Google Summer of Code

GSoC 2021 proposal

NetworkX - Community Detection Algorithms

Papageorgiou Dimitrios - dkpapageo@ece.auth.gr

"I often compare open source to science. To where science took this whole notion of developing ideas in the open and improving on other people's ideas and making it into what science is today and the incredible advances that we have had. And I compare that to witchcraft and alchemy, where openness was something you didn't do."

– Linus Torvalds, *Creator of the Linux kernel*

1. Student Details

1.1 Personal Info

Full Name: Papageorgiou Dimitrios (Dimitris)
Birth Date: 15/06/1997
Email: dkpapageo@ece.auth.gr
Github: https://github.com/z3y50n
Stack Overflow: https://stackoverflow.com/users/11263279/z3y50n
Native Language: Greek
Location: Thessaloniki, Greece
Timezone: UTC +2 EET Eastern European Time

1.2 Studies Info

Degree: Electrical and Computer Engineering
University: Aristotle University of Thessaloniki
Year: 6th
Current Grade: 7.51 / 10

1.3 Skills

In order to provide a more clear view of my skills, I split the technologies I am familiar with into two groups:

- **Working experience**: I have implemented a project using the specific technology that can be found in a Github repository.
- **Theoretical knowledge**: I am just familiar with the technology, having used it in minor academic exercises.

Obviously in my years of programming not only academically, but also professionally and in my personal time I have used many more libraries and it's impossible to list all of them here. After all, the most important thing a programmer can have as a skill is to be able to search quickly through the docs and figure out how a specific technology works.

Category	Working Experience	Theoretical Knowledge
Programming Languages	Python, C, C++, Javascript, CUDA, GNU Bash	Java, Javascript, Scheme
Web Development	HTML, CSS, NodeJS, React	PHP, JQuery, Ajax
Frameworks - Libraries	Flask, BeautifulSoup, Selenium, Kivy, Numpy, Matplotlib, NetworkX, ROS, MPI, openMP, Cilk, pthreads, BLAS, Express	Scipy, Pandas, Django
Databases	MySQL, MongoDB, DynamoDB	postgreSQL
Typesetting	JSON, Markdown, XML	
Version Control	Git	Maven
Buildsystem - Generators	CMake, Makefile, Docker	
IDEs	Vim, VS Code, Sublime, Eclipse, Spyder, R-Studio	
Scientific Programming	Matlab, Octave, R, K-means, Neural Networks	Data Analysis, Dynamic Programming, Greedy Algorithms, Probabilistic Algorithms, Graph Theory

1.4 About Me

Currently I am in my last year of studying Electrical & Computer Engineering at Aristotle University of Thessaloniki. My main activities as of now are my diploma thesis (create a health focused SmartMirror) and working as a part-time software engineer at QuantamixSolutions where I've been employed there since last summer. In my part-time job I managed to get a lot of experience regarding the software development process, as well as writing code in a real production environment. Aside from my programming interests I am a pretty social person, I like to exercise, read books, hike in nature and explore new things and technologies.

1.5 Commitment to GSoC - Project

Before the start of GSoC (main Coding Period), I will have finished the majority of my obligations, which are mainly school projects, so that I am absolutely sure, I will respect

my timeline and I will succeed the goals of the project. My university exams won't be a problem since I plan on being free from required subjects until the program starts. Also currently I am working on my diploma thesis which won't have an impact on the timeline because I am confident that I will have most of the work (if not all of it) complete by the time GSoC starts. From the time I apply to Networkx, I will try to contribute as much as I can in order to stay in touch with the library, regardless of my acceptance. This work can be either to implement some new algorithm or work on some bugs from the open issues list that will give me an even better understanding of the library.

2. Project Details

2.1 Abstract

NetworkX is a Python library for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. Basically it includes a lot of graph representation and functionalities that can help study those networks.

One very interesting field of networks is community detection, that is to group different nodes together based on similar characteristics or the number of edges between inner and outer nodes. It can help us reveal the hidden relations among the nodes in the network. Networkx has already implemented a few algorithms that perform community detection, but currently there is no implementation for the Louvain Community Detection algorithm even though there have been previous attempts to integrate it inside the library (they can be seen here and here. Louvain is a very efficient greedy algorithm for community detection that seems to run in O(nlogn) according to this.

2.2 Technical Details

The main implementation of Louvain Community Detection algorithm is described in this paper which was written by the authors of the algorithm. Also the algorithm itself is written in C++ by the same authors and can be found here. There has been an implementation in python as well, that is a library located here and up to this point anyone who wants to use Louvain inside python needs this library. The main purpose of the project is to write Louvain inside the networkx allowing the users to use it directly through the library.

Louvain itself is based on modularity optimization and therefore a lot of helper functions are needed to be implemented to help with the algorithm. Also since it is a heuristic approach the user should be able to select how deep into the community layers he wants to go (obviously the bottom of the tree is the best partition according to its clusters' modularity). Finally, one major decision that needs to be taken is regarding the output of the algorithm, in order to be consistent with other community detection algorithms and also to be utilized easily and efficiently by other parts of networkx. There have been a lot of discussions about the output format with major opinions preferring to return a list of sets of nodes (most if not all of the existing algorithms in NetworkX return the same output structure), so extra attention should be given to handle the output correctly. The current implementation in the

python-louvain package returns a dictionary with nodes as keys and community number as their value.

3. Project Timeline

A brief timeline of the project follows. It should be noted that during the implementation of the project, some priorities might change after the interaction with the mentors. But, of course, this has been taken into consideration and in the timeline, some days are dedicated in the case something changes or stalls.

The project can be divided into two sections, one for the actual implementation of the algorithm and its helper functions, and one for implementing a wrapper to control the output in order to be consistent with other implemented algorithms and the needs of the community.

Special attention is given to documentation writing which will be taking place immediately after each commit. Providing documentation is of equal importance as is development, in order for the end-user to be able to efficiently use the new tools. That is one of the organization's highest priorities and it will be handled accordingly. Therefore, during each phase there is also time allocated for documentation and proper testing.

TIMELINE

-- April 13 - May 17 (Period until accepted student proposal announced)

During this period I will make sure that I finish my other obligations and stay in touch with networkx, either by working on some issues, or implementing some algorithm. Also it could be beneficial to strengthen my discrete mathematics background during this time. Furthermore, I should also study current implementations of the Louvain algorithm in order to get an even better understanding of the final code architecture that I will implement in the summer.

-- May 17 - June 7 (Community Bonding Period)

During my proposal, I have contacted the mentors of the project many times and they have responded with willingness as they have provided me with really detailed feedback, which is a great indicator that our cooperation will be excellent. As I am working my way towards understanding the library better it could be beneficial to discuss my questions (which could be many given the size of Networkx) and also get my mentors' help in implementing a smaller algorithm so that I can become more familiar with the way Networkx works. In that way, there will not be any time loss on the main Coding Period and I will be able to focus on it right away.

-- June 7 - July 12 (Phase 1)

In this phase the main Louvain algorithm should be implemented. All the helper functions along with the functions that return the community clusters are done, documented and proper tests have also been implemented.

-- July 12 - July 16 (First Evaluation)

At this stage the first Pull Request will be opened and receive feedback from the mentors and rest of the community. Further discussions are needed for the wrapper class that will handle the output format of the algorithm.

-- July 16 - August 16 (Phase 2)

In this period the wrapper class of the Louvain Algorithm is implemented. Also there could be some issues that arose regarding the progress so far and therefore extra care is needed in order to fix them. At the end of this month, the full Louvain Algorithm should be ready to be merged, documented and tested well.

-- **August 16 - August 23**

Final discussions regarding my implementation are made at this stage, getting feedback from my mentors and correcting possible minor bugs. At this stage though bugs should be minimal. Further development is discussed and suggested issues are formed for future work beyond GSoC. One possible next step is to implement the parallel version of Louvain's Algorithm which is presented here

Note: Reviewing periods are very crucial and could or should take the same time as the development process. Code should be handled with professionalism and no margin for errors should be left.

4. Motivation

4.1 Why GSoC?

My motivation to take part in Google Summer of Code can be concluded in the points below:

- **Become an active open source contributor**: Open source software plays a vital role in every developer's coding life. Sharing ideas and providing free software to people is a virtue on its own. As I graduate from my school, a dream of mine is to become an open source contributor and broaden my software knowledge in this way.
- **Spend summer months coding**: In summer people lay low from their jobs here in Greece. I would like to take advantage of these three loose months and enhance my coding skills by working on something that is very interesting to me while making some income in parallel, that will aid me in future studies.
- **Interact with other people**: In GSoC, I will have the opportunity to gain a lot of knowledge and experience from my mentors. Interacting with people from the open source community is extremely valuable to me. From knowledge sharing to everyday talks with highly skilled developers around the world, GSoC seems to be a thrilling experience.

4.2 Why NetworkX?

From the organizations listed in the main GSoC page, NumFocus caught my eye since it was a lot about mathematics and had a lot of different sub-organizations and projects to look into. I have always had an interest in algorithms and their actual implementation which is why I chose NetworkX specifically, given that the whole project is to write code for one. Also, NetworkX is a very big library used by a lot of people and I find it quite exciting to be able to contribute to an organization like this and at the same time help the people who use this library to work better. I have the feeling that my interests fully conform with the core subject of NetworkX and thus GSoC could operate as a starting step in a long-lasting relationship.

4.3 Why this specific project?

Louvain Community Detection Algorithm is a very important one since it is very efficient and also produces good results. Therefore, implementing it inside NetworkX could be very good and also an interesting project. Since I have studied some graph theory and discrete mathematics at my university, I found this project very exciting and also very feasible for my programming level and experience. I also love python, so it can be a great exercise to improve my skills as a programmer. In conclusion, I feel that I am very capable of understanding certain concepts of community detection algorithms and therefore that this project fully corresponds to my potential.

4.4 Why choose me?

I am a fast learner and try to work as efficiently as I can. In order to cope with the high academic level of the courses, the high amount of workload at each course and the intense examination periods, I have sharpened my ability to work uninterruptedly under pressure through demanding school projects all these years in the environment of a demanding computer engineering school. Also the past few months working for a real company made me understand that good, quality code is very important. My willingness to contribute to an **open source python** project is huge and I plan on sticking with the team and become a regular member of NetworkX. I have already devoted a lot of time studying and using NetworkX and what proves my urge to contribute is that **I have already made my first pull request** that has been merged! I plan on working on some other issues or implementing some smaller algorithms as well until the coding period starts and stick with the organization regardless of my acceptance in GSoC. I am confident that I have what it takes to see the project through by the end of summer and that is why you should trust me with this assignment.