictive power to determine the popularity of a post on Instagram. It analyzes a set of 3411 images that use a given hashtag, prioritizing landscape photographs, where the authors sought to find posts whose engagement was neutral, consistently based on how nice the picture is, unlike images depicting people, animals or texts.

The available results of the evaluation of social metadata suggest that the number of comments has the strongest predictive power, corresponding positively to the popularity of a publication. This makes perfect sense, since a greater number of comments indicates a greater interaction of the users, since they cannot unlike a post. In fact, this is a critical point of the article because it is as if the authors used a result as input to predict another result. Using comments as input criteria to predict the number of likes seems inconsistent, and in this approach, it is unfeasible to predict, for example, the popularity of a publication before it is published.

Results that deal with the visual composition analysis of the images, produced by a convolutional neural network, show that it was unable to provide reliable results. The reason may be in the uniformity of the dataset, since the photos have strong similarities between them. In this sense, training a neural network to recognize subtle aesthetic differences is a truly complex task, and in the context of the article, in many cases, significant differences in metadata between photos would have greater predictive power than in visual composition data.

REFERENCES

[1] C. Olston and M. Najork, Web Crawling, in Foundations and Trends in Information Retrieval Vol. 4, No. 3 (2010) 175–246.

[2] GraphQL – Get started with a query language for your API [Web page]. Retrieved September 21, 2018, from https://graphql.org/

[3] Instagram API [Web page]. Retrieved September 21, 2018, from https://developers.facebook.com/docs/instagram-api/overview/

[4] What it is ETL [Web page]. Retrieved September 21, 2018, from https://www.sas.com/en_us/insights/data-management/what-is-etl.html

[5] The Face API Service [Web page]. Retrieved September 21, 2018, from https://docs.microsoft.com/en-us/azure/cognitive-services/face/overview

[6] What is Power BI [Web page]. Retrieved September 21, 2018, from https://powerbi.microsoft.com/en-us/what-is-power-bi/

[7] Ulag, A. M. (2018, February 27). Gartner recognizes Microsoft as a leader in BI [Web blog post]. Retrieved September 21, 2018, from https://powerbi.microsoft.com/pt-br/blog/gartner-recognizes-microsoft-as-a-leader-in-analytics-and-bi-platforms-for-11-consecutive-years.

[8] Instagram Terms of Use [Web page]. Retrieved September 21, 2018, from https://help.instagram.com/581066165581870/

[9] Instagram Data Policy [Web page]. Retrieved September 21, 2018, from https://help.instagram.com/519522125107875/

[10] K. Michael and M. G. Michael, Computing Ethics – No Limits to Watching, in Com. of the ACM Magazine Vol. 56, No. 11 (2013) 25-28.

[11] C. J. Qian, J. D. Tang, M. A. Penza, C. M. Ferri, in Instagram Popularity Prediction via Neural Networks and Regression Analysis [Web page]. Retrieved September 21, 2018, from http://cjqian.github.io/docs/instagram_paper.pdf

AUTHORS

José Luís Vieira Sobrinho, Computer Engineer graduated from the Federal University of Goiás (Brazil) and former exchange student at the State University of New York at Oswego (United States) pursuing a Master's Degree.

Gélson da Cruz Júnior, He holds a bachelor's degree in Electrical Engineering from Universidade Estadual Paulista Júlio de Mesquita Filho (1990), a Master's degree in Electrical Engineering from the State University of Campinas (1994) and a PhD in Electrical Engineering from the State University of Campinas (1998). He holds a post-doctorate from INESC-Porto and is currently a full professor of the postgraduate course in Electrical and Computer Engineering at the School of Electrical, Mechanical and Computing Engineering of the Federal University of Goiás.

Cássio Dener Noronha Vinhal, He holds a degree in Electrical Engineering by the Federal University of Uberlândia (1990), Master's Degree in Electrical Engineering by UNICAMP (1994) and PhD in Electrical Engineering by UNICAMP (1998). He is currently a Full Professor at the School of Electrical, Mechanical and Computer Engineering at the Federal University of Goiás. He was postdoctoral fellow at the Institute of Systems and Computer Engineering, University of Porto, Portugal (2006-2007).