Figueira Filho *et al. Journal of Internet Services and Applications* (2015) 6:17
DOI 10.1186/s13174-015-0032-6

Journal of Internet Services and Applications
a SpringerOpen Journal

RESEARCH Open Access

A study on the geographical distribution of Brazil's prestigious software developers

Fernando Figueira Filho[1*], Marcelo Gattermann Perin[2], Christoph Treude[1], Sabrina Marczak[3], Leandro Melo[1], Igor Marques da Silva[1] and Lucas Bibiano dos Santos[1]

Abstract

Brazil is an emerging economy with many IT initiatives from public and private sectors. To evaluate the progress of such initiatives, we study the geographical distribution of software developers in Brazil, in particular which of the Brazilian states succeed the most in attracting and nurturing them. We compare the prestige of developers with socio-economic data and find that (i) prestigious developers tend to be located in the most economically developed regions of Brazil, (ii) they are likely to follow others in the same state they are located in, (iii) they are likely to follow other prestigious developers, and (iv) they tend to follow more people. We discuss the implications of those findings for the development of the Brazilian software industry.

Keywords: Collaborative software development; Software engineering; Social network analysis; Brazil

1 Introduction

Information Technology (IT) has been playing a major role in rapidly growing economies and emerging markets such as the BRIC countries (Brazil, Russia, India, and China), Mexico, Malaysia, Indonesia, and others [32]. The development of information and communication technologies has long been referred to as a "strategic tool" and a pre-requisite for economic growth and social development, especially in developing nations [5].

In Brazil, the investment of resources into fostering the development of IT industries and services has been rising. Public funding for research has steadily increased over the past decade from 1 to 1.17 % of the GDP, slightly lower than in Russia and China but the highest among Latin American countries [10]. Recent initiatives from the Brazilian government include mobility programs such as Science without Borders [14], which is sending hundreds of thousands Brazilians to study at prestigious universities abroad, and the Greater TI program (TI Maior [15]), which has significant focus on boosting the domestic IT sector.

Brazil's software market grew 26.7 % in 2012, ranking seventh globally and surpassing China [16]. Although growing at fast rates, the Brazilian software industry still lags behind in export revenue and most of its production is consumed in the domestic market. To improve Brazil's global competitiveness, recent policies from the Brazilian government have aimed at fostering innovation with public incentives, which include increasing funds for R&D projects and providing tax breaks for key industrial sectors such as IT, biotechnology, and energy. Despite these efforts, Brazil ranks 64th in the 2014 World Economic Forum's Global Innovation Index, behind Russia, China, and Chile.

We hypothesize that socio-economic characteristics are essential in determining the success of a country's IT industry. Using Brazil as a case study, in this paper we report on our investigation on how socio-economic characteristics of different states are related to the prestige of the developers that reside in those states. In addition, we investigate who these developers interact with, using their *follow*-relationships as data source.

We collected and analyzed social network data from over four thousand active GitHub users who explicitly stated in their profiles where in Brazil they are located. For each of these users, we measured their network prestige based on their *follow*-relationships, i.e. who follows whom on GitHub. The *follow*-relationships among GitHub users indicate how useful they are to others and how valuable their activities and contributions are [2]. We analyzed

*Correspondence: fernando@dimap.ufrn.br
[1] Departamento de Informática e Matemática Aplicada, Universidade Federal do Rio Grande do Norte, Campus Universitário, 59078-970 Natal, RN, Brazil
Full list of author information is available at the end of the article

how developers' prestige correlates with Brazil's socio-economic and demographic data per state, including GDP, percentage of urban population, and number of educational institutions, as well as who these developers interact with through their *follow*-relationships. Our findings show that (i) prestigious developers tend to be located in the most economically developed regions of Brazil, (ii) they are likely to follow others in the same state they are located in, (iii) they are likely to follow other prestigious developers, and (iv) they tend to follow more people.

This paper is organized as follows. Section 2 introduces the concept of prestige according to social network literature, GitHub, demographics and statistics about Brazil and its IT industry, and our research hypotheses. Section 3 presents how we collected data to test our hypotheses. Section 4 reports our findings and Section 5 discusses them. Section 6 presents the limitations of our work and Section 7 discusses related work. Section 8 ends the paper with our final considerations.

2 Background

2.1 Network prestige

In this paper, we are interested in measuring the prestige of GitHub users located in Brazil. In social network analysis, *prestige* can be measured based on directional relations among actors. We measured prestige using a graph of *follow*-relationships, in which there is an arc linking user A to B if A *follows* B. On GitHub, this implies that user A receives notifications from B's development activities, which means that there is interest from A in assessing B's contributions.

There are different network measures that can be computed to quantify the prestige of an actor in a social network. The simplest actor-level measure of prestige is the *in-degree* of a vertex i [43] in a graph, which is often referred to as *i's popularity*. However, popularity is a very restricted measure of prestige because it takes only direct choices into account. With popularity it does not matter whether choices are received from popular people. The overall structure of the network is disregarded [6].

Another prestige measure is *proximity*. It defines an *influence domain* of actor i as the set of actors from whom i is reachable and considers the distance these actors are from i. It ignores actors who cannot reach i, thus it is defined even if the network is not connected (when some actors are not reachable from other actors) [43].

We used Pajek [17] to calculate the proximity prestige for each vertex in the graph of *follow*-relationships described above. In Pajek, the *proximity prestige* of a vertex is the proportion of all vertices (except itself) in its input domain divided by the mean distance from all vertices in its input domain.

Maximum proximity prestige is achieved if a vertex is directly chosen by all other vertices. This is the case, for example, in a star-network in which all choices are directed to the central vertex. Then, the proportion of vertices in the input domain is 1 and the mean distance from these vertices is 1, so proximity prestige is 1 divided by 1. Vertices without input domain get minimum proximity prestige by definition, which is zero [6].

2.2 GitHub

GitHub is a web-based hosting service that allows developers to host their software project repositories using the Git revision control system. Since its launch in April 2008, GitHub has become one of the most popular source code hosting services with over twenty million projects maintained by over eight million registered developers [18]. It is now the largest code host in the world [9].

In addition to revision control, GitHub acts as a social network site that enables developers to connect and collaborate with each other. Developers can search for software projects that they are interested in, easily *fork* those projects to make their own contributions, and follow the work of others. We are particularly interested in *follow*-relationships, as they represent a deliberate interest from one developer in another's work and denote the prestige of a developer in GitHub's social network.

The site organizes software repositories by software developer or organization, rather than by project, showing a list of each developer's repositories and their activity on GitHub in a news feed. For a developer, this effectively turns their GitHub profile into an easily accessible public portfolio of their open source development activities [36].

A GitHub user profile includes information on their repositories (i.e., projects) and their recent public activities, such as committing code to a repository or opening an issue report, which are usually not visible in other development environments. The profile page also shows several statistics that are often used on social networking sites, such as the number of other developers following a user or the number of projects they are watching. Such transparency is an interesting feature of GitHub and other social coding sites [42].

GitHub is particularly attractive for researchers because it provides access to its internal data stores through an extensive REST API [19], which researchers can use to access a rich collection of unified and versioned process and product data [9].

Before we detail how we accessed data on Brazilian developers using the GitHub API, we introduce basic demographics about Brazil in the next section to frame our research.

2.3 Brazil's demographics

Brazil is the fifth-biggest country in the world in terms of area and population. With more than 200 million inhabitants, it is also the biggest country in South America and covers almost half (47.3 %) of the entire continent. Except for Chile and Ecuador, Brazil shares a border with every other country in South America.

Roughly 90 % of Brazil's inhabitants live in states on the eastern and southern coasts of Brazil, where the population density varies from 20 to 300 residents per square kilometre. The rest of Brazil, i.e., the Amazon and the mountain regions, offers a lot more space with a population density of less than 5 residents per square kilometre in some cases. In contrast, the Federal District of the capital Brasília and the state of Rio de Janeiro have population densities of more than 300 inhabitants per square kilometre.

Brazil is divided into 26 states and a Federal District, which can be divided into five major regions:

North. The North accounts for almost half of the area of Brazil (45 %), but it is the region with the fewest inhabitants. In particular the Northwest is not industrially developed. Instead, the region is home to the Amazon basin, the largest ecosystem on earth. The following states are in the North: Acre (AC), Amapá (AP), Amazonas (AM), Pará (PA), Rondônia (RO), Roraima (RR), and Tocantins (TO).

Northeast. Almost a third of Brazilians live in the Northeast, a region that is culturally very diverse. It is characterized by Portuguese, African, and indigenous influences. The following states are in the Northeast: Alagoas (AL), Bahia (BA), Ceará (CE), Maranhão (MA), Paraíba (PB), Pernambuco (PE), Piauí (PI), Rio Grande do Norte (RN), and Sergipe (SE).

Center-West. The Center-West of Brazil owes its importance to its wealth in raw materials. Nevertheless, the region is not particularly well developed. However, intensive efforts, such as the move of the capital to Brasília, are being made to strengthen the region. The following states are in the Center-West: Distrito Federal (DF), Goiás (GO), Mato Grosso (MT), and Mato Grosso do Sul (MS). The capital, Brasília, is located in the DF.

Southeast. The Southeast of Brazil is home to more people than any other South American country. With the metropolitan areas of São Paulo and Rio de Janeiro, this region is the economic engine of the country. The following states are in the Southeast: Espírito Santo (ES), Minas Gerais (MG), Rio de Janeiro (RJ), and São Paulo (SP).

South. The South is the smallest region of Brazil with climatic conditions similar to those of southern Europe. The region shows significant cultural influences from German, Polish, and Italian immigrants. The following states are in the South: Paraná (PR), Santa Catarina (SC), and Rio Grande do Sul (RS).

Brazil's most populous metropolitan areas are São Paulo with about 20 million inhabitants, Rio de Janeiro with about 12.5 million inhabitants, and Belo Horizonte with about 5 million inhabitants, making São Paulo the largest city in the southern hemisphere.

Nowadays, Brazil's economy is the seventh largest in the world in terms of nominal gross domestic product (GDP), and the seventh largest in terms of purchasing power parity. A member of the BRIC countries, Brazil had one of the world's fastest growing major economies until about 2010 with economic reforms that gave the country new international reputation and influence. However, the economy has slowed down to modest growth over the last four years.

2.4 Brazil's IT industry

According to a recent study [1], Brazil ranked 7th in IT investments worldwide and 1st in Latin America, with an investment of 61.6 billion US dollars in 2013. Of this, 10.7 billion came from the software market and 14.4 billion from the services market.

The domestic market is operated by approximately 11,230 companies, dedicated to the development, production and distribution of software and services. From those companies, about 93 % can be categorized as micro and small enterprises. Finance, Services and Telecom accounted for almost 51 % of the user market, followed by Industry, Government and Commerce.

The study also pointed out the regional concentration of investments in the IT market. The Southeast region of Brazil met the largest volume of funds allocated to the sector in 2013, with 64.6 %. The North of the country was the least invested in the sector, with a percentage of 2.2 %; the Northeast recorded 8.6 %; South and Center-West accounted for 13.4 % and 11.0 % respectively.

2.5 Hypotheses

There are many challenges associated with the Brazilian software industry and its growth. The vast majority of software companies are located in the Southeast and South regions of Brazil. In 2008, these two regions accounted for 84.3 % of all software companies in the country with more than 20 employees [37]. This result emphasizes the well-known inequality across regions in Brazil.

Assuming that the uneven distribution of software n and their employees across Brazilian states is related to the socio-economic situation in each of these states, our first hypothesis tests whether developers' prestige is associated with the development level of the state they are located in:

H1. Developers' prestige is associated with the development level of the state they are located in.

Our following hypotheses explore the *follow*-relationships between Brazilian developers on GitHub in more detail. Previous work on *follow*-relationships found that developers tended to connect with people with similar levels of performance and experience [35]. In fact, the presence of *homophily*, i.e. the tendency of individuals to associate and bond with similar others, has been discovered in many other network studies in sociology (see McPherson et al. [31] for a review).

For our paper, we examined the homophily of *follow*-relationships by focusing on two different attributes: *geographic location* and *programming language choice*. In particular, our second hypothesis investigates whether developers tend to follow other developers located in the same state:

H2. Developers tend to follow developers located in the same state.

As an alternative explanation of why developers might follow each other, we investigate whether they might use common programming languages as a decision factor:

H3. Developers tend to follow developers who use the same programming languages.

Prestige itself might be a factor for a developer when deciding who to follow. In a study on Twitter's *follow*-relationships, Hopcroft et al. [13] found that the likelihood of two prestigious users creating a reciprocal relationship is nearly 8 times higher than the likelihood of two ordinary users. Our forth hypothesis tests whether, in Brazil, a prestigious developer tends to follow other prestigious developers.

H4. In Brazil, the prestige level of a developer who is following is associated with the prestige level of a developer who is being followed.

In a study with open-source software communities, Shen and Monge found that project leaders tend to follow more people, showing that project leaders are more well-connected than developers in other roles [35]. Our fifth hypotheses does a similar test by focusing on the association between network prestige and the number of people developers follow:

H5. In Brazil, developers' prestige is associated with the number of developers they follow.

3 Method

3.1 Data collection

To obtain data on software developers in Brazil, we accessed the GitHub API using the PyGithub module [20] to search for users who publicly stated their location in their GitHub profile [21]. We ran two different queries, both on November 5, 2014. The first query searched for users who had created their accounts between January 1, 2009 and November 2, 2014, and whose location contained the word *"Brasil"*. The second query was similar to the first one, but searched for a different spelling: *"Brazil"*. Whereas the first spelling is the Portuguese way of spelling Brazil, the latter is the one used in English.

The first query returned 8,815 unique users, and the second query returned 12,064 unique users. Merging these lists resulted in a total of 20,875 users that had either indicated *"Brazil"* or *"Brasil"* in their profile. Of these, 8,634 did not specify their location any further, i.e., their location only consisted of one word indicating the country. Since we are interested in state-specific information of developers in Brazil, we eliminated those from the dataset, leaving 12,241 users.

We collected additional data per state in order to be able to test our hypothesis H1 (Section 2.5). Data collected for testing H1 is in Table 1 at the end of the paper.

HDI. The Human Development Index (HDI) is a composite measure for education, income, and longevity indices, calculated in order to measure social and economic development within countries. It consists of a number between 0 and 1 wherein the development is considered higher when closer to 1. The corresponding data for Brazil's states was taken from Atlas Brazil [22].

GDP. The Gross Domestic Product (GDP) is defined by OECD as an aggregate measure of production that is equal to the sum of the gross values added of all resident institutional units engaged in production (plus any taxes, and minus any subsidies, on products not included in the value of their outputs). The corresponding data for Brazil's states was taken from the Brazilian Institute of Statistical Geography (IBGE) [23] and represents the *GDP per capita* in Reais[1] of each Brazilian state in 2011.

Urbanized population ratio. The urbanized population ratio is the part of a state's population that lives in urban areas as opposed to rural areas. The corresponding data for Brazil's states was taken from IBGE's 2010 census [24].

Population density. Population density measures the number of individuals living in a given area. The corresponding data for each state was taken from IBGE's 2010 census [25].

Number of higher education institutions. The number of higher education institutions per state was taken from the Brazilian Ministry of Education (MEC) [26].

Internet speed. Internet speed refers to the allocated bandwidth available in a given state, measured in data per second. The corresponding data for Brazil's states was taken from CTWatch [27].

To test our third hypothesis, we used the Github API to get the number of lines of code by programming language for each repository owned by a given user. We summed up

Table 1 Socio-economic data on all Brazilian states

State	Developers	HDI	GDP (in 1000)	Urban %	Pop. density	Higher edu. inst.	Internet speed (in Gbps)
SP	1424	0.783	1,349,465	95.9	166.2	446	10
RJ	526	0.761	462,376	96.7	365.2	110	10
RS	337	0.746	263,633	85.1	38.0	81	2.5
MG	295	0.731	386,156	85.3	33.4	243	10
SC	271	0.774	169,050	84.0	65.3	74	2.5
PR	256	0.749	239,366	85.3	52.4	153	2.5
DF	133	0.824	164,482	96.6	444.8	57	10
PE	131	0.673	104,394	80.2	89.6	73	2.5
CE	119	0.682	87,982	75.1	56.8	38	2.5
BA	89	0.660	159,869	72.1	24.8	93	2.5
PB	74	0.658	35,444	75.4	66.7	22	0.034
RN	59	0.684	36,103	77.8	60.0	17	0.034
GO	54	0.735	111,269	90.3	17.7	53	0.034
ES	34	0.740	97,693	83.4	76.3	69	0.034
AM	32	0.674	64,555	79.1	2.2	18	<0.034
AL	30	0.631	28,540	73.6	112.3	19	0.034
MS	25	0.729	49,242	85.6	6.9	33	0.034
PA	22	0.646	88,371	68.5	6.1	26	0.034
TO	20	0.699	18,059	78.8	5.0	16	<0.034
MA	17	0.639	52,187	63.1	19.8	19	0.034
SE	16	0.665	26,199	73.5	94.4	11	0.034
PI	16	0.646	24,607	65.8	12.4	27	0.034
MT	15	0.725	71,418	81.8	3.4	47	0.034
RO	14	0.690	27,839	73.6	6.6	21	<0.034
RR	4	0.707	6,951	76.6	2.0	8	<0.034
AC	2	0.663	8,794	72.6	4.5	7	<0.034
AP	1	0.708	8,968	89.8	4.7	11	<0.034

the number of lines of code by programming language and then assigned the language associated with the highest number of lines of code to that user.

3.2 Data preparation

To ensure that only active users were included in our dataset, we further filtered out GitHub users that had not made any contribution to a public repository within the last three months, i.e., after August 2, 2014. While we may ignore GitHub users that only contribute to private repositories this way, the decision was made in order to avoid noise in our dataset from individuals with a GitHub account that do not contribute to projects. We used GitHub's definition of a contribution [28] in this step: GitHub considers it a contribution when a user pushes to a repository (*PushEvent*), when a user makes a pull request (*PullRequestEvent*), or when a user creates an issue (*CreateIssueEvent*). Of the 12,241 users left after the previous step, 7,977 had not made any contribution to a public repository on GitHub within the last three months, leaving us with 4,264 active users.

Next, we tried to associate state information with each user, i.e., we tried to find out which of the 26 states (or the Federal District) the user indicated in their profile information. Since the location information on GitHub is free-form text and GitHub does not validate this information in any way, some parsing was required in order to semi-automatically attach state information to each user:

1. We normalized location strings by replacing accented letters with their non-accented equivalents. For example, *"ç"* was replaced by *"c"*, and *"á"* was replaced by *"a"*.
2. Strings were transformed to uppercase to make sure that different case did not affect our analysis.
3. We replaced special characters (*"("*, *")"*, *"-"*, *"<"*, *">"*, *"|"*, *"'"*, *"/"*, *":"*, *";"*) with a space (*" "*).

4. We added one space, i.e. *" "*, to the beginning and end of each location string. This was done to make it easier to distinguish words in situations where one word is a substring of another. For example, the normalized string *"PARA"* (from the northern Brazilian state of Pará) is a substring of the normalized string *"PARANA"* (from the southern Brazilian state of Paraná).
5. We normalized the country name to *"BRAZIL"*.
6. We replaced multiple consecutive spaces with one single space.

As an example, our pre-processing steps would transform *"Franca/SP - Brasil"* into *"FRANCA SP BRAZIL"* and *"João Pessoa, PB, Brasil"* into *"JOAO PESSOA PB BRAZIL"*.

In the last step, we tried to match a Brazilian state to each of the 4,264 active users left after the previous step. To do so, we attempted to match the states using both the abbreviated (e.g., *"RN"*) and unabbreviated names (e.g., *"RIO GRANDE DO NORTE"*) of each Brazilian state. In addition, we also tried to match the capital cities of states in case users had included the capital city instead of the state name. For example, our method would have matched *"Recife, Brazil"* to the state Pernambuco. For 248 users, we were unable to assign a state, leaving us with a total of 4,016 active users in Brazil for which we were able to obtain state information. Table 2 summarizes the number of users left after each step of our data preparation method.

Finally, we generated a table containing all *follow*-relationships among these 4,016 active developers in Brazil. The generated table contained tuples with unique user identifiers in the form <*user1,user2*> if User 1 followed User 2.

4 Findings

4.1 Developers in Brazil

Figure 1 illustrates the geographical distribution of active developers in Brazil per state. The state of São Paulo has the highest numbers of developers (1,424), followed by Rio de Janeiro (526) and Rio Grande do Sul (337). We were able to assign at least one developer to each of the 26 states and the Federal District, but the states at the lower end of the spectrum have very few developers: Roraima (4), Acre (2), and Amapá (1). It is worth noting that the latter states are all in the north region of the country.

Table 2 Collecting data on Brazilian GitHub users

Criteria	Number
Contains *"Brasil"*	8,815
Contains *"Brazil"*	+12,064
Contains either, without duplicates	20,875
Contains only country name	−8,634
Contains more than country name	12,241
Inactive users	−7,977
Active users	4,264
No state information	−248
Final number	4,016

Figure 2 shows how the active developers are distributed over the five regions of Brazil. The Southeast region has most of the Brazilian developers, whereas the Center-West and North regions have the least.

Network prestige was not equally distributed across the Brazilian states. São Paulo is the state with the most significant levels of network prestige (0.06672), followed by Rio de Janeiro (0.06213) and Rio Grande do Sul (0.06573). JavaScript is the main programming language among Brazilian active developers (21.0 %), followed by Java (14.3 %) and Ruby (11.9 %).

4.2 Hypotheses tests

To verify the correlations proposed in hypotheses H1, H4, and H5, we first evaluated the normality of our data by applying the Kolmogorov-Smirnov test [44]. The tests were significant for all variables in analysis for these three hypotheses, confirming the non-normality of data. Therefore, following the Bishara and Hittner recommendation [4], we applied Spearman's rho (r_s) to evaluate the association between variables.

H1. Developers' prestige is associated with the development level of the state they are located in.

As presented in Section 3, the development level of each state is reflected in its characteristics of GDP, HDI, percentage of urban population, population density, number of higher education institutions, and speed of the Internet connection.

Table 3 and Fig. 3 show the correlation between prestige level (proximity prestige index) and all other variables related to hypothesis H1. Note that all variables were positive and significantly associated with proximity prestige index by Spearman's rho. However, because the magnitudes of all correlations were not strong, indicating a possible important set of other variables with influence on the prestige in the social network, we conclude that **H1 is supported with a very weak correlation**.

H2. Developers tend to follow developers located in the same state.

We first summarized for each state the number of *follow*-relationships where the followed developer was from the same state as the follower. Then, we divided this number by the number of possible relationships between developers from the same state. We also summarized for each state the amount of *follow*-relationships where the follower developer was from the state and the followed developer was from a different state. Accordingly,

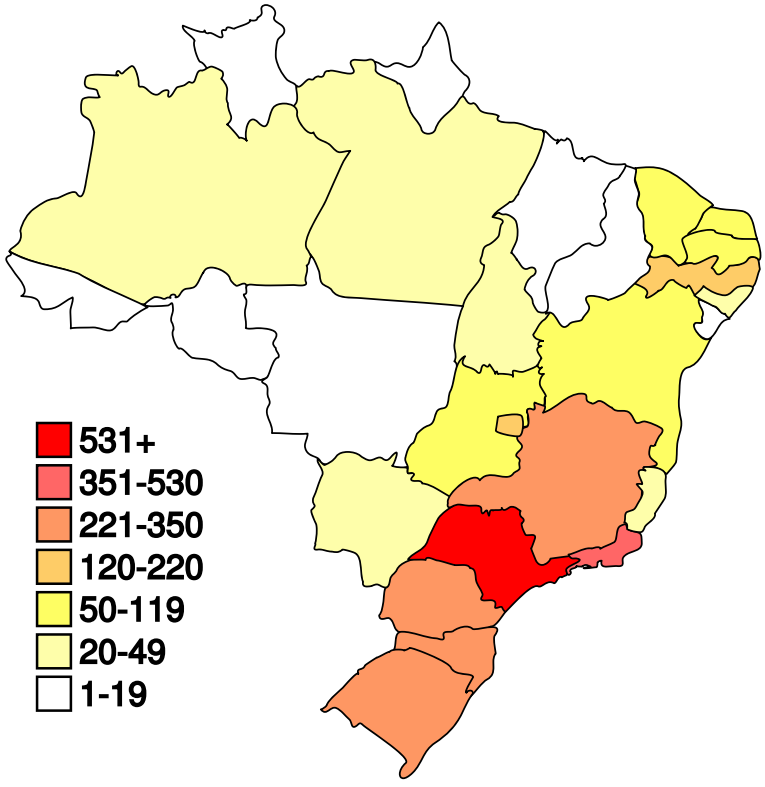

Fig. 1 Developers per state in Brazil

we divided this number by the number of possible relationships between follower developers of the state and followed developers from different states. In other words, we created a coefficient for *follow*-relationships in the same state and *follow*-relationships across states.

To evaluate H2, we applied a pairwise Wilcoxon signed-rank test. Result showed a significant difference between the two generated coefficients ($t = -4.211$; $p < 0.001$). All coefficients of followers in the same state were greater than coefficients of followers of developers from different states, except for three states (Roraima, Paraíba, and Mato Grosso). Therefore, we conclude that H2 is **supported**.

H3. Developers tend to follow developers who use the same programming languages.

To verify H3, we followed the same approach applied in H2. Similarly, we generated a coefficient of relationships for each programming language for *follow*-relationships with developers of the same programming language and *follow*-relationships from one programming language to others. Result pointed to a non-significant difference between the two coefficients ($t = -1.461$; $df = 39$; $p = 0.152$). This result did **not support** H3.

H4. In Brazil, the prestige level of a developer who is following is associated with the prestige level of a developer who is being followed.

For this hypothesis we considered the above mentioned table of *follow*-relationships. We estimated the association between the proximity prestige indices of each developer in existent following relationships. As shown in Fig. 4 (left), although Spearman's rho was positive and significant ($r_s = 0.231$; $p < 0.001$), its magnitude was considered weak [7]. Therefore, we conclude that **H4 is supported with a weak correlation**.

H5. In Brazil, developers' prestige is associated with the number of developers they follow.

To test H5, we related the prestige level of developers with their out degree centrality level. As shown in Fig. 4 (right), result for the correlation test were positive and significant ($r_s = 0.394$; $p < 0.001$), but its magnitude was at most moderate [7]. Hence, **H5 is supported with a moderate correlation**.

5 Discussion

The recent rise of social media use by developers [39] and the effects of leveraging social transparency [40] in virtual communities bring exciting possibilities to software engineering [38]. In particular, GitHub has dramatically improved the level of collaboration and participation among people who build software [2]. Nurturing relationships among software developers is a phenomenon

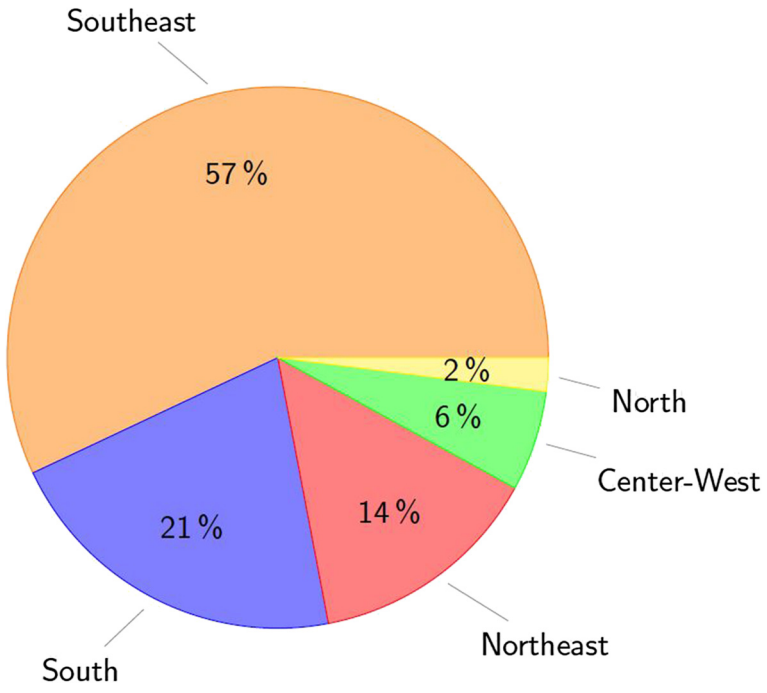

Fig. 2 Developers per region

of increasing interest in Software Engineering research [38] because of its potential for fostering even further innovation in software products and services.

On GitHub, users follow interesting developers, listen to their activities, and find new projects. Social relationships between users are utilized to disseminate projects, attract contributors, and increase projects' popularity [30]. Thus, the prestige of a developer in a social network of *follow*-relationships indicates how useful they are to others and how valuable their activities and contributions are. In this context, prestigious developers act as hubs of information and knowledge flow in software development.

Our findings for hypothesis H1 suggest that prestigious developers in Brazil are more likely to be found in developed states, although the correlations were very weak. The strongest positive correlations were found between developers' prestige and state GDP, population density, and urban percentage, respectively. These are indicators that reflect economic development and industrial capacity. Another state characteristic that correlates with developers' prestige is the number of higher education institutions. Although the correlation does not indicate causation, this finding might suggest that investment in higher education could play an important role for the next generation of software developers in the country.

Table 3 Correlation between prestige level and state characteristics

State characteristic	Spearman's rho
State GDP	.094**
State HDI	.062*
State Urban %	.091**
State Population Density	.094**
State Higher Education Institutions	.082**
State Average Internet Speed	.091**

*indicates significant correlation at the 0.05 level (2-tailed)
**indicates significant correlation at the 0.01 level (2-tailed)

The effects of geographical distance have been studied in several contexts, including software engineering [12, 33, 34]. Our finding for hypotheses H2 indicates that developers tend to follow others located in the same state they are located in. Bell et al. [3] found that institutional-level ties are valuable in knowledge transmission only when such ties are geographically proximate. Our finding shows that this might be true at the individual-level as well, i.e., individual developers have a greater interest in others that are geographically proximate. This indicates that developers' *follow*-relationships might develop primarily as a result of collocated interactions in local communities, e.g. among co-workers, classmates, and colleagues.

Our finding for hypothesis H3 demonstrates that *follow*-relationships among developers might have causes other than the similarity of their interests in programming languages.

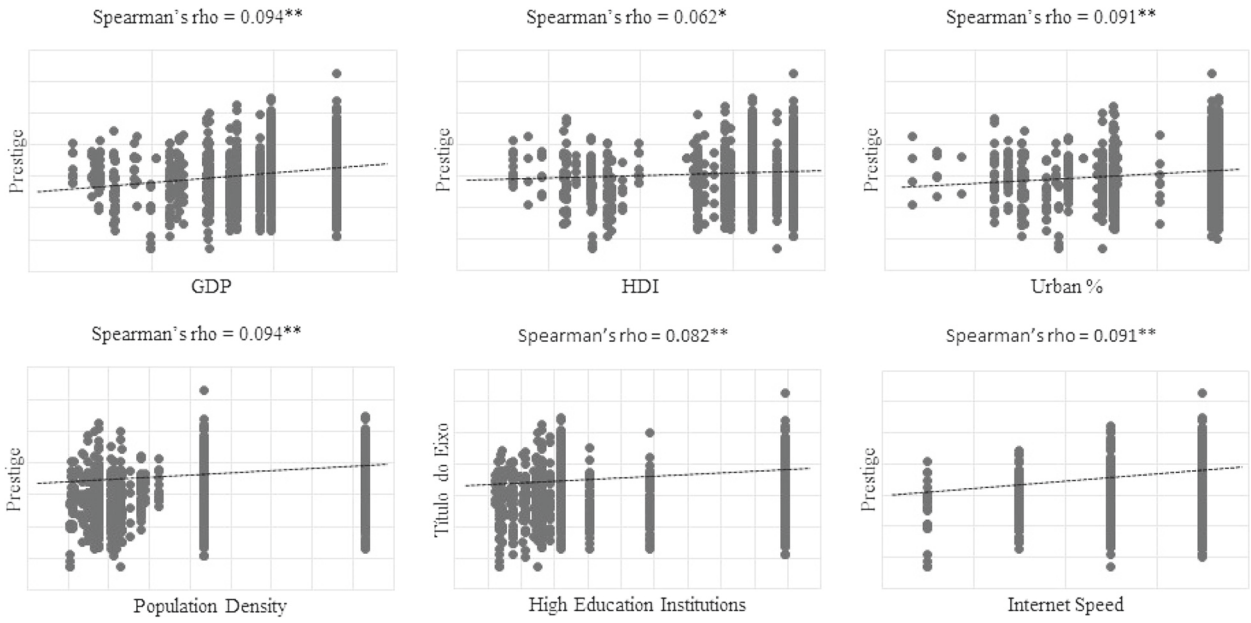

Fig. 3 Correlation between prestige level and state characteristics

Our finding for hypothesis H4 indicates that, in Brazil, prestigious developers tend to follow other prestigious developers, while hypothesis H5 indicates that prestigious developers are likely to follow a larger number of people. These findings suggest that prestigious developers make extensive use of the social networking features available on social coding sites. Most importantly, it shows a preferential attachment among highly prestigious software developers. Considering our findings for H2 and H4, we conclude that this preferential attachment is positively influenced by the level of prestige of software developers in their networks, but also by the geographical distance among them.

Understanding the demographics of a software developer population can inform a variety of initiatives for nurturing the IT industry in Brazil. First, our findings suggest that government efforts should be targeted at promoting high-tech industries in the least developed regions of the country, i.e. Northeast, Center-West, and North. However, the success of those efforts depends on expanding educational infrastructure and promoting economic growth in those regions. Second, in order to boost the IT sector in Brazil, the Brazilian government would have to consider the influence of geographical distance on the preferential attachment displayed by prestigious developers. This means investing into the decentralization of key industrial sectors by creating conditions for attracting highly skilled labor to the least explored regions of Brazil.

6 Limitations

As with any research method, there are limitations related to our choice of research methods. These can be divided into threats to external, construct, and internal validity.

6.1 External validity

External validity reflects the extent to which the results of a study can be generalized to other settings. We cannot

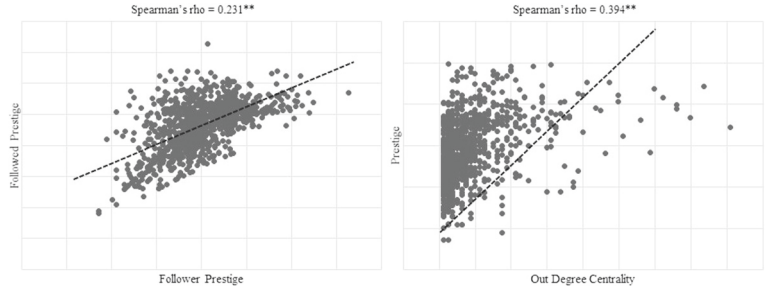

Fig. 4 Correlations associated with the hypotheses H4 (left) and H5 (right)

claim that the results of our study are generalizable beyond Brazil's borders, however, many emerging economies (in particular the BRIC-countries) face challenges similar to those of Brazil, and it is plausible that many of our conclusions apply to those countries as well.

In addition, our results are limited to the population of software developers found on GitHub, and cannot readily be generalized to every software project, open source or not. We chose to study GitHub's population of users because their geographical location information was available on their profiles which enabled us to automatically mine it. However, GitHub is now the largest code host in the world [9], and we are not aware of any other publicly available data source for our study.

6.2 Construct validity

Construct validity reflects the extent to which our study actually measures what we claim to measure. For this study, the issues related to construct validity are mostly connected to the way in which users on GitHub specify their location. A GitHub user can provide any text as location, and the texts we collected may not represent a valid location in the world. In fact, we were not able to find state information for 248 of the 4,264 active GitHub users located in Brazil that had included additional information about their location.

In addition, we investigated the participation of developers based on the information they provided on their GitHub profile page. This information may not be frequently updated by the user. This means that the user may be working in other locations while contributing to projects on GitHub. Therefore, in our study, locations are merely indicators of where developers lived when they signed up for GitHub, and may not correspond to the their actual locations. However, a manual approach would have been infeasible for the amount of data needed to generate statistically significant conclusions, and we believe that the location information entered by GitHub users is at least a good approximation of where Brazil's developers are actually located.

Another issue is the construct validity of the network prestige measure. For our study, this measure considered only the *follow*-relationships among developers. This measure may not correspond to the actual prestige of a developer in his social world or community. Further work is needed to assess whether *follow*-relationships are good predictors of one's prestige among peers in development communities such as GitHub.

Finally, for the investigation of H3, we assigned the most used programming language for each user in our dataset. A developer on GitHub can use a variety of programming languages, and we ignored those languages except for the most used one. However, we believe our procedure offered a good approximation of developers' preferences regarding programming languages.

6.3 Internal validity

Internal validity reflects the extent to which a causal conclusion based on the study and its methods is possible. Based on our statistical tests, we cannot infer causal relationships between socio-economic data and developers' prestige, for example. However, because of our use of well-recognized statistical techniques, we are confident that the correlations we found hold for the data we collected and analyzed. To investigate whether the socio-economic situation caused developers to become more prestigious or vice versa, will be a goal of future work.

Also, we must point out that the correlation level estimated regarding H1, H4 and H5 should be considered as very weak, weak and moderate respectively. Hence, although the correlations were both positive and significant, those results must be considered with some care.

7 Related work

Previous work has examined the structure of social relationships in the GitHub community. Thung et al. [42] extracted information about 100,000 projects from GitHub and identified their most influential developers. Jiang et al. [30] examined *follow*-relationships among GitHub users. They discovered that social relationships are not reciprocal and that social links play a notable role in project dissemination.

A few articles have focused on the geographical distribution of GitHub users. Takhteyev and Hilts [41] analyzed the geographical distribution of GitHub developers worldwide. They found that developers are highly clustered and concentrated primarily in North America, and Western and Northern Europe. Heller et al. [11] applied visualization techniques for analyzing the effect of geographic distance on developer relationships and social connectivity.

Gonzalez-Barahona et al. [8] estimated the geographical origin of more than one million individuals by analyzing SourceForge's [29] mailing lists archives from several large open source projects, such as GNOME and FreeBSD. Their results show that most developers are in North America and Europe.

To the best of our knowledge, our work is the first to study the geographical distribution of Brazil's prestigious software developers.

8 Final considerations

Many rapidly growing economies and emerging markets, such as the BRIC countries (Brazil, Russia, India, and China), use information technology (IT) as a key driver for progress, development, and success. To be able to

compete in a global market, these countries are innovating and implementing strategic initiatives to attract and nurture IT professionals.

To shed light on the challenges and opportunities faced by decision makers when trying to develop a country's IT potential, we have studied the geographical distribution of Brazil's software developers by using GitHub data and correlating it with socio-economic information about different regions and states within Brazil. Our findings show that prestigious developers—measured in terms of their proximity prestige in the social network of *follow*-relationships on GitHub—tend to be located in states that are more economically developed. In the case of Brazil, these are the states in the southern part of the country with high GDPs, a substantial number of higher education institutions, and fast Internet. In addition, we find that Brazil's prestigious developers are likely to follow others that are located in the same state, that they are likely to follow other prestigious developers, and that they tend to follow more developers in general.

In future work, we plan to complement the results presented here by analyzing the movements of IT professionals over time as their socio-economic environment changes. We also plan to look beyond Brazil's borders to see how Brazil's software developers participate in the global software development community and to compare our findings to those of other countries. Finally, we plan to investigate *follow*-relationships in order to identify the factors influencing developers' decisions when choosing who to follow.

Endnote

[1] The real (plural *Reais*) is the present-day currency of Brazil.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
FFF lead and coordinated the writing process, and elaborated the hypotheses. MGP designed the study and performed statistical analysis. CT helped to draft the manuscript and to improve it. SM drafted the manuscript and participated in the study design. LM collected and pre-processed social network data. IMS collected socio-economic data. LBS helped to prepare the figures we included in this paper. All authors read and approved the final manuscript.

Acknowledgements
We thank CAPES–Brazil for financially supporting Leandro Melo, and Nancy Songtaweesin for suggesting some socio-economic indicators we used in our research.

Author details
[1] Departamento de Informática e Matemática Aplicada, Universidade Federal do Rio Grande do Norte, Campus Universitário, 59078-970 Natal, RN, Brazil. [2] Programa de Pós-Graduação em Administraçáo, Faculdade de Administração, Contabilidade e Economia, Pontifícia Universidade Católica do Rio Grande do Sul, Av. Ipiranga, 6681, 90619-900 Porto Alegre, RS, Brazil. [3] Programa de Pós-Graduaćão em Ciência da Computação, Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Av. Ipiranga, 6681, 90619-900 Porto Alegre, RS, Brazil.

Received: 25 November 2014 Accepted: 21 July 2015
Published online: 11 August 2015

References

1. ABES Software (2014) Brazilian software market: scenario and trends. http://www.abessoftware.com.br/dados-do-setor/dados-2014
2. Begel A, Bosch J, Storey MA (2013) Social networking meets software development: perspectives from GitHub, MSDN, Stack Exchange, and TopCoder. IEEE Softw 30(1):52–66
3. Bell GG, Zaheer A (2007) Geography, networks, and knowledge flow. Organ Sci 18(6):955–972
4. Bishara AJ, Hittner JB (2012) Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. Psychol Methods 17(3):399
5. Castells M (1999) Information technology, globalization and social development. Tech. rep., United Nations Research Institute for Social Development, September. UNRISD Discussion Paper No 114. Geneva, Switzerland
6. De Nooy W, Mrvar A, Batagelj V (2011) Exploratory social network analysis with Pajek, Vol. 27. Cambridge University Press, New York, NY, USA
7. Fallik F, Brown BL (1983) Statistics for Behavioral Sciences. The Dorsey Press, Homewood, Illinois
8. Gonzalez-Barahona JM, Robles G, Andradas-Izquierdo R, Ghosh RA (2008) Geographic origin of libre software developers. Inf Econ Policy 20(4):356–363. Empirical Issues Open Source Software
9. Gousios G, Spinellis D (2012) Ghtorrent: GitHub's data from a firehose. In: Proceedings of the 9th IEEE working conference on Mining Software Repositories, MSR '12. IEEE Press, Piscataway. pp 12–21
10. Gupta N, Weber C, Peña V, Shipp S, Healey D (2013) Innovation policies of Brazil. Tech. rep., Institute for Defense Analyses, IDA Paper P-5039
11. Heller B, Marschner E, Rosenfeld E, Heer J (2011) Visualizing collaboration and influence in the open-source software community. In: Proceedings of the 8th working conference on Mining Software Repositories, MSR '11. ACM, New York. pp 223–226
12. Herbsleb JD, Mockus A, Finholt TA, Grinter RE (2001) An empirical study of global software development: Distance and speed. In: Proceedings of the 23rd International Conference on Software Engineering, ICSE '01. IEEE Computer Society, Washington, DC. pp 81–90
13. Hopcroft J, Lou T, Tang J (2011) Who will follow you back?: Reciprocal relationship prediction. In: Proceedings of the 20th ACM international Conference on Information and Knowledge Management, CIKM '11. ACM, New York, USA. pp 1137–1146
14. Programa Ciência sem Fronteiras. http://www.cienciasemfronteiras.gov.br
15. TIMaior - Programa Estratégico de Software e Serviços de Tecnologia de Informação. http://timaior.mcti.gov.br
16. Surpassing China, Brazil's IT Industry is a Force to Reckon With. http://pulsosocial.com/en/2013/08/26/surpassing-china-brazils-it-industry-is-a-force-to-reckon-with
17. Program Package Pajek/PajekXXL. http://pajek.imfm.si/doku.php
18. GitHub Press. https://github.com/about/press
19. GitHub Developer. https://developer.github.com/
20. PyGithub documentation. http://jacquev6.net/PyGithub/v2/index.html
21. GitHub - Searching Users - User Documentation. http://help.github.com/articles/searching-users/#search-based-on-the-location-where-a-user-resides
22. Consulta - Atlas do Desenvolvimento Humano no Brasil. http://www.atlasbrasil.org.br/2013/consulta/
23. IBGE - Produto Interno Bruto 2011. http://www.ibge.gov.br/home/presidencia/noticias/images/2522_3643_173712_106392.gif
24. IBGE - Censo Demográfico 2010 - Distribuição percentual da população. http://www.censo2010.ibge.gov.br/sinopse/index.php?dados=9&uf=00
25. IBGE - Censo Demográfico 2010 - Densidade demográfica. http://www.censo2010.ibge.gov.br/sinopse/index.php?dados=10&uf=00
26. e-MEC - Instituições de Educação Superior e Cursos Cadastrados. http://emec.mec.gov.br/

27. CTWatch Quarterly - Cyberinfrastructure for Multidisciplinary Science in Brazil. http://www.ctwatch.org/quarterly/articles/2006/02/cyberinfrastructure-for-multidisciplinaryscience-in-brazil/3/
28. GitHub - Introducing contributions. https://github.com/blog/1360-introducing-contributions
29. SourceForge. http://sourceforge.net
30. Jiang J, Zhang L, Li L (2013) Understanding project dissemination on a social coding site. In: 20th working conference on reverse engineering. IEEE, Piscataway, New Jersey, USA. pp 132–141
31. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Annu Rev Sociol 27:415–444
32. Murugesan S (2011) The rise of emerging markets: opportunities and challenges for IT. IT Prof 13(1):6–8
33. Nguyen T, Wolf T, Damian D (2008) Global software development and delay: does distance still matter? In: IEEE international conference on global software engineering. IEEE, Piscataway, New Jersey, USA. pp 45–54
34. GM Olson, JS Olson (2000) Distance matters. Hum-Comput Interact 15(2):139–178
35. Shen C, Monge P (2011) Who connects with whom? A social network analysis of an online open source software community. First Monday 16(6). http://firstmonday.org/ojs/index.php/fm/article/view/3551/2991
36. Singer L, Figueira Filho F, Cleary B, Treude C, Storey MA, Schneider K (2013) Mutual assessment in the social programmer ecosystem: an empirical investigation of developer profile aggregators. ACM, New York
37. Softex (2012) Observatório softex. Tech. rep., Softex
38. Storey MA, Singer L, Cleary B, Figueira Filho F, Zagalsky A (2014) The (r) evolution of social media in software engineering. In: Proceedings of Future of Software Engineering, FOSE 2014, New York. pp 100–116
39. Storey MA, Treude C, van Deursen A, Cheng LT (2010) The impact of social media on software engineering practices and tools. In: Proceedings of the FSE/SDP workshop on Future of Software Engineering Research, FoSER '10. ACM, New York. pp 359–364
40. Stuart HC, Dabbish L, Kiesler S, Kinnaird P, Kang R (2012) Social transparency in networked information exchange: a theoretical framework. In: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, CSCW '12. ACM, New York. pp 451–460
41. Takhteyev Y, Hilts A (2010). Investigating the geography of open source software through GitHub. http://takhteyev.org/papers/Takhteyev-Hilts-2010.pdf. Accessed: November 19,2014
42. Thung F, Bissyande TF, Lo D, Jiang L (2013) Network structure of social coding in GitHub. In: Proceedings of the 17th European Conference on Software Maintenance and Reengineering, CSMR '13. IEEE Computer Society, Washington, DC. pp 323–326
43. Wasserman S, Faust K (1994) Social network analysis: methods and applications, Vol. 8. Cambridge University Press, New York, NY, USA
44. Yap B, Sim C (2011) Comparisons of various types of normality tests. J Stat Comput Simul 81(12):2141–2155