Diversity on GitHub: a Data Visualization Study

•••

Jun, 2018

Karina Kohl Silveira

School of Technology,
PUCRS
Porto Alegre, Brazil
karina.kohl@acad.pucrs.br

Cesar Liedke

School of Technology,
PUCRS
Porto Alegre, Brazil
cesar.liedke@acad.pucrs.br

Motivation:

GitHub is a rich field to get great datasets.

Vasilescu et al.[1] presented for the first time a large data set of social diversity attributes of programmers in GitHub teams.

The Diversity data set contains data from 23,493 projects and a total of 93,056 rows of content, and outcomes of their teams of contributors from 2009 to 2013.

The data is available in CSV format at: https://github.com/bvasiles/diversity

Related Work

"Open Source Resume (OSR): A Visualization Tool for Presenting OSS Biographies of Developers" [2]

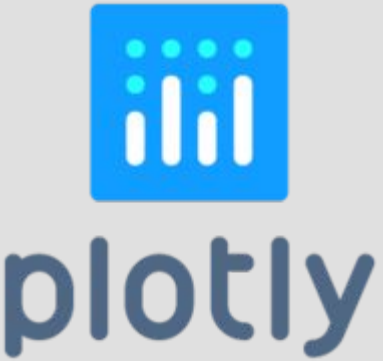
Ways to visualize GitHub developers data to support the process of recruit appropriate developers for software projects once traditional resume only shows experiences claimed by developers, and very few evidence or information regarding their actual development activities can be obtained.

"DemographicVis: Analyzing demographic information based on user generated content" [3]

Reddit.com data to model process that connects the demographic groups with features that best describe the distinguishing characteristics of each group.

Visualization Techniques and Data Interaction


jupyter


plotly

Plotly for Python Jupyter

Plotly for Python is a graphing library that supports to make interactive graphs

Jupyter is an open-source web application that allows to create and share documents that contain live code, equations, visualizations and narrative text.

Some uses for Jupyter are: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, etc.

Case Studies for Diversity

Diversity

Accordingly to Vasilescu et al. [1], diversity in a team can be identified when team members are different from each other with respect to some attribute e.g., gender, tenure, nationality, etc.

We can use aggregate measures of diversity, such as the Blau index to capture how diverse groups are, i.e., the higher the measures, the more diverse team members are with respect to a given attribute.

Blau Index

A **diversity index** is a quantitative measure that reflects how many different types (such as species) there are in a dataset (a community), and simultaneously takes into account how evenly the basic entities (such as individuals) are distributed among those types.

Gini–Simpson index

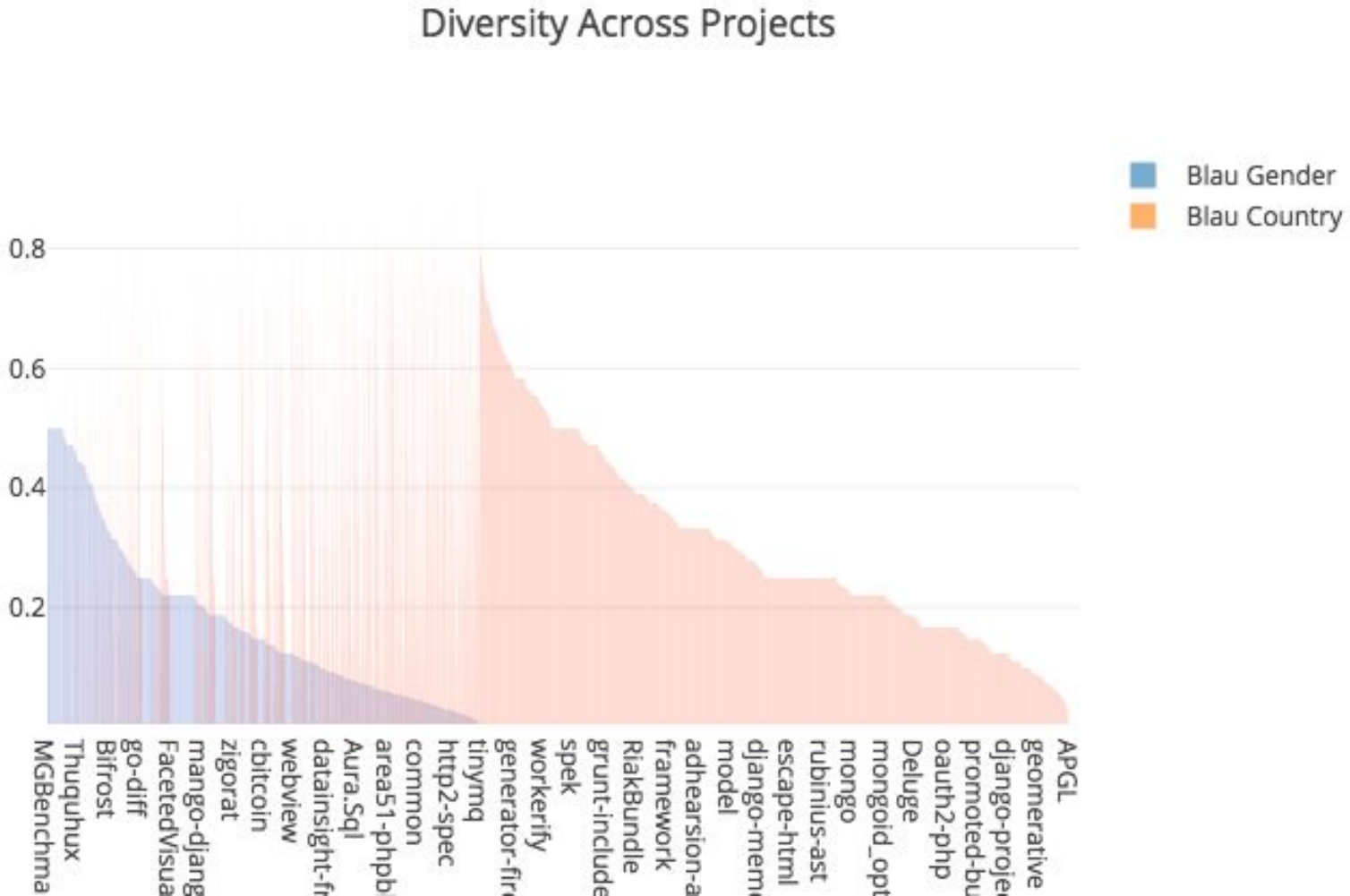
The original Simpson index λ equals the probability that two entities taken at random from the dataset of interest (with replacement) represent the same type.

The **Gibbs–Martin** index of sociology, psychology and management studies,which is also known as the **Blau index**, is the same measure as the Gini–Simpson index.

Diversity Across Projects - Gender and Country

Case Study 1

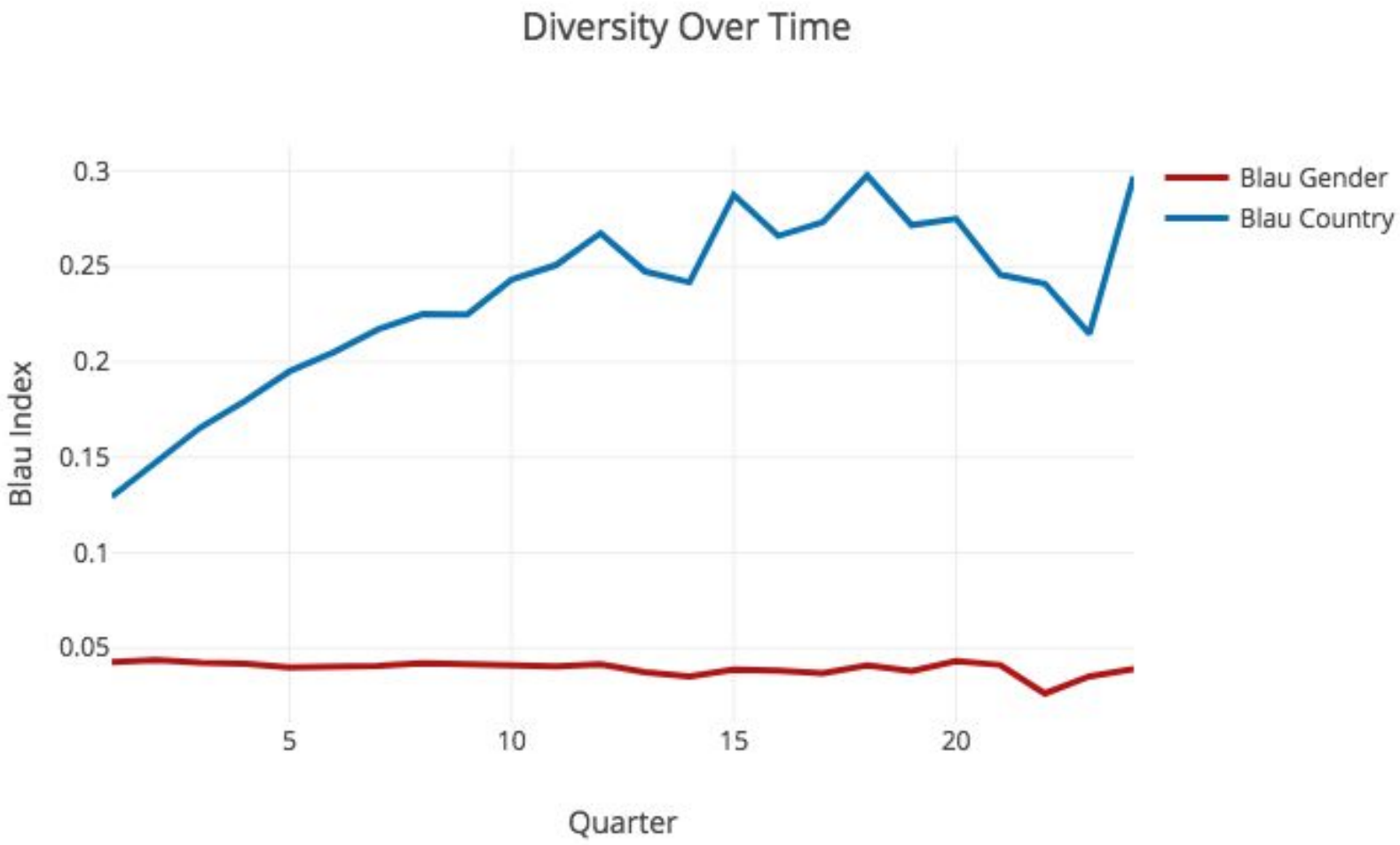
- A team consisting only of same gender members cannot be consider gender diverse;
- However, a team can reach its maximal gender diversity by having equally many female and male members (assuming a simplified, binary gender), regardless of team size;
- Using the Blau index:
  - gender uniformity is encoded as 0
  - maximal gender diversity is encoded as 0.5 (since the data set considers only two possible values, male and female).


Diversity Across Projects
Blau Gender
Blau Country
0.8
0.6
0.4
0.2
APGL
geomerative
django-projec
promoted-bu
oauth2-php
Deluge
mongoid_opt
mongo
rubinius-ast
escape-html
django-meme
model
adhearsion-a
framework
RiakBundle
grunt-include
spek
workerify
generator-fire
tinymq
http2-spec
common
area51-phpb
Aura.Sql
datainsight-fr
webview
cbitcoin
zigorat
mango-djang
FacetedVisua
go-diff
Bifrost
Thuquhux
MGBenchma

Diversity Over Time

Case Study 2

- Plot of evolutionary setting as the evolution of the Blau index of gender diversity in a project during six years (24 quarters) of history;
- One can observe how the project's team of contributors was gender balanced in the first two years (Blau index values above 0.4);
- However, with time, as the contributor team grew (team size shown dotted for comparison), it also became male-dominated;


Diversity Over Time
Blau Gender
Blau Country
Blau Index
0.3
0.25
0.2
0.15
0.1
0.05
5
10
15
20
Quarter

Discussion

Expressiveness Effectiveness

Mackinlay, 1986

Expressiveness

- Identify the graphic presentations that express the desired information
- Prioritize important information and avoid false inferences
- Consistent visual mappings

Effectiveness

- Determine graphic presentations that allow us a faster and precise data interpretation.
- Minimizes the cognitive effort
- Highlights the main elements.

The proposed visualizations allow us:

faster understanding the information under the Diversity data set used

For the first Case Study

- We can easily infer which projects were the most gender diverse through the entire period of time considered.
- With the interaction techniques provided, we can choose specific samples from the less to the most diverse projects in terms of gender and country diversity.

For the second Case Study

- We can select a very specific period of time and check, in that period, how gender and country diversity changes over the time independent from projects.

Conclusion & Future Work

Graphic presentation of the data help on the understanding of the information and it complements the textual evaluation of the information contained in the data.

Once the data set was partially used, it is possible to provide other correlations and visualizations to extract additional information.

* [.]

References:

[1] B. Vasilescu, A. Serebrenik, and V. Filkov. 2015. A Data Set for Social Diversity Studies of GitHub Teams. In 2015 IEEE/ACM 12th Working Conference on Mining Software Repositories datasets.
[2] T. Jaruchotrattanasakul, X. Yang, E. Makihara, K. Fujiwara and H. Iida, "Open Source Resume (OSR): A Visualization Tool for Presenting OSS Biographies of Developers," 2016 7th International Workshop on Empirical Software Engineering in Practice.
[3]W. Dou, I. Cho, O. ElTayeby, J. Choo, X. Wang and W. Ribarsky, "DemographicVis: Analyzing demographic information based on user generated content," 2015 IEEE Conference on Visual Analytics Science and Technology.
[4] Gibbs, Jack P.; William T. Martin (1962). "Urbanization, technology and the division of labor". American Sociological Review. 27: 667–677.