Networks of Knowledge: The World of Stack Overflow Users

University of Pittsburgh

Name: Riley Vetere Jones
Class: INFSCI 2415
Instructor: Dr. Lingfei Wu

1 Figures

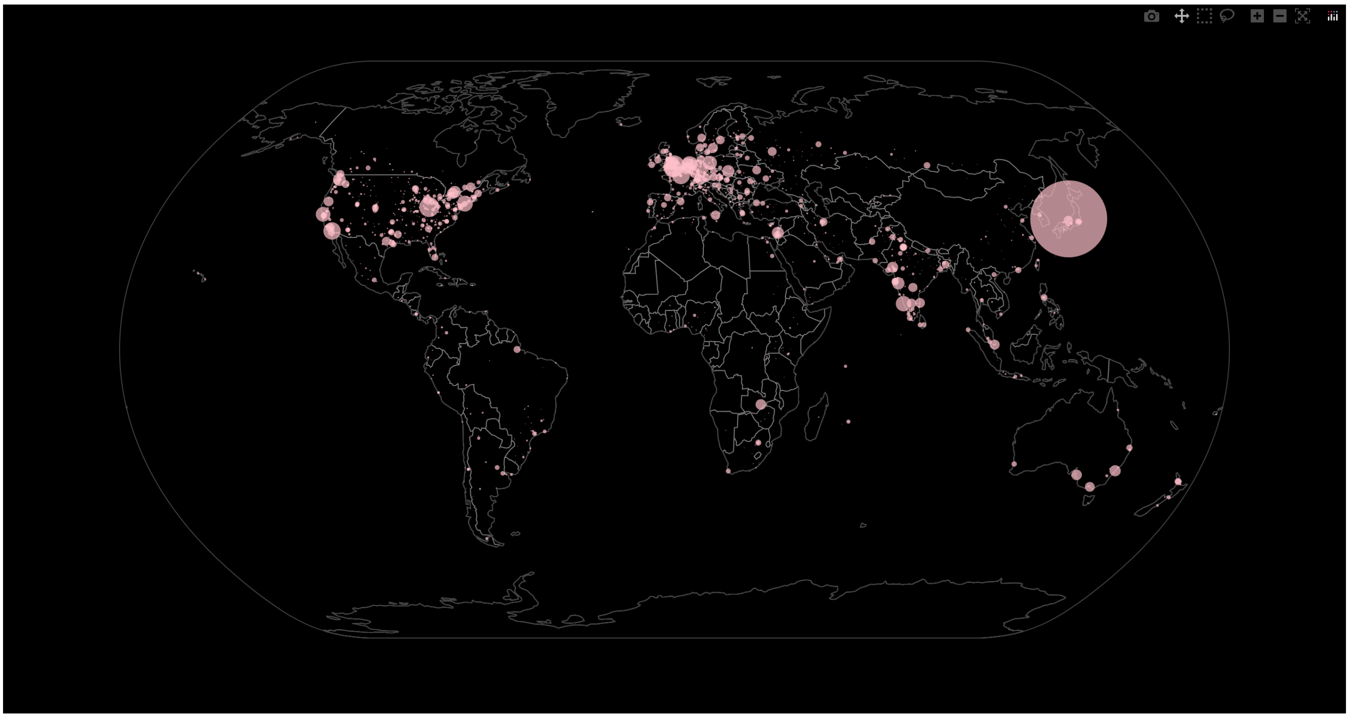

Figure 1: Map of Top 50,000 Stack Overflow Commenters

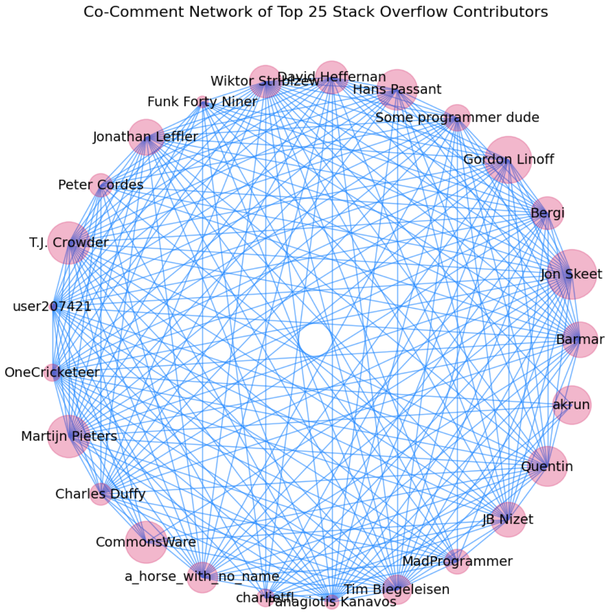

Figure 2: Co-Comment Network of Top 25 Commenters

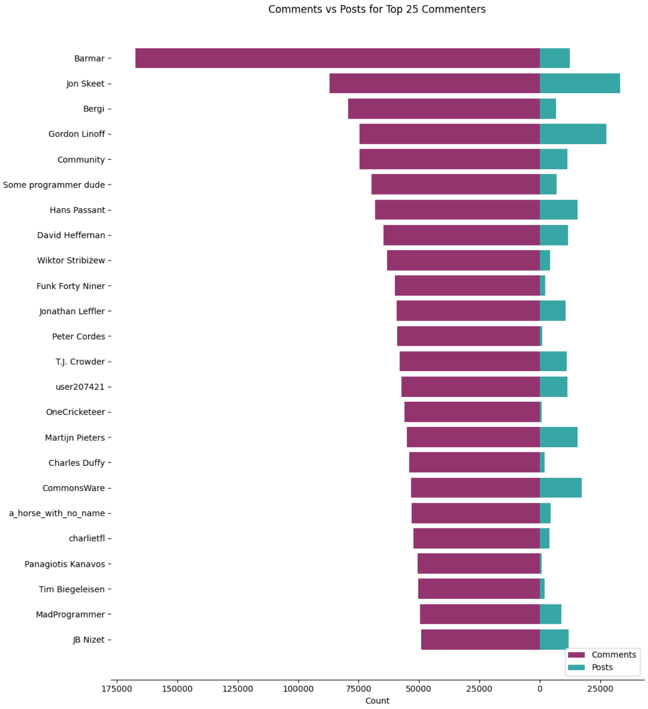

Figure 3: Mirrored Bar Plot for Top 25 Commenters

1.1 Legend

In Figure 1, each point represents a location of a top 50,000 commenter on Stack Overflow. The sizes of the points are scaled by how many top users are located at that point.

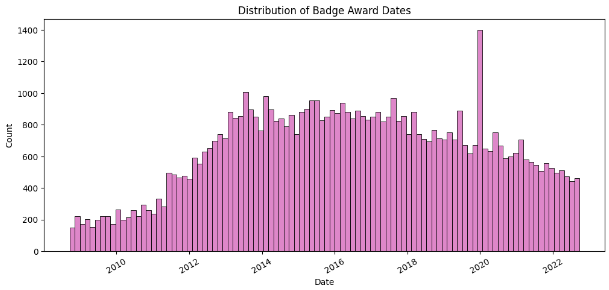

Figure 4: Time Series Bar Chart of Badge Award Dates for Top 25 Commenters

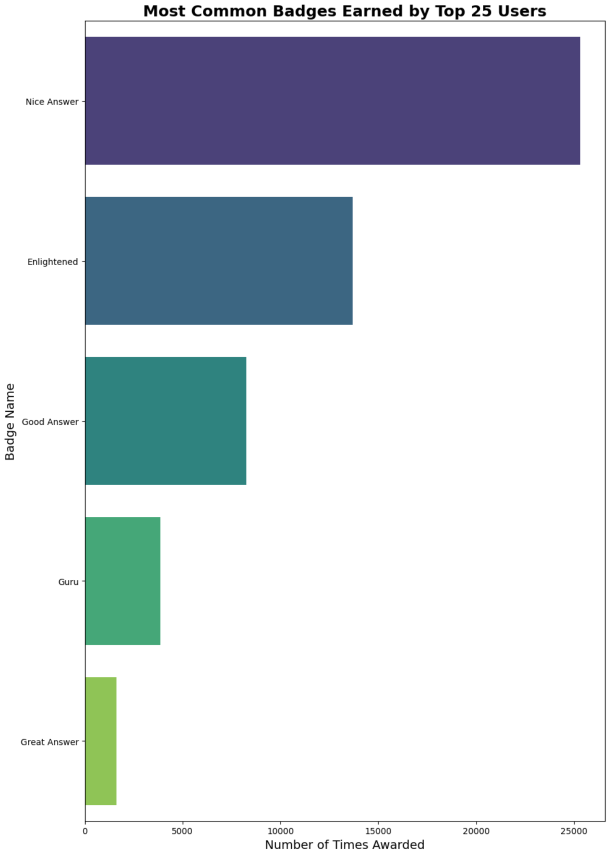

Figure 5: 5 Most Common Badges Awarded to Top 25 Users

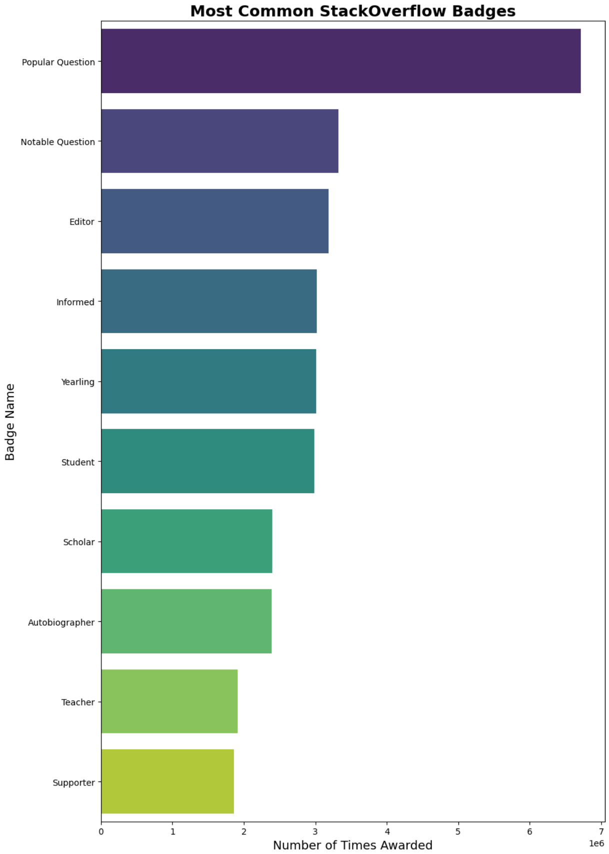

Figure 6: 10 Most Common Badges Awarded to All Users

In Figure 2, each node represents a top 25 commenter by comment count on Stack Overflow, with their username overlaid on the node. An edge connects users who commented on the same post. Node sizes are scaled by reputation, a measure of community trust.

In Figure 3, each bar shows the count of comments on the left in purple and the count of posts on the right in teal.

In Figure 4, each bar represents a count of how many badges were awarded to top 25 users on the date given on the x-axis.

In Figure 5, each bar represents the count of how many badges of that type were awarded to top 25 users. The bars are arranged in descending order, and the top 5 most common are shown.

In Figure 6, each bar represents the count of how many badges of that type were awarded to all users. The bars are arranged in descending order, and the top 10 most common are shown.

2 Findings

2.1 Figure 1

- Stack Overflow users are spread out globally.
- Nonoichi, Japan is the largest hub of top commenters, with 13,333.
- The global North has a higher concentration of top commenters.

2.2 Figure 2

- Stack Overflow is highly collaborative.
- Few isolated nodes—top commenters form a cohesive network.
- Collaboration shows little relationship with reputation - even small nodes have many edges

2.3 Figure 3

- Status as a top-commenter is not correlated with a high volume of posting.
- There is a little variability in the comment count of top commenters, with the exception of the top commenter, Barmar, being an outlier.

2.4 Figure 4

- Badge awarding for these users steadily increases over time, tracking the top 25 users' rise in esteem over the years.
- There is a significant mode in 2020, which could indicate more badge awarding due to the effect of the Covid 19 pandemic.
- The badge awarding for the top 25 users is on the decline, potentially pointing to a new set of users that overtake them as top users.

2.5 Figures 5 & 6

- Unsurprisingly, the badges for top commenters are all for commendable answering of questions. This differs from the badges for all users, which are more related to posting questions.
- There is a significant mode in 2020, which could indicate more badge awarding due to the effect of the Covid 19 pandemic.
- The badge award for the top 25 users is on the decline, potentially pointing to a new set of users that overtake them as top users.

3 Data & Methods

The data is from the Stack Overflow data on Kaggle. I used the users (11 million rows), comments (75 million rows), and badges (33 million rows) tables. I also used the Science Cities dataset that is available on the course canvas page. The entire project was completed in Kaggle's Notebook Editor.

Each graph contains data that has been joined with other tables and queried using SQL. SQL queries were created with work found on Kaggle and my own introductory knowledge of SQL. The query for Figure 1 was limited to the top 50,000 to limit run time and ensure readability of the final figure.

For Figure 1, ChatGPT created an algorithm with the annoy package to fuzzy match user locations to the cities in the provided data set (OpenAI (2025))(Bernhardsson (2018)).

4 Significance

Understanding the behavior of Stack Overflow users provides valuable insight into the dynamics of online communities. By analyzing badge awards, comment date distributions, and geographical location, this can reveal patterns of user engagement, motivation, and participation across time and space. Such insights can inform the design of more effective social networking systems, improve community moderation strategies, and guide developers in fostering sustained, high-quality contributions. Additionally, identifying regional or temporal trends in activity can help researchers better understand how people around the world interact with and contribute to the sharing of knowledge.

5 Materials

- Data: Kaggle : StackOverflow (2018)
- SQL Query: Elior Ben Harush (2022)
- GitHub: https://github.com/rveterejones/InfoViz_Project

References

Bernhardsson, E. 2018, Annoy: Approximate Nearest Neighbors in C++/Python. https://pypi.org/project/annoy/

Elior Ben Harush, David Malchin, B. B. R. K. A. Z. 2022, Stack Overflow Project, https://www.kaggle.com/code/d4isdavid/stack-overflow-project/notebook

Kaggle : StackOverflow. 2018, Stack Overflow Data on Kaggle, https://www.kaggle.com/datasets/stackoverflow/stackoverflow

OpenAI. 2025, GPT-5.1: ChatGPT, https://chat.openai.com/