Draft: Exploration on the Use of WDQS

Breakdown by Geography, User Agent and Referer Class

Chelsy Xie (Analysis & Report)

31 August 2016

Executive Summary

Wikidata Query Service(WDQS) was launched publicly on September 7, 2015. As the first anniversary coming up, we want to take a look into who is using WDQS, and how they are using it. In this report, we focus on the successful (HTTP status codes 200 & 304) web requests to the SPARQL endpoint, their breakdown by country, user agent, referer class, and their pattern over time.

Data

Extracting successful (HTTP status codes 200 & 304) web requests to the SPARQL endpoint from July 1st to August 29, 2016, we count the number of queries and users(identified by the combination of client IP and user agent) by country, user agent and referer class. See data.R for more details.

Cross-Sectional

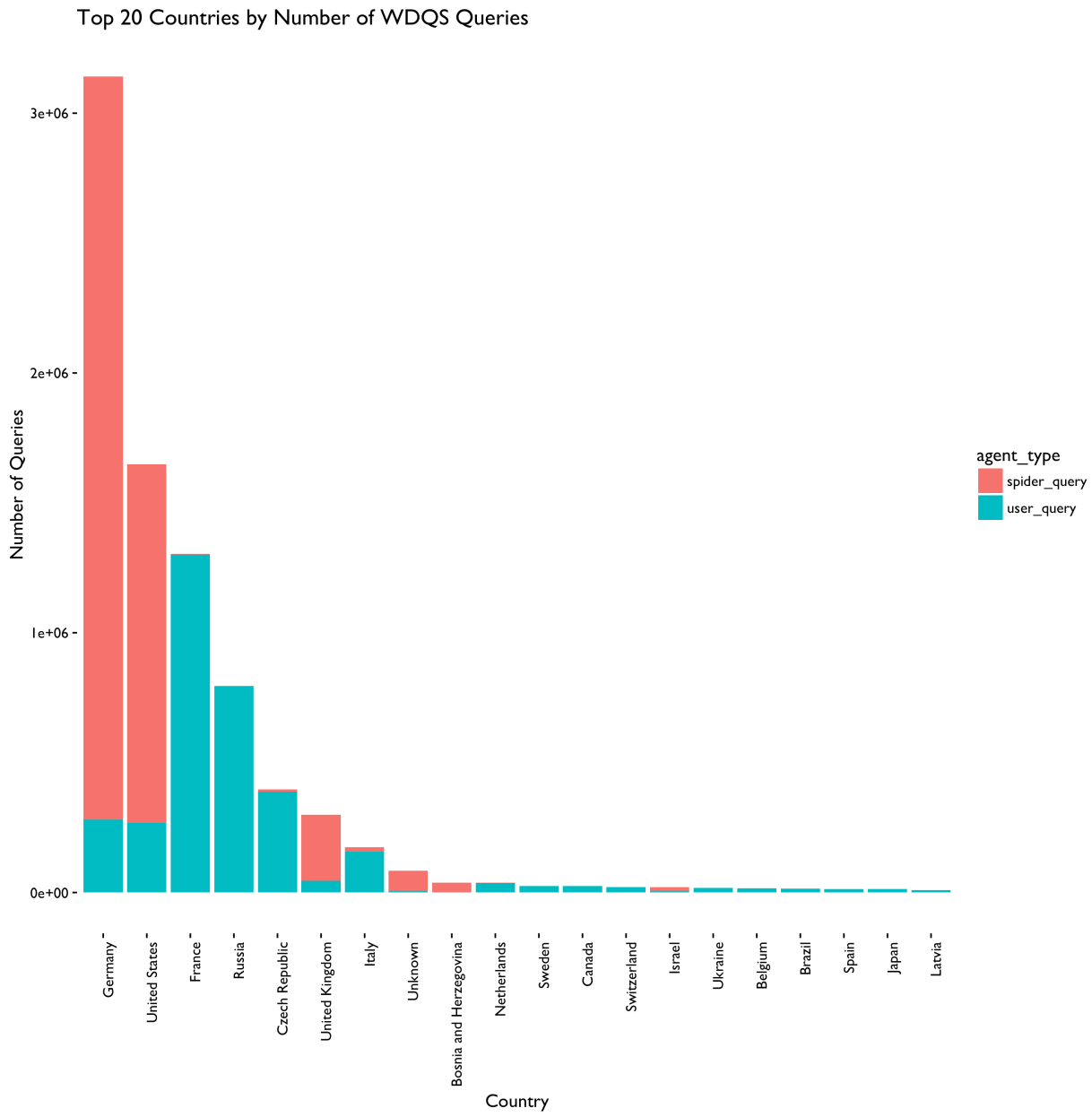

Figure 1: Germany, United States and France take the first 3 places on the rank. While most of them are spider queries in Germany and US, France has the most user queries.

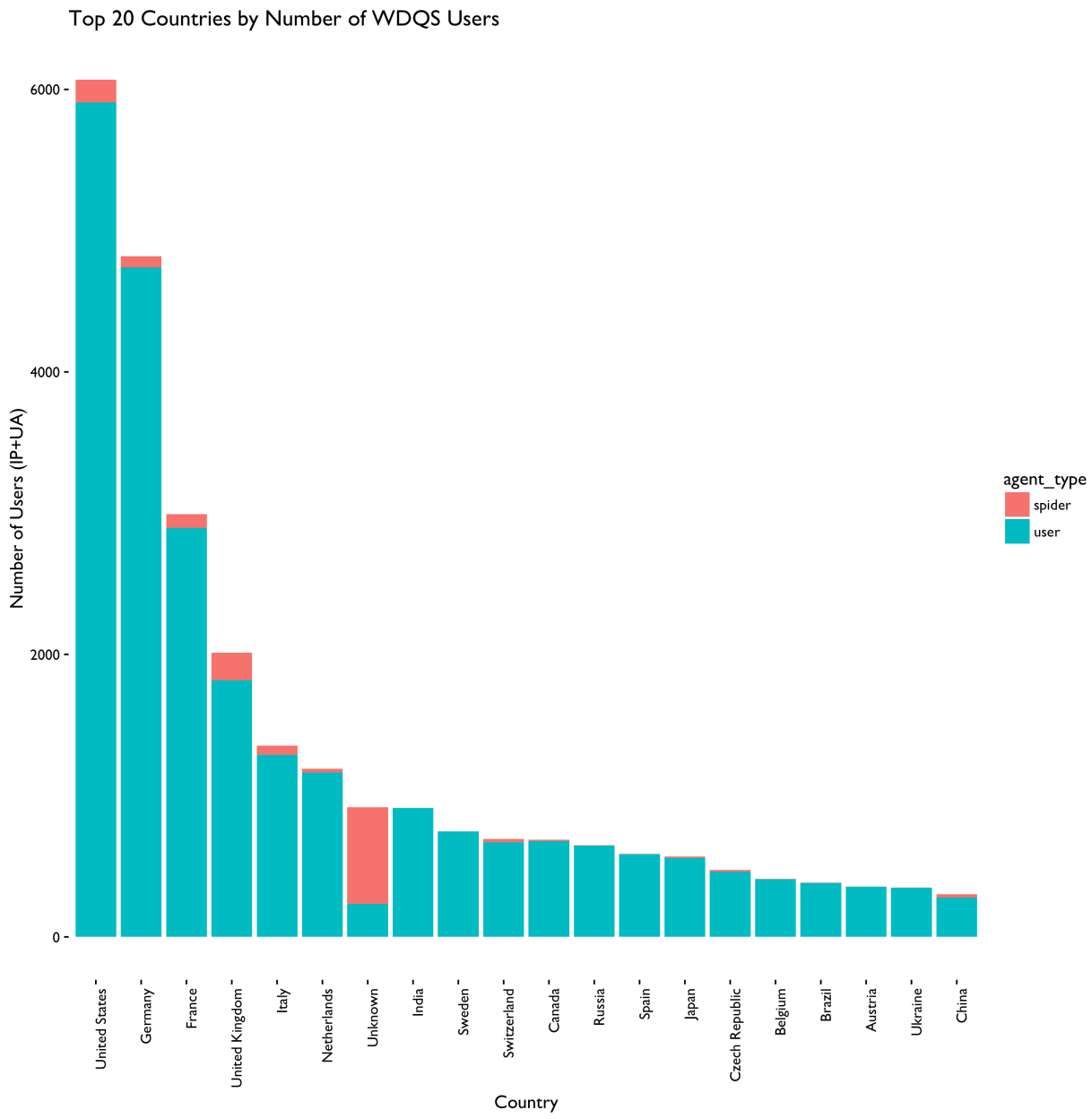

Figure 2: United States, Germany and France have the most number of WDQS users.

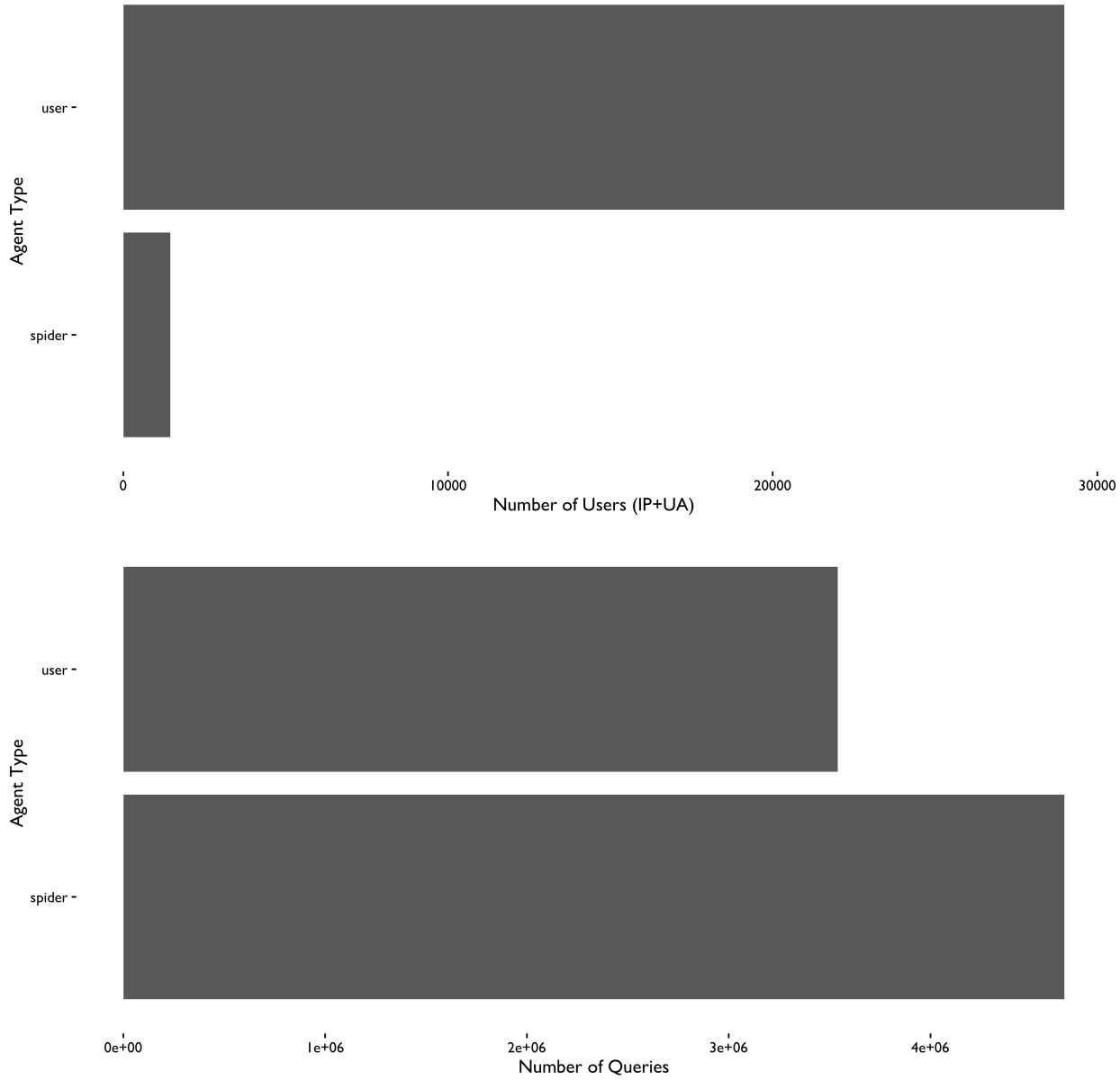

Figure 3: Number of WDQS Queries and Number of WDQS Users by Agent Type.

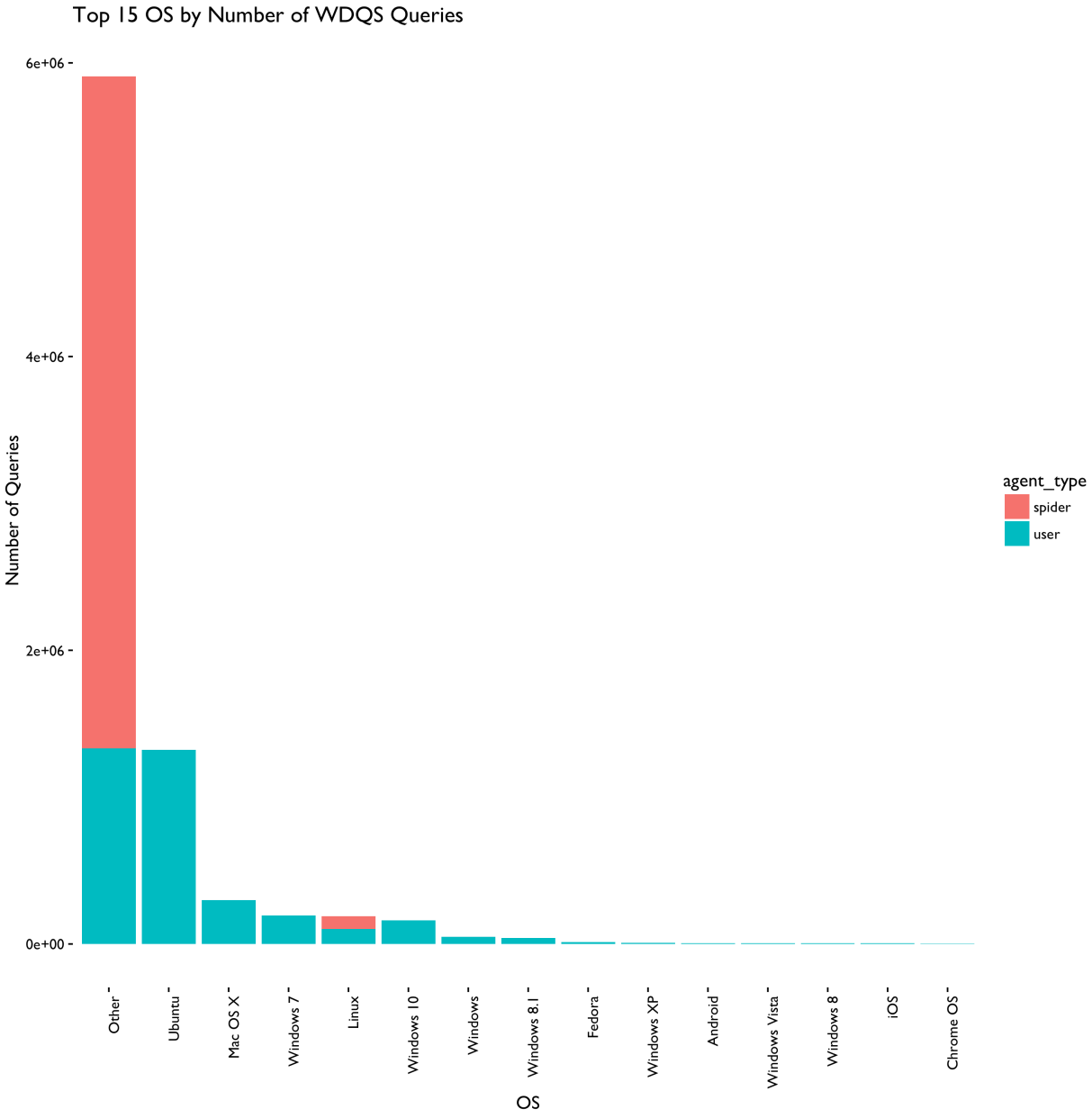

Figure 4: Top 15 OS by Number of WDQS Queries.

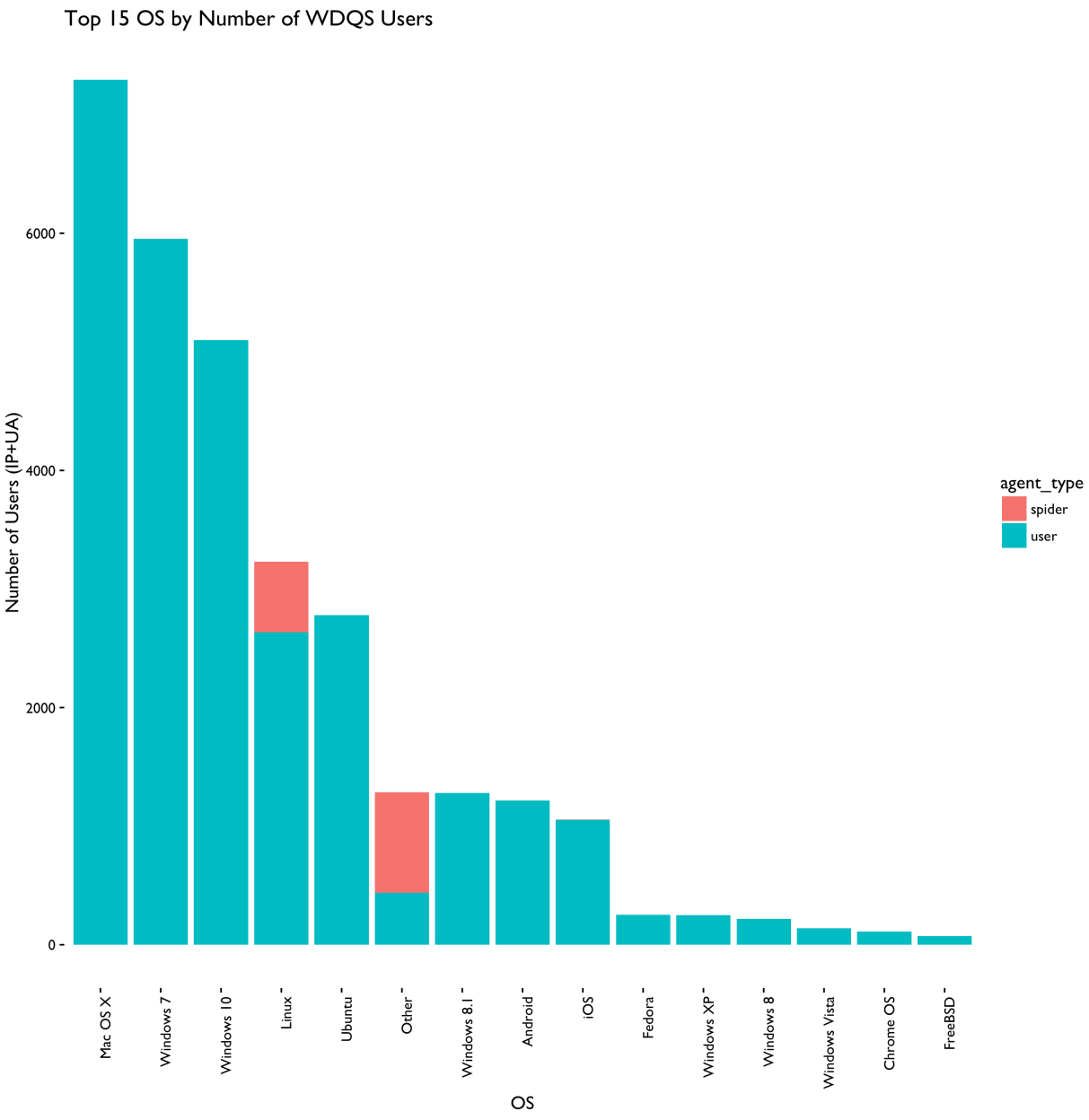

Figure 5: Top 15 OS by Number of WDQS Users.

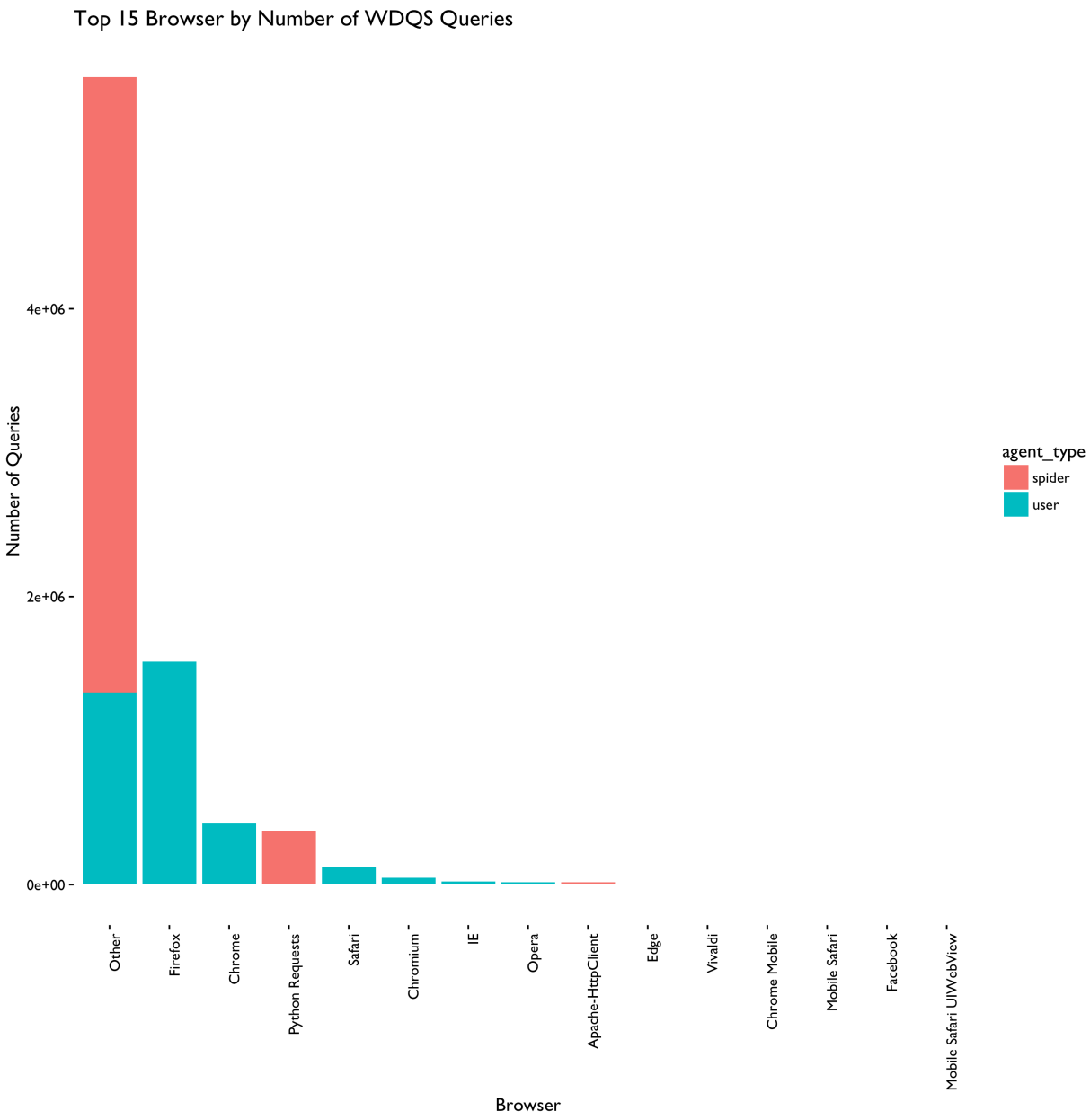

Figure 6: Top 15 Browser by Number of WDQS Queries.

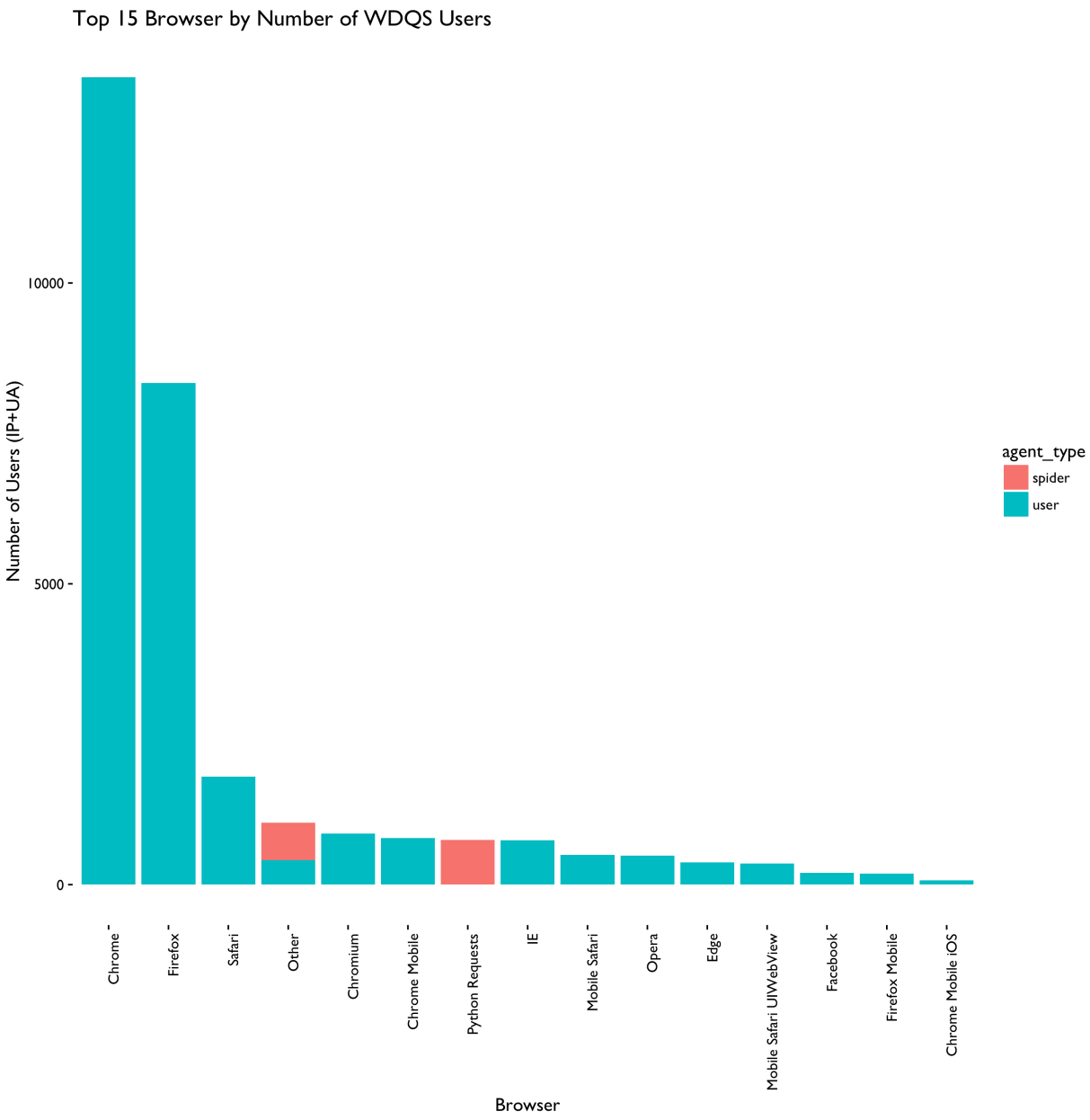

Figure 7: Top 15 Browser by Number of WDQS Users.

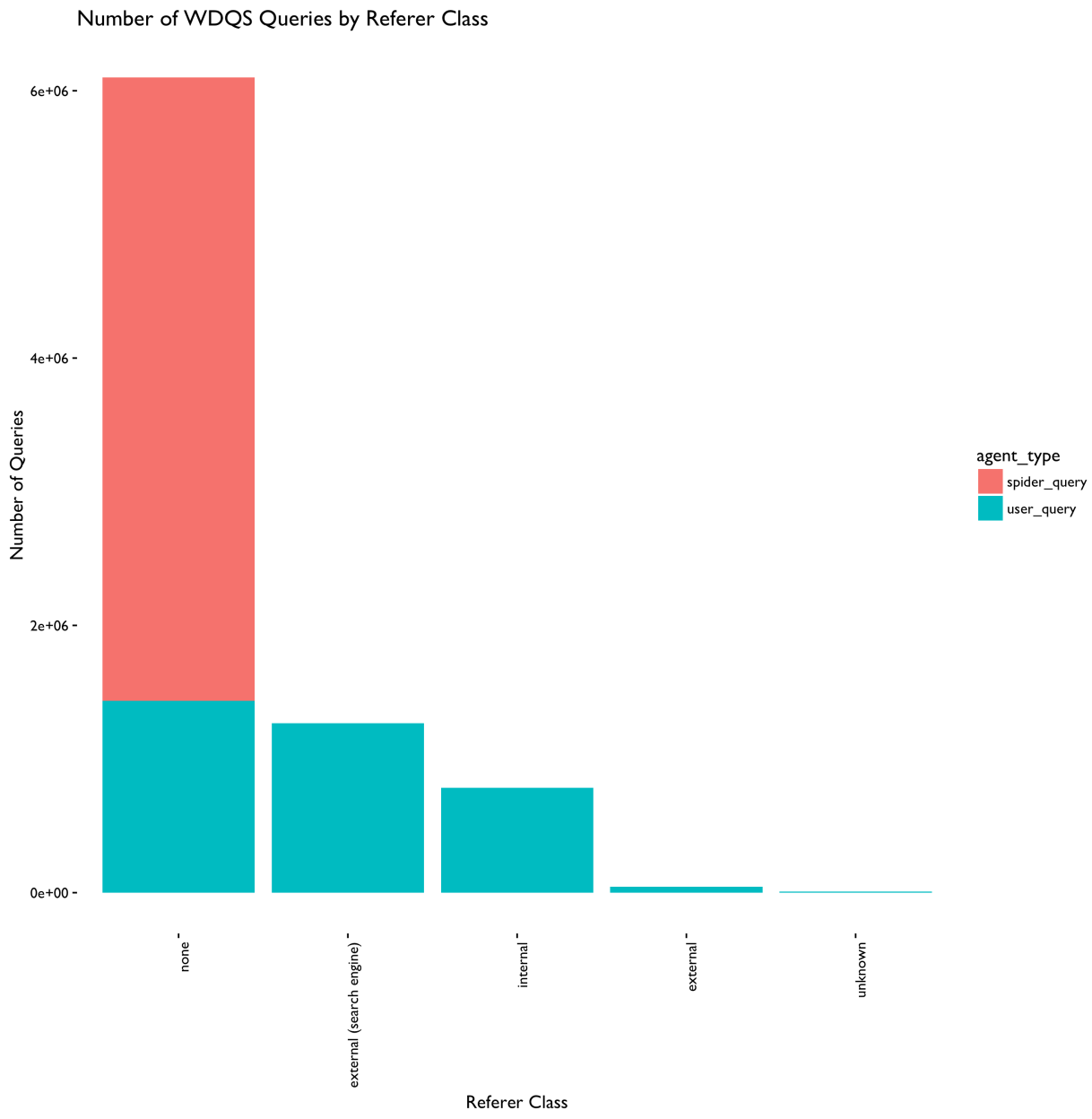

Figure 8: Most requests have no referer, followed by those referred from search engine.

Longitudinal

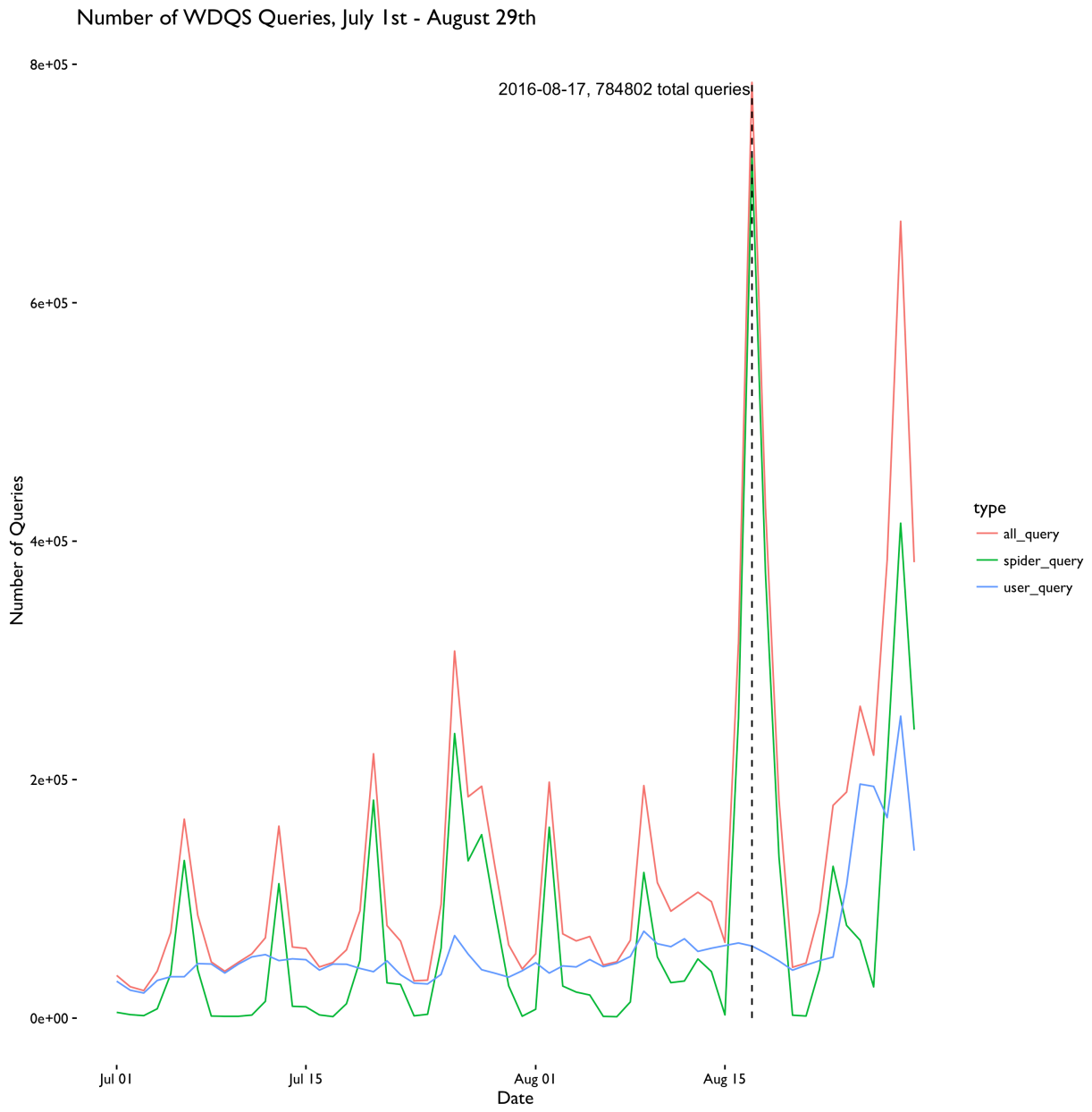

Figure 9: There seems to be a weekly cycle in the number of spider queries. After the spike(August 16-19), both type of queries saw an increase. We will do more investigation later in the report.

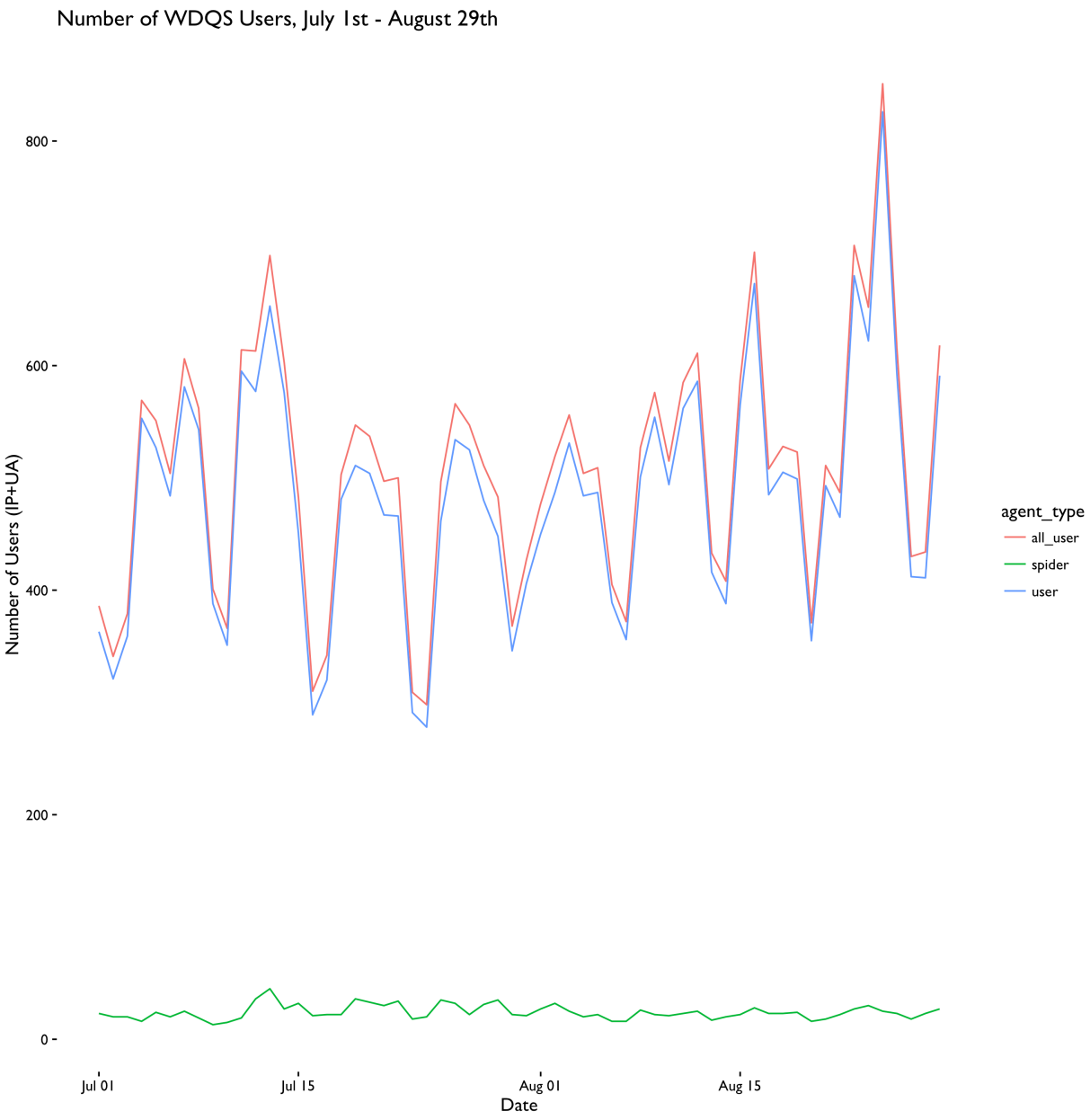

Figure 10: There seems to be a weekly cycle in the number of users.

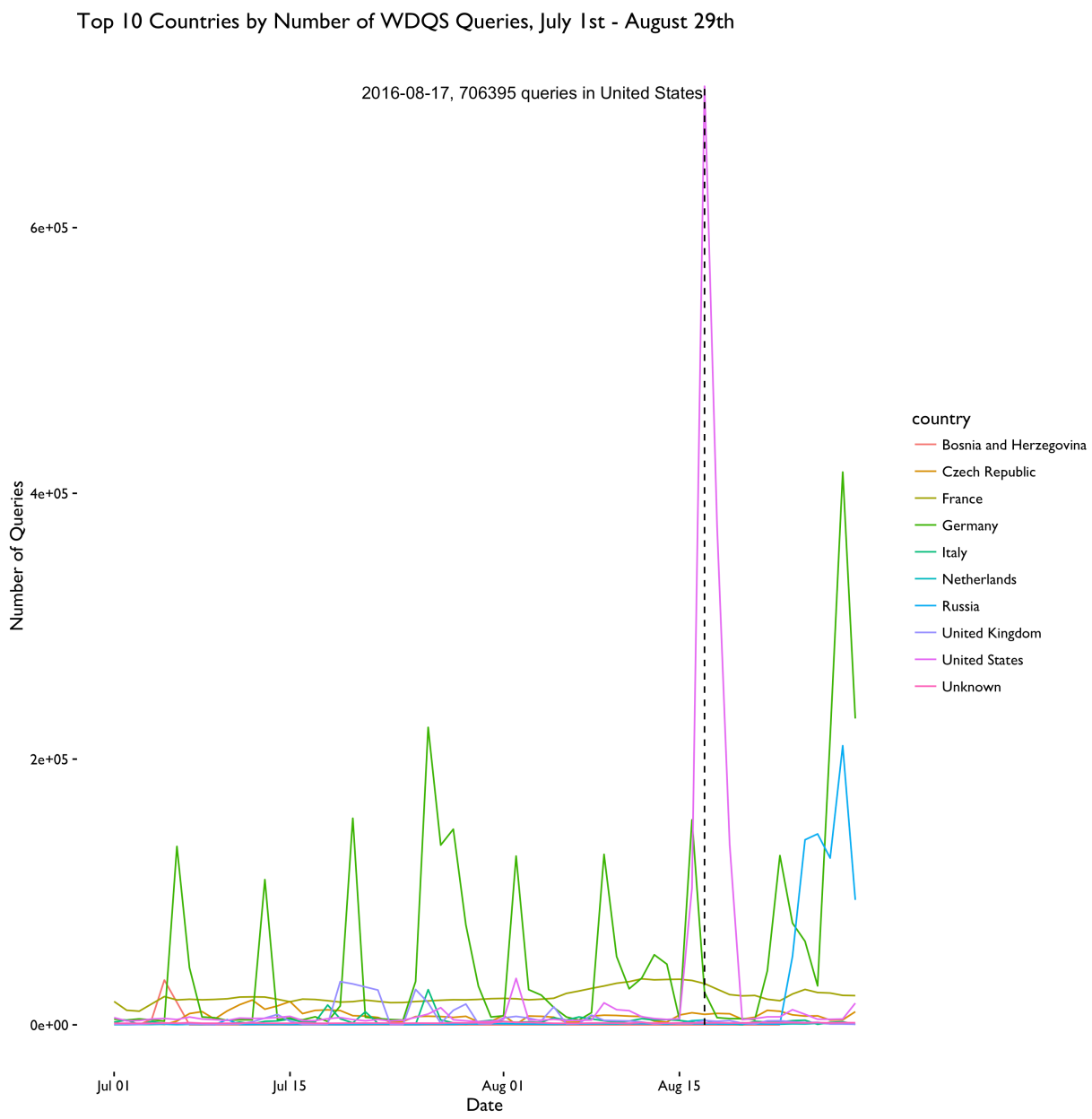

Figure 11: Further breakdown by country. The spike was contributed by the US. Germany seems to dominate the weekly cycle.

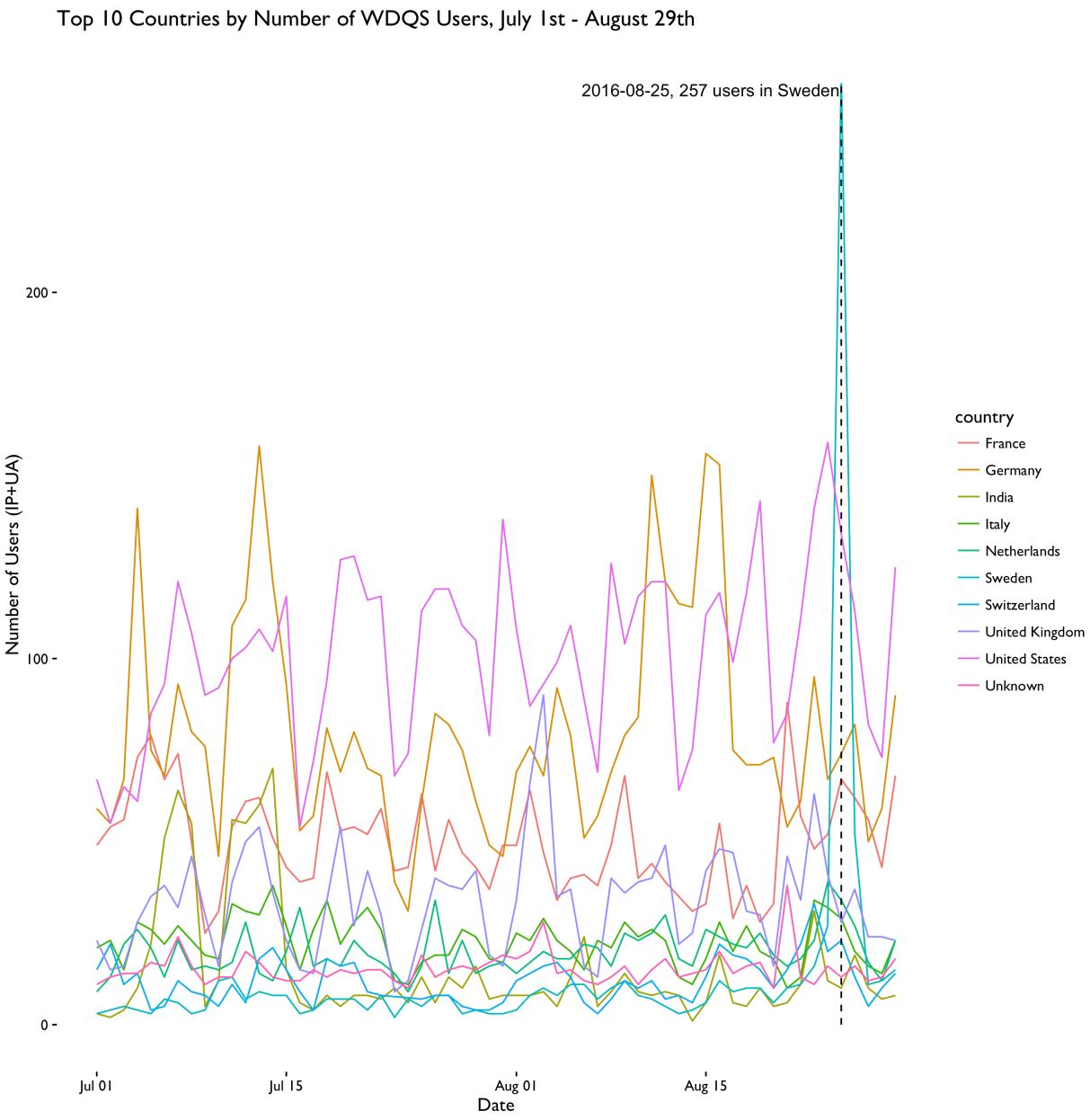

Figure 12: Top 10 Countries by Number of WDQS Users, July 1st - August 29th, 2016.

Next, we excluded the spider queries in US from August 16 to 19, then implemented BFAST method on the query data. BFAST(Breaks For Additive Season and Trend) integrates the decomposition of time series into trend, season, and remainder components with methods for detecting and characterizing change within time series. First, it decompose the series into trend and seasonal components with the STL method, then it use OLS-MOSUM test on each components to see if there is any significant break point. Next, BFAST fit the two components and the detected break points with linear regression. BFAST iteratively estimates the time and number of changes, and characterizes change by its magnitude and direction, until the number and position of the breakpoints are unchanged.

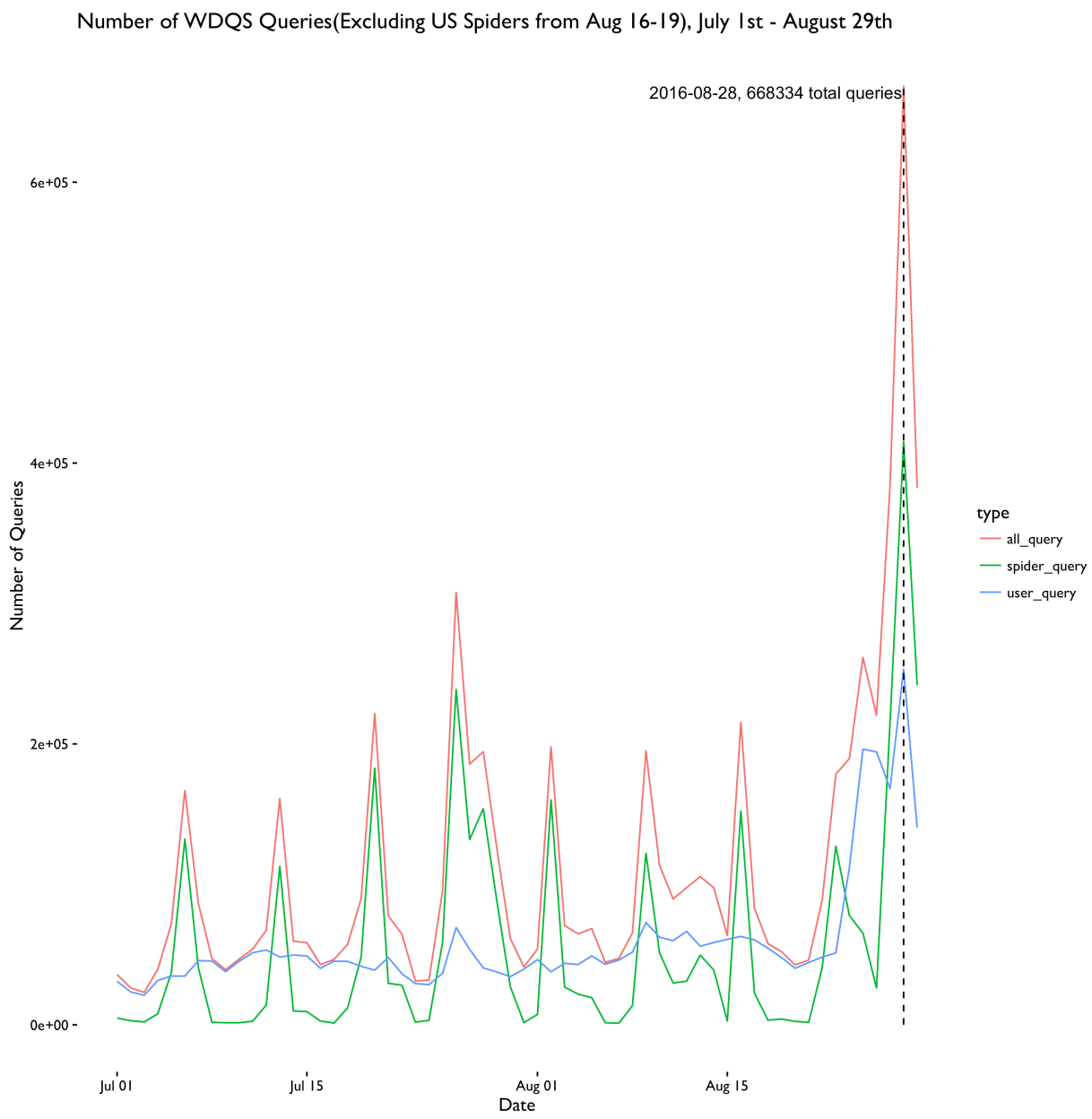

Figure 13: After excluding the spider queries from US Aug 16-19, the weekly cycle seems to hold for those days. Further investigation is needed to find out whether this spike is contributed by a particular spider.

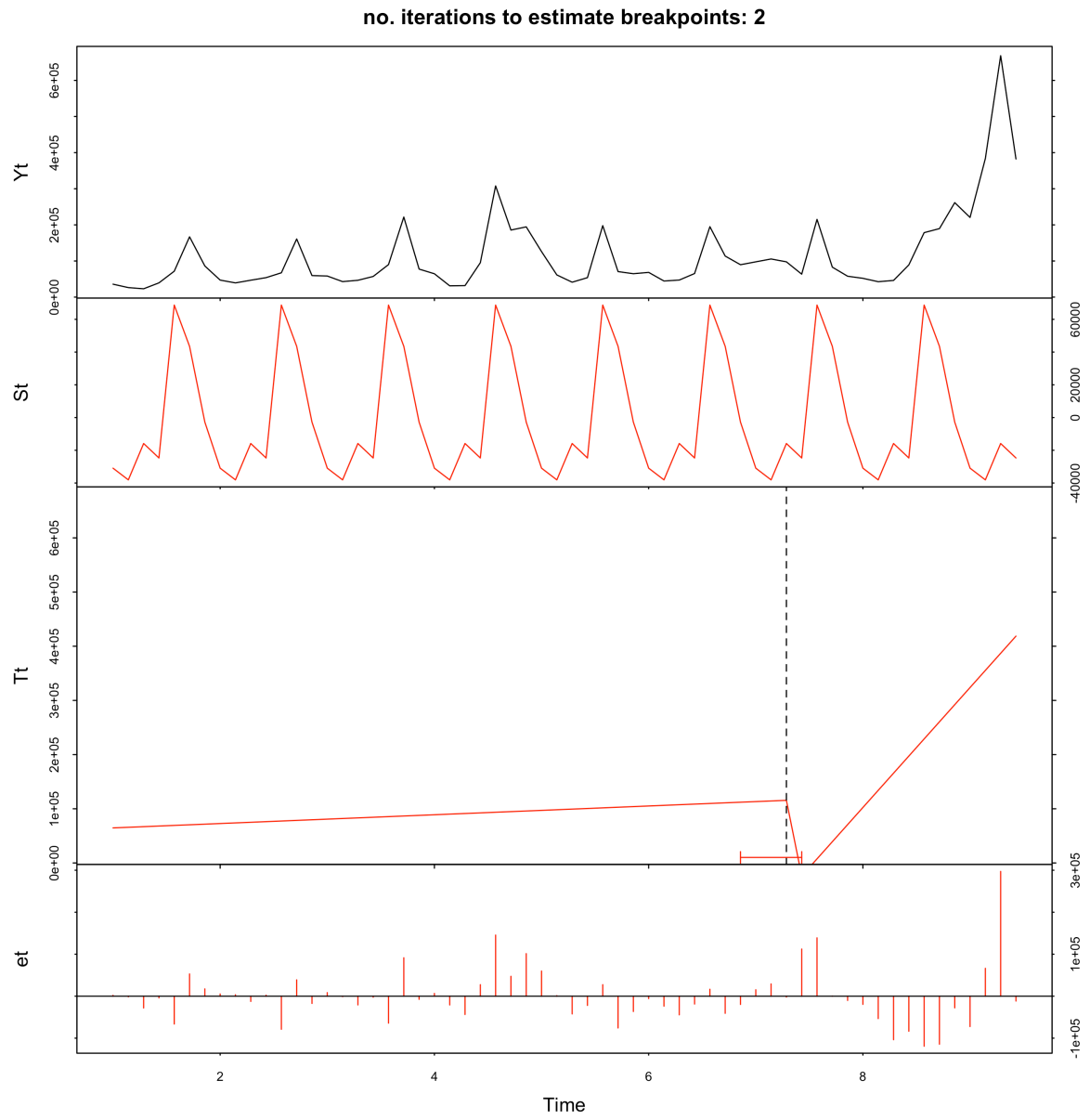

Figure 14: Adjusted number of queries decompose. St depicts the weekly cycle. BFAST method detect a change point on Aug 14 in the trend component(Tt). At the change point, the decrease may be a result of our adjustment(excluding US spider), and more observations is needed to confirm the increase afterwards.

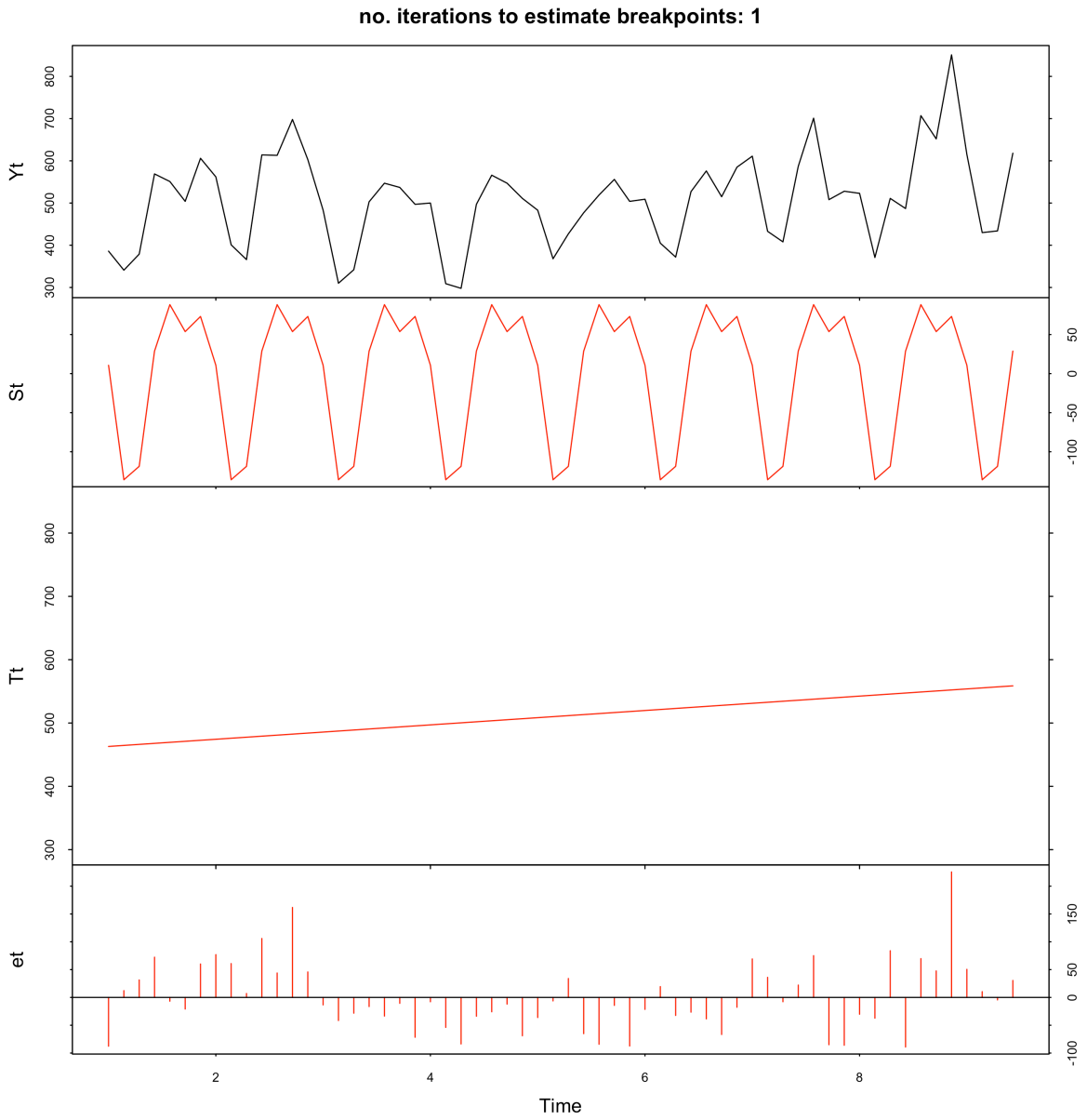

Figure 15: Number of users decompose. St depicts the weekly cycle. There is no change point detected. We also see a slightly increasing trend.

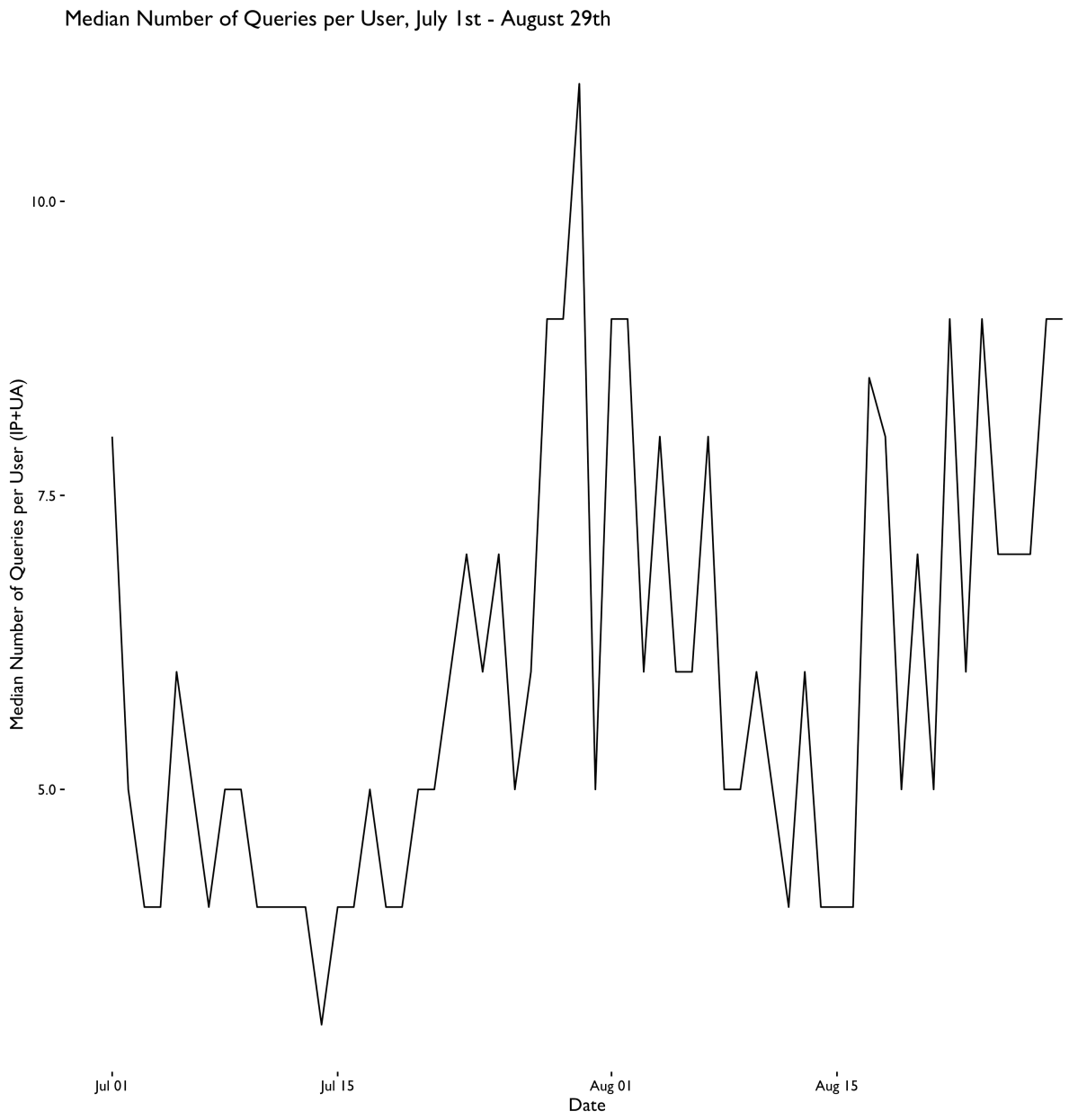

Figure 16: Median Number of Queries per User.

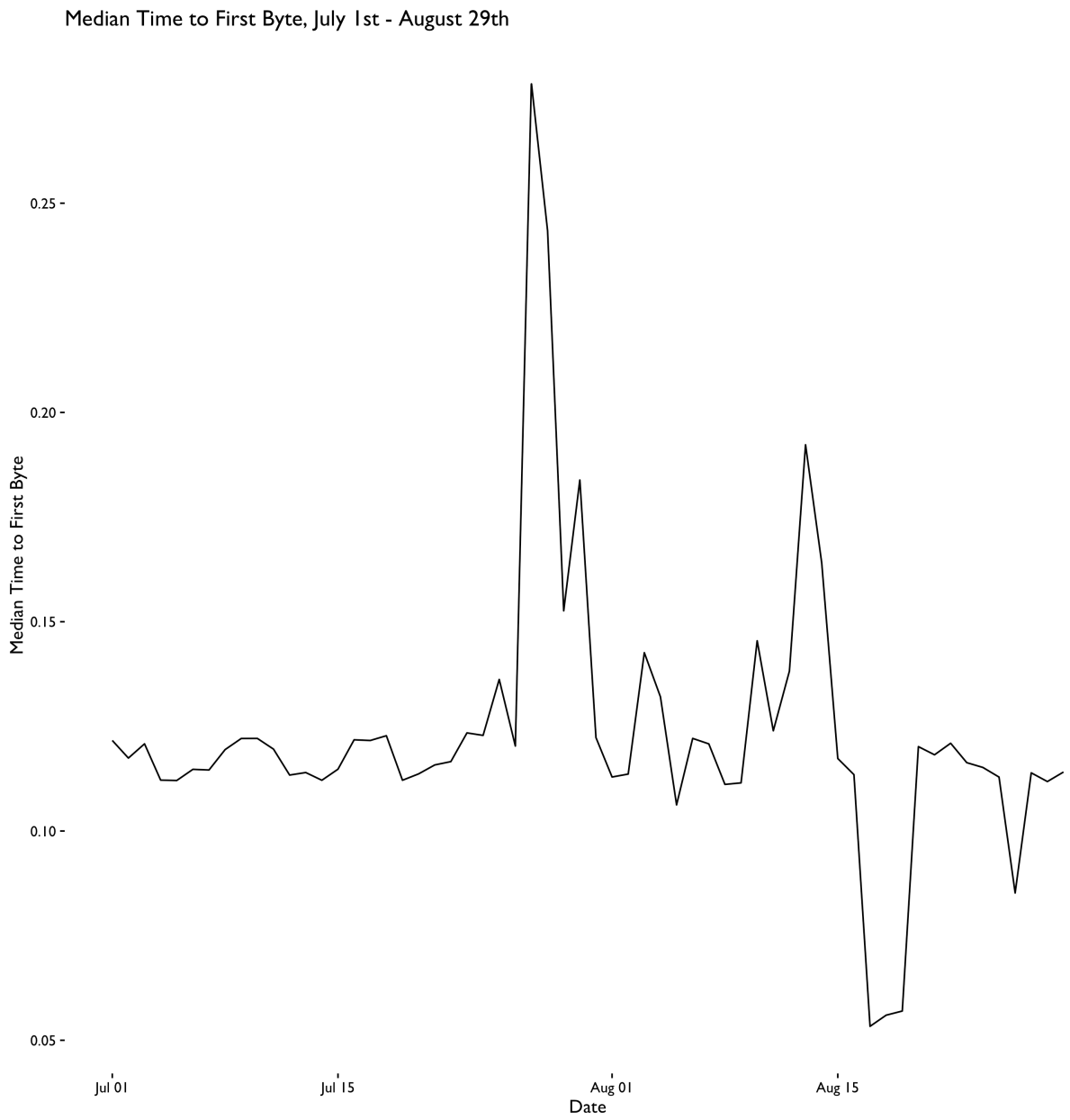

Figure 17: Median Time to First Byte.

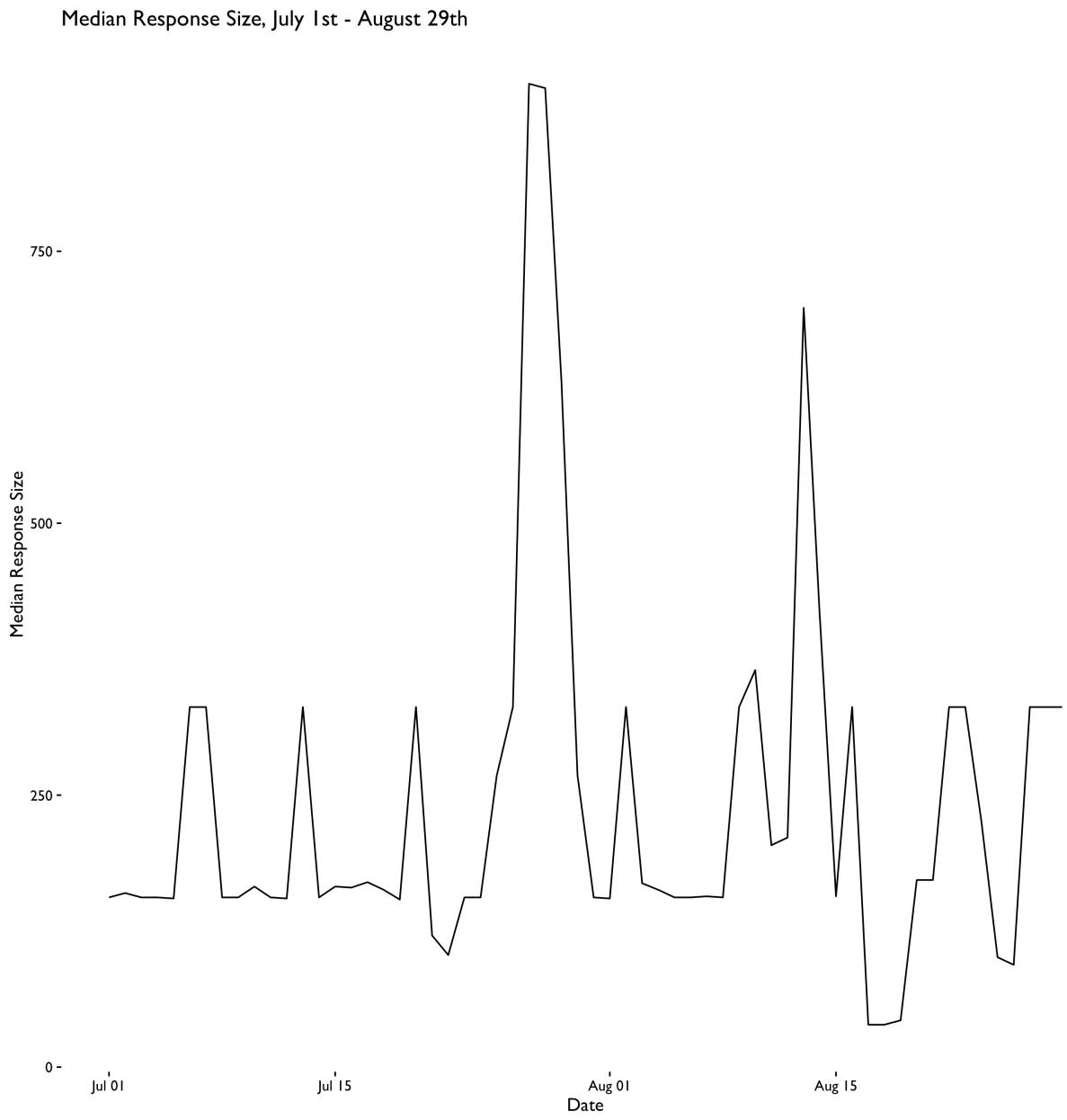

Figure 18: Median Response Size.