

# Structure and Evolution of GitHub Collaboration Network

Zagaria Simone  
Matr. 2145389



Data-Driven Modeling  
of Complex Systems

# Introduction

## What?

- Our objective is to analyze the **Collaboration Network** between users on Github to understand its **structure** and **evolution**

## Why?

- To **understand the dynamics** of the interactions between users on github and how it has grown over time
- Seemed an interesting research study

## How?

- **Github Membership Network**<sup>1</sup> (177.000 nodes, 440.000 edges)

[1] Source: <http://konect.cc/networks/github/>

# The Starting Dataset

|                |                                        |
|----------------|----------------------------------------|
| Node meaning   | User, project                          |
| Edge meaning   | Membership                             |
| Network format | <b>B</b> Bipartite, undirected         |
| Edge type      | <b>—</b> Unweighted, no multiple edges |

The initial Dataset was a **bipartite, undirected graph** containing projects on one side and users on the other.

The Dataset also had **timestamps** for the **creation of projects**, ranging from mid-2008 to mid-2009

However, to analyze collaborations between users, we projected this bipartite network onto the user set, creating a **Unipartite, undirected weighted graph**

|                            |                 |
|----------------------------|-----------------|
| <a href="#">Size</a>       | $n = 177,386$   |
| <a href="#">Left size</a>  | $n_1 = 56,519$  |
| <a href="#">Right size</a> | $n_2 = 120,867$ |
| <a href="#">Volume</a>     | $m = 440,237$   |

# Building the Collaboration Graph

~ 110K nodes

~ 11 milion  
edges

The Final Collaboration Graph was constructed as follows:

**Nodes:** represent **Users**.  
The attribute '*num\_projects*' indicates the number of unique projects for each user.

**Edges:** represent **Collaborations between users**. The weight is proportional to the number of projects shared by the users.

**Isolated nodes** were deleted for a more insightful analyses.

# Research Questions

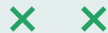
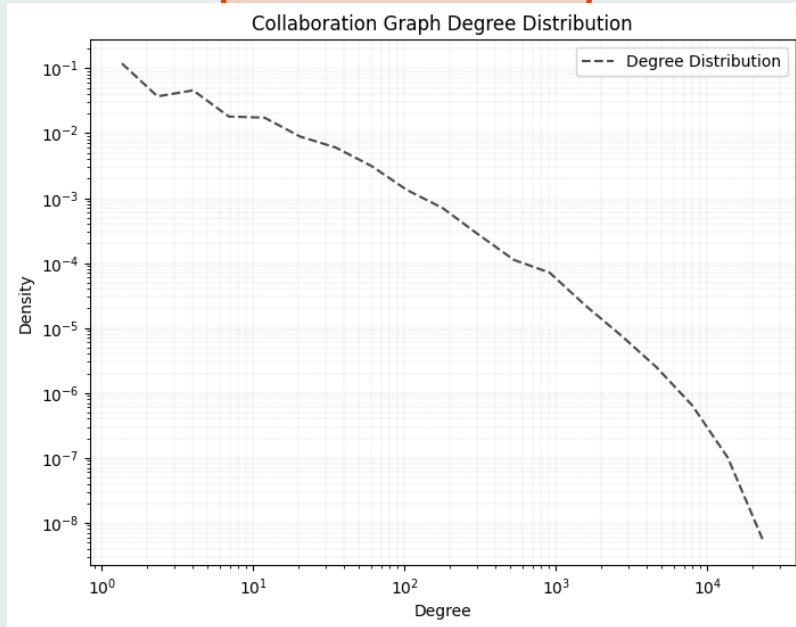
**1. How is the Collaboration Network among GitHub users structured?**

**2. How does the Collaboration Network evolve over time? Are collaborations typically recurrent or users tend to seek new collaborators?**

# Overview of Graph Metrics



(Log) Degree  
Distribution:



Largest  
component  
size:  
**99907 nodes**

Number of  
connected  
components:  
**2948**

Density:  
**0.0019**

# Scale-Free Property Analysis



The graph showed characteristics consistent with a scale-free network



=== Scale-Free Analysis ===

Gamma: 2.3089

p-value: 0.0000

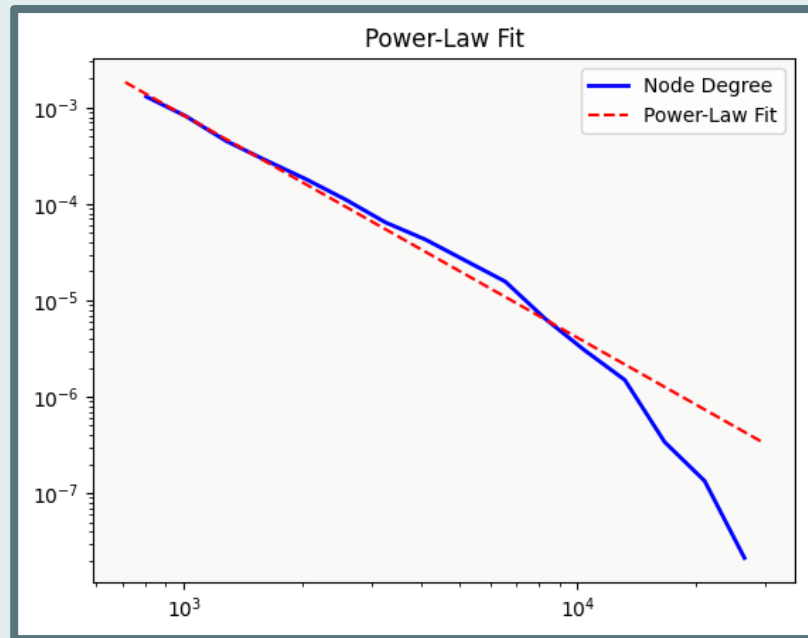
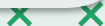
R: 11.9656

There is significant evidence to suggest the graph follows a power-law distribution

The estimated exponent of the power law for this network is  $\gamma=2.3$ , which falls in the typical range for scale-free networks

A comparison with exponential distribution showed a **p-value** of **0.00** and a **high R value**, indicating a significant and strong tendency towards a Scale-Free distribution

From the chart, we observe that the curve is consistent with a power law.



# Community Detection



N° Communities  
detected:  
**3382**

Louvian algorithm  
Modularity:  
**0.2811**

For this analysis we chose the **Louvian Community Detection algorithm**, as it's one of the fastest and most scalable:

**3382 communities** detected, with more central users often acting as **hubs**.

Among the communities found, there was a balance between smaller and larger communities. However, the modularity value indicates a **relatively weak community structure**.



# Centrality Metrics

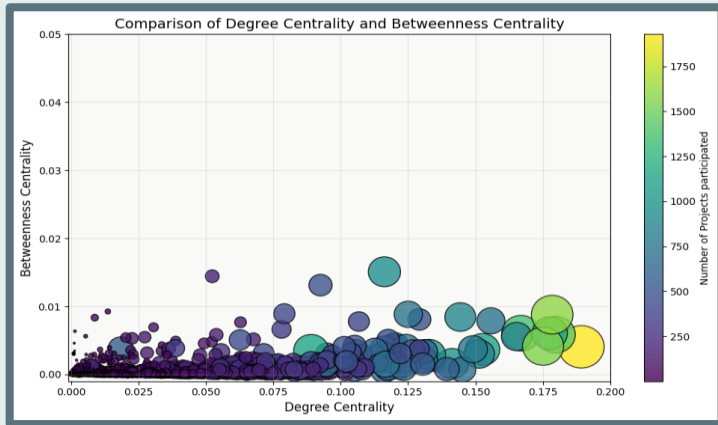


| id       | degree centrality | betweenness centrality |
|----------|-------------------|------------------------|
| user_17  | 0.272642          | 0.028591               |
| user_299 | 0.189293          | 0.004054               |
| user_607 | 0.179747          | 0.005841               |
| user_645 | 0.178488          | 0.008774               |
| user_8   | 0.177577          | 0.006048               |

Degree centrality highlighted users with the most connections in the network, confirming their role as **hubs**.

Betweenness centrality identified users who act as **bridges** connecting different parts of the network.

The chart confirms that most users cluster around low centrality measures and that users who participated in more projects tend to be more central.

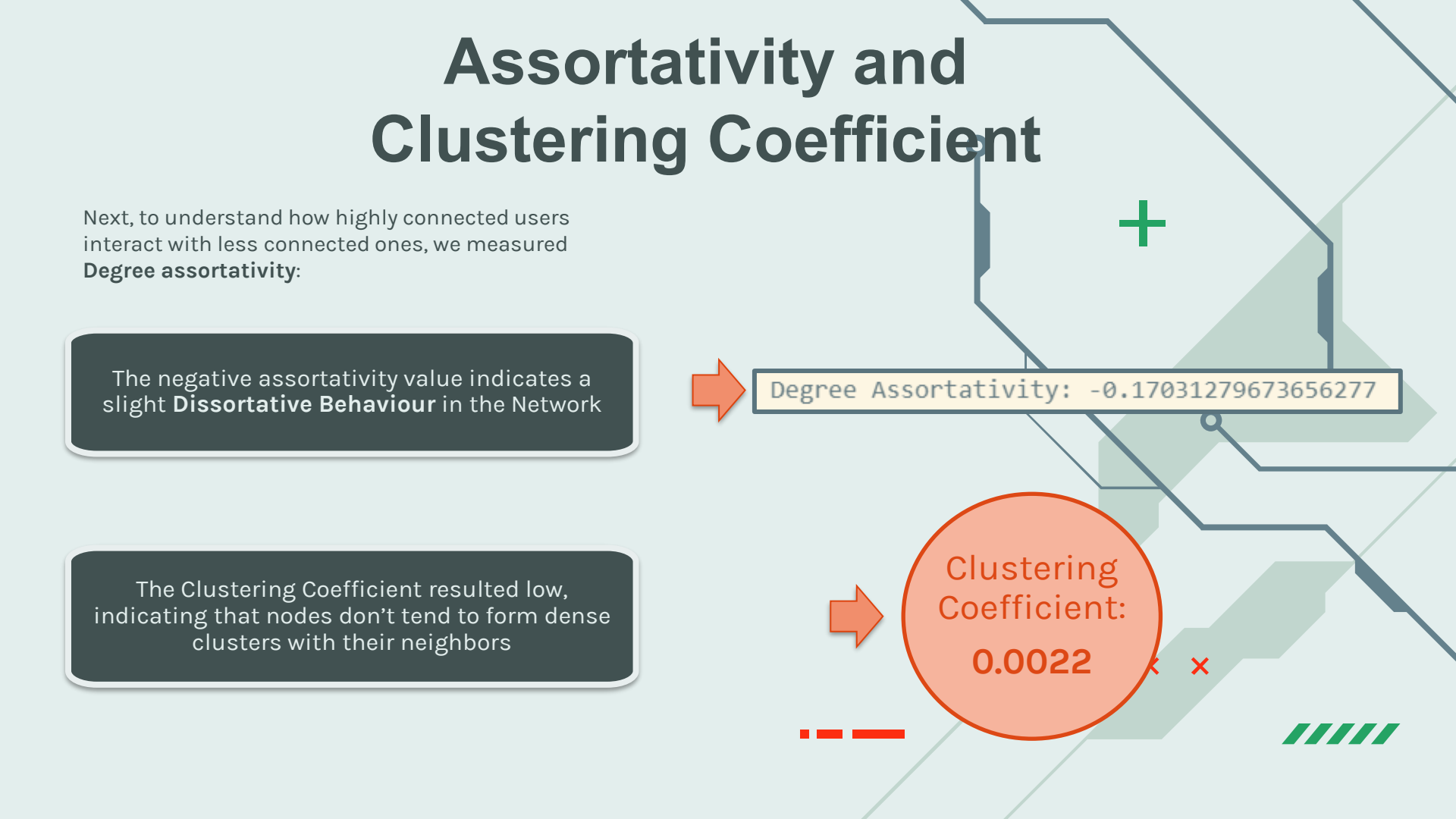


# Assortativity and Clustering Coefficient

Next, to understand how highly connected users interact with less connected ones, we measured **Degree assortativity**:

The negative assortativity value indicates a slight **Dissortative Behaviour** in the Network

The Clustering Coefficient resulted low, indicating that nodes don't tend to form dense clusters with their neighbors



Degree Assortativity:  $-0.17031279673656277$

Clustering Coefficient:  
 $0.0022$



## x x Research Question 2

**How does the Collaboration Network evolve over time? Are collaborations typically recurrent or users tend to seek new collaborators?**



# The Temporal Collaboration Network

To answer this research question, we used the timestamps of the project creations of the original bipartite graph to recreate the collaboration network at **specific times t**.

In particular, we chose to compare **3 stages** of its life to the final graph, in order to understand how it has evolved over time. We called them:

Early Graph:  
t = 2008-11-15

Middle Graph:  
t = 2009-02-09

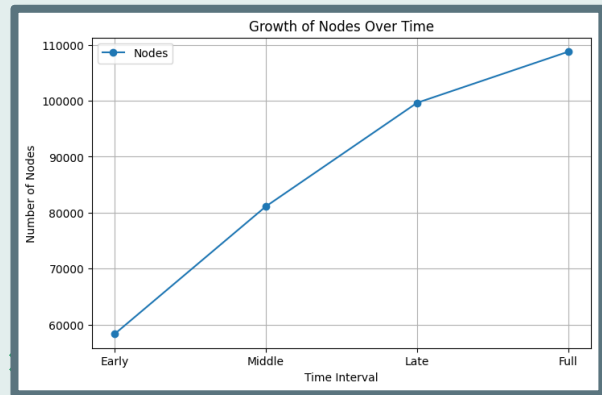
Late Graph:  
t = 2009-05-18

# Early, Middle, Late graphs

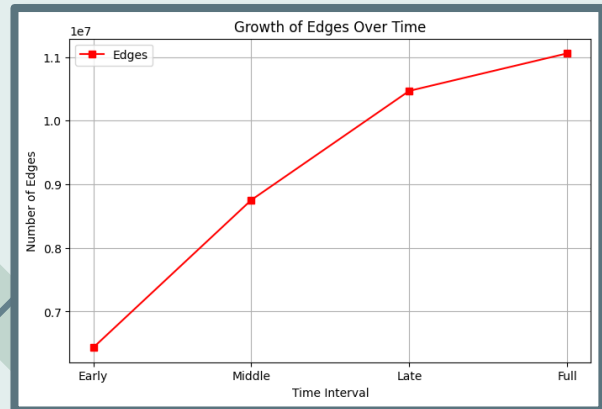
Firstly, we observed that the number of nodes and edges **grew almost linearly** over time, slowing down only during the last interval.

The dates span 3 months from each other.

We notice the graph nearly doubled its size in just the span of 6 months.



✕ ✕



Early Graph:

Number of nodes: 58346

Number of edges: 6434649

Middle Graph:

Number of nodes: 81123

Number of edges: 8753082

Late Graph:

Number of nodes: 99623

Number of edges: 10468553

Full Graph:

Number of nodes: 108748

Number of edges: 11058194

# Comparison of Clustering Coefficient and Degree Distribution

Estimated Clustering Coefficient:

Early: 0.0039

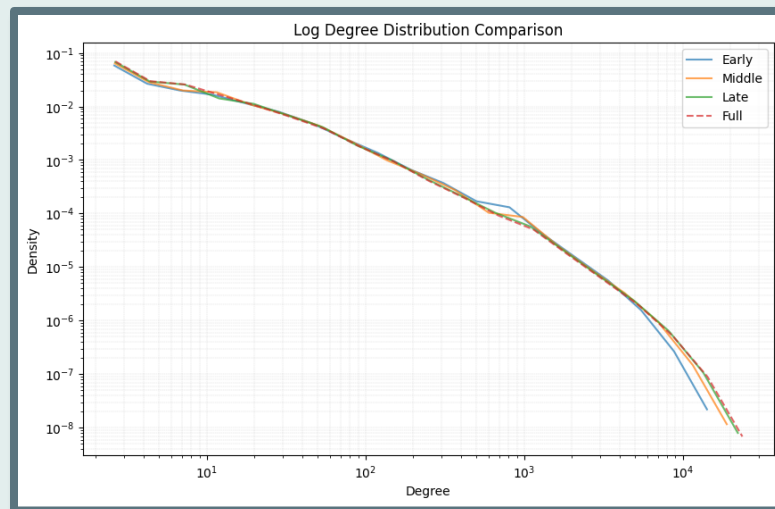
Middle: 0.0030

Late: 0.0025


The **clustering coefficient decreases over time**, indicating that the network becomes less locally clustered as it grows. This is consistent with the decreasing density and modularity over time.

Comparing the degree distributions, we observe that the graph always maintained a similar trend, with slight variations to the high degree nodes.

This suggests that the graph has always been prone to the **formation of hubs**.



# Community Detection Comparison




| Interval | Modularity |
|----------|------------|
| Early    | 0.336062   |
| Middle   | 0.298589   |
| Late     | 0.284773   |
| Full     | 0.280181   |

Community Detection with Louvian Algorithm was used for each graph, and then the divisions for each interval were **compared to the Final graph**:

**Modularity decreased over time.** This was expected as the network grows in terms of nodes and edges, while the density of the graph decreased, reducing overall cohesion.

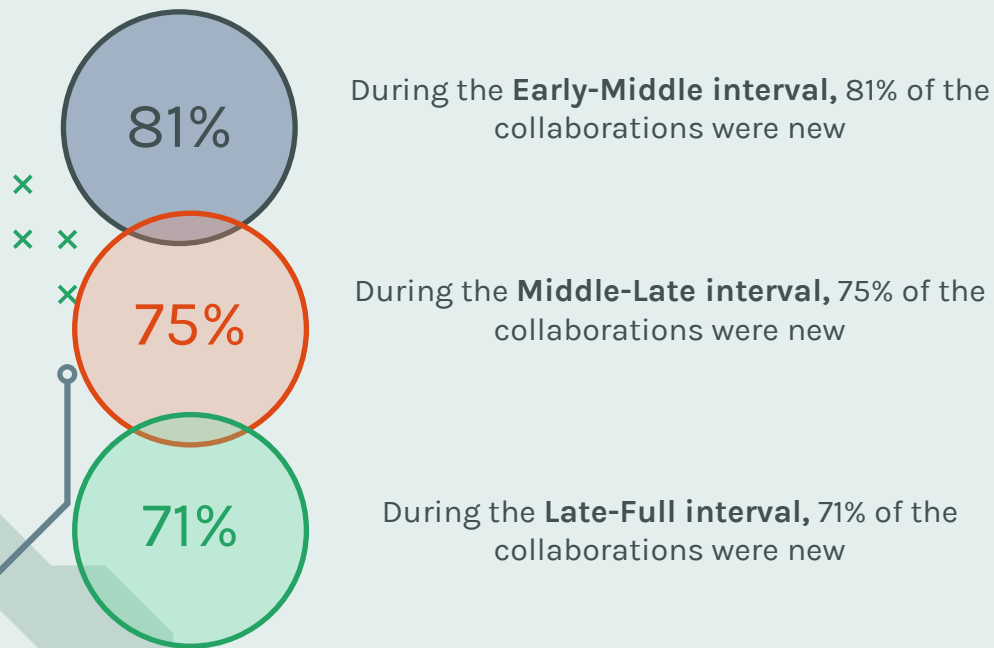
The number and size of communities divisions also varies a lot over time, indicating that **communities are not stable but evolved and changed as the network grew.**



# Collaboration Recurrency

Lastly, we analyzed the new edges created during each interval to find out whether users tend to prefer recurrent collaborations over new collaborations.

The results showed that, **compared to the previous interval**:



Thus, this shows that users on Github tend to **prioritize new collaborations over recurrent ones.**



**Thank you  
for the  
attention!**

