# Open vs. closed innovation: using online network data to measure innovation

*Benjamin Snow and Oliver Bott*

*14 November 2014*

## 1 Background

This research project examines the potential benefit of using open knowledge data in the form of collaborative online network data as an innovation indicator. By doing so, this work critically assesses current innovation indicators, namely patent data, in the hope of offering new alternatives for measuring and understanding innovation. The stated research question is: *To what extent can open innovation network data add to the measurement of innovation performance?*

For the full examination of the previous literature on the subject and reasoning for the purpose, motives, and plan for this study, please see the Research Proposal. This contribution focuses specifically on outlining the data gathering, data cleaning, and merging process. It will also examine the constructed dataset using basic descriptive statistics, as well as run preliminary inferential statistical models, offering explanation and context throughout the process. Finally, steps to improve the analysis for the final version will be discussed.

## 2 Data Gathering

To examine open network data against patent data, this study relies on two data key sources and uses the statistical tool *R* (R Core Team 2014) for the data analysis.

The first data set was obtained by using the Application Programming Interface (API) data for open networks. To examine open data innovation, data is obtained from the the git repository web-based hosting service GitHub[1]. The *R* (R Core Team 2014) packages *httr* (Wickham 2014), *dplyr* (Wickham and Francois 2014) and *rjson* (Couture-Beil 2014) allow for compiling data on the follower counts and locations associated with different users and reponsitories. In our analysis we decided to look at three follower categories (users with x followers per 10,000 population). The variable of no followers is a general indicator of GitHub use in a given location. We also include users with a follower range of 1-24 as an indicator for medium collaboration. The third category of GitHub users with more than 25 followers, which for us presents an indicator of innovative activity.

As an indicator of closed innovation we use city-level patent registration data from the Organization for Economic Co-operation and Development[2]. We use Patent Cooperation Trety (PCT) patent applications per 10,000 inhabitants on the city-level. From the same database, we also use GDP, employment and environmental data as other variables which could prove significant in explaining differences in innovation. The *R* (R Core Team 2014) package *rsdmx* (Blondel 2014) is necessary for obtaining the OECD dataset.

All data obtained via the GitHub API and OECD database can be linked to individual cities (n=120) in a total of 15 countries, allowing for an analysis on the regional level. The code used for gathering and cleaning the data is stored in a separate .R file and can be accessed via this link.

[1] Online accessilbe via https://github.com/.
[2] Online accessible via http://stats.oecd.org.

# 3 Data Sources

Table 1 depicts the variables used in the study. We use data on the city-level for a total of 120 cities based in OECD countries. As can be seen, we base our analysis on cross-sectional data with varying time frames. All of the OECD data is from 2008 except for the pollution data, which stems from 2005. The GitHub API data is taken from 2014. This inconsistency presents a limitation to the interpretation of our findings. **Limitations: GitHub only software, patent all innovation**

| Variables | Year | Source |
|---|---|---|
| Patents | 2008 | OECD |
| GDP | 2008 | OECD |
| Population | 2008 | OECD |
| Greenspace | 2008 | OECD |
| Employment | 2008 | OECD |
| Pollution | 2005 | OECD |
| No Following | 2014 | GitHub API |
| 1-24 Following | 2014 | GitHub API |
| >25 Following | 2014 | GitHub API |

# 4 Data Selection

Several potential explanatory variables are collected besides the patent and GitHub data. An overview is depicted in Table 2. These variables were selected as they were thought to potentially show cause for why innovation, be it open or closed, occurs in a certain city, but needed to be variables that would not introduce endogeneity to our model.

**Greenspace**: The Greenspace indicator specifically shows the urban greenspace in $m^2$ per capita. It was deemed potentially relevant in that with a choice of city to innovate in (assuming some level of geographic labor flexibility) there might be a recreational value necessary for attracting talent. Put another way, green cities could attract innovators.

**Pollution**: The Pollution indicator, measured in the annual average of population exposure to air pollution PM2.5 expressed in mirco gram per cubic metre, is taken both as a broad proxy for industrialization (leaving aside a discussion of to what degree pollution is from industry vs cars, etc), but also related to the Greenspace variable, that it seemed worth exploring whether a certain level of pollution discouraged talent attraction of innovators on the city level.

**Employment**: The Employment indicator, showing the employment as share of the national total, is taken largely as an indication of that city´s significance within its national context. Understanding whether a city would likely be viewed as the most prominent or significant, and whether this effects innovation, or whether innovation takes place in smaller provincial cities, is worther understanding. Additionally, seeing if the type of innovation (open vs. closed) depends on the significance of the city. An assumption which could be confirmed or disabused, for instance, is that closed innovation is more likely in prominent cities, whereas open innovation, which might require less physical presence, is more likely in less significant cities.

**GDP**: A GDP per capita indicator explores whether the size of the economy, or wealth generally, encourages innovation on the city-level, and if it is indicator of one type of innovation over another.

**Population**: A Population variable explores whether there is a necessary city size threshold which corresponds to innovation, and also was taken for controlling for across cities, to find patent data or GitHub data per a number of people in a city. Without this control, GitHub followers and patent data would likely simply correspond to the population of the city, which would be less instructive.

## 5 Prelimianary Descriptive Statistical Analysis

The summary statistics of the variables show wide ranging distributions. Since the data cleaning eliminated all values equal to or lower than zero, a log transformation seems to be a strategy that could strengthen our analysis.

Table 2: Summary statistics

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Patents | 120 | 1.67 | 1.36 | 0.07 | 6.28 |
| GDP | 120 | 36,877.00 | 8,288.00 | 17,665.00 | 61,804.00 |
| Population | 120 | 2,205,763.00 | 3,710,055.00 | 500,350 | 34,482,742 |
| Greenspace | 120 | 634.40 | 969.70 | 1.13 | 5,081.00 |
| Employment | 120 | 4.51 | 7.98 | 0.18 | 39.39 |
| Pollution | 120 | 16.15 | 5.20 | 5.85 | 31.44 |
| nofollowing | 120 | 3.81 | 3.59 | 0.16 | 23.16 |
| medfollowing | 120 | 1.92 | 1.94 | 0.06 | 10.78 |
| hifollowing | 120 | 0.20 | 0.26 | 0.01 | 1.45 |

The *car* package (Fox and Weisberg 2011) is used to examine the relationship, distribution, and normality of all variables included in the model, to understand which log transformation and regression model would be most appropriate. The distribution of many variables are highly skewed (see Figure 1). This finding supports our attempt to undertake a log transformation of the variables used.

The residual plot between patents and users with high follower numbers (see Figure 2) depicts a relatively random pattern, which indicates that a linear regression model provides a decent fit to the inferential statistics of the data set.

## 6 Inferential Statistics

This study plans to use an ordinary least squares (OLS) model to examine the relationship between patent and highly followed open data sources. We use this model for our regression analysis:

$$logP_i = \beta_0 + \beta_1 logF_i + \beta_2 logGDP_i + \beta_3 logPop_i + \beta_4 logG_i + \beta_4 logE_i + \beta_4 logPol_i + \epsilon_i$$

$P$ is the patent intensity expected in a given city $i$. While in our setting it is not possible to show causal effects of the relationship between the variables, demonstrating that one is acting on the other is already an important finding in itself. The findings endorse the implicit hypothesis of the study that open data sources do seem to show innovation in a similar but perhaps distinct manner to patent data. This found relationships offers a glimps into the ‘throughput’ of innovation rather than the ‘output’ which patent data reflects. However, this will need to be examined further in the final paper.

Now that a relationship between patent and follower data has been shown, further inferential statistical analysis attempts to find the common predictor or cause of innovation in both patent and open data is necessary. This is done by running identical OLS regressions, using the the log of both patent and highly followed GitHub users as the dependant variables, and the log of each of the previously mentioned independant variables (GDP, Population, Greenspace, Employment, and Pollution), to examine if any of these variables prove significantly (and presumably, similarly) correlated to our dependant variables.

This process has allowed us first directly compare our previously established (patent) and newly hypothesized (open - github) measures of innovation. After comparing directly, attempting to find how they differently

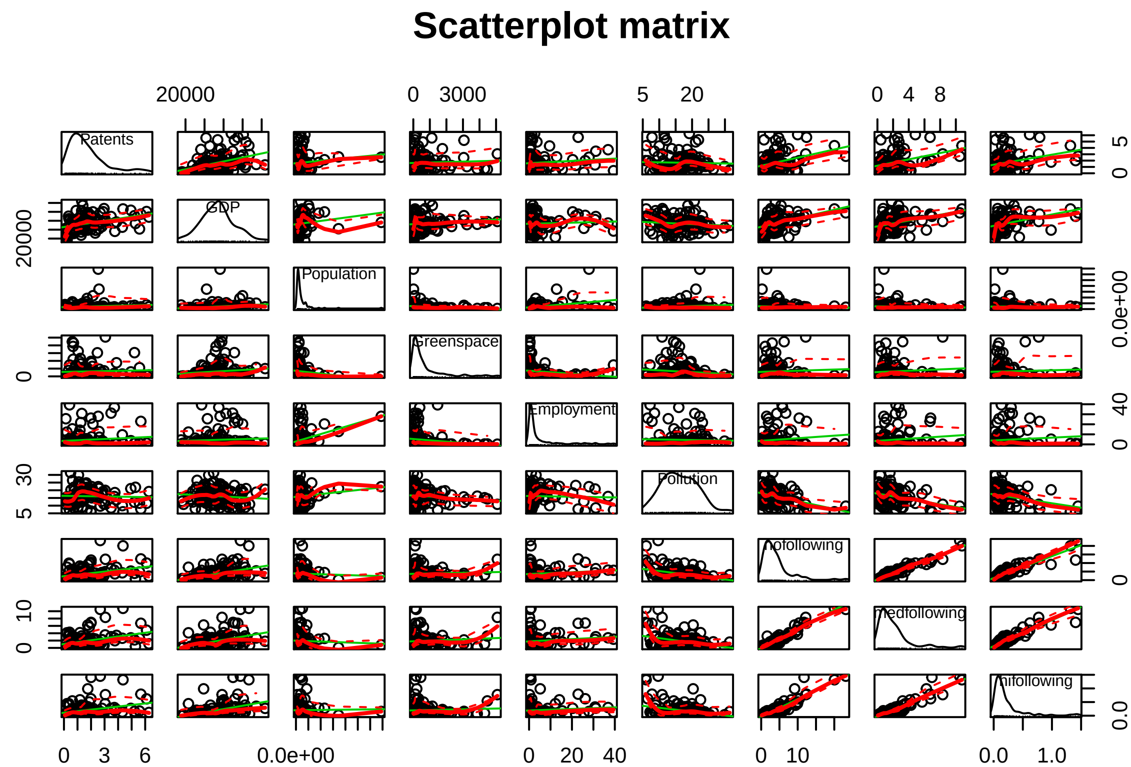


Figure 1: plot of chunk unnamed-chunk-5

Table 3: Regression Estimates of Patent Activity

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | log(Patents) | | | |
| | (1) | (2) | (3) | (4) |
| (Intercept) | 0.32*** (0.08) | | | |
| No Following | | 0.34*** (0.08) | | |
| 1-24 Following | | | 0.34*** (0.07) | 0.15* (0.08) |
| >25 Following | | | | 1.76*** (0.41) |
| GDP | | | | −0.06 (0.11) |
| Population | | | | 0.05 (0.05) |
| Greenspace | | | | 0.08 (0.06) |
| Employment | | | | 0.49** (0.21) |
| Pollution | −0.13 (0.11) | 0.10 (0.08) | 0.89*** (0.17) | −18.73*** (3.92) |
| Observations | 120 | 120 | 120 | 120 |
| $R^2$ | 0.11 | 0.15 | 0.15 | 0.34 |
| Adjusted $R^2$ | 0.10 | 0.14 | 0.15 | 0.31 |
| Residual Std. Error | 0.87 (df = 118) | 0.85 (df = 118) | 0.85 (df = 118) | 0.76 (df = 113) |
| F Statistic | 14.72*** (df = 1; 118) | 20.50*** (df = 1; 118) | 21.58*** (df = 1; 118) | 9.88*** (df = 6; 113) |
| *Note:* | | | | *p<0.1; **p<0.05; ***p<0.01 |

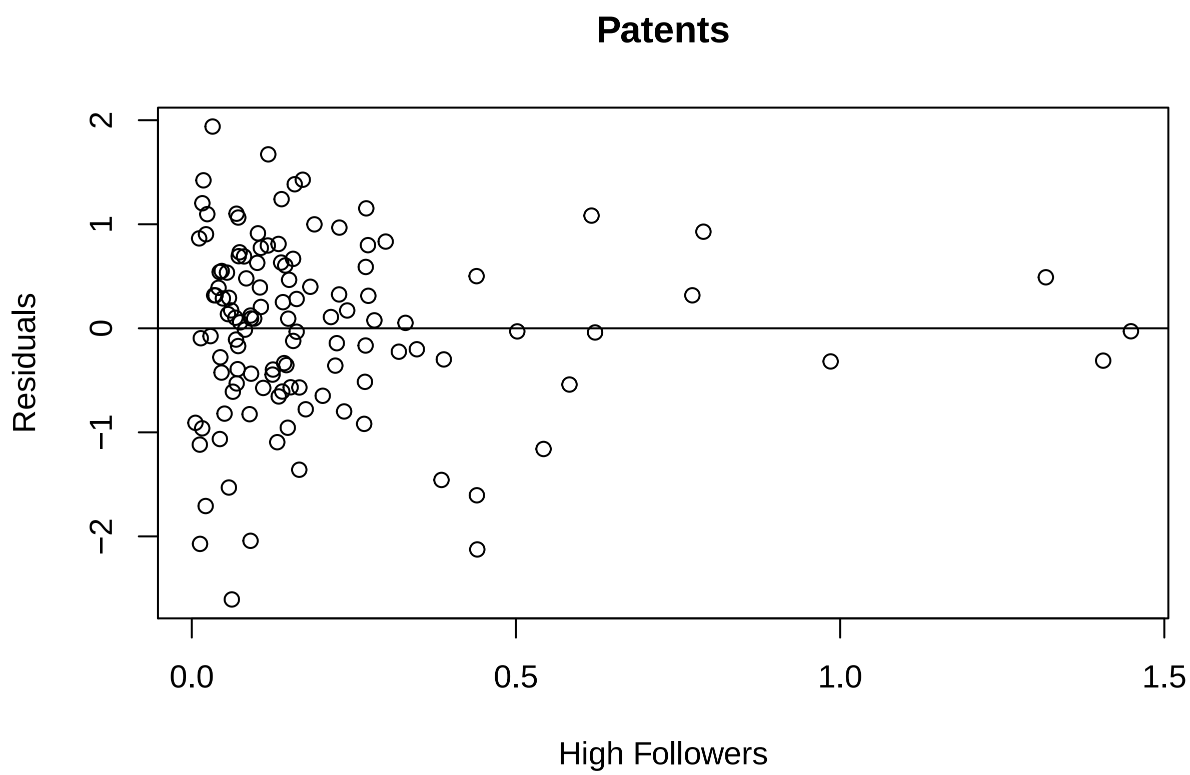


Figure 2: plot of chunk unnamed-chunk-6

reflect innovation, as well as attempting to find the common cause of innovation for both the innovation throughput which is open data, and the innovation output which is patent data, is necessary.

In a second model, we estimate the effect of various variables on follower numbers, again using log transformations:

$$logF_l = \beta_0 + \beta_1 logP_l + \beta_2 logGDP_l + \beta_3 logPop_l + \beta_4 logG_l + \beta_4 logE_l + \beta_4 logPol_l + \epsilon_l$$

Table 4: Regression Estimates of Follower Numbers

| | *Dependent variable:* | |
|---|---|---|
| | log(hifollowing) | |
| | (1) | (2) |
| (Intercept) | 0.46*** (0.10) | 0.18* (0.10) |
| Patents | | 2.21*** (0.45) |
| GDP | | −0.19 (0.12) |
| Population | | 0.04 (0.05) |
| Greenspace | | 0.09 (0.07) |
| Employment | | −0.65*** (0.23) |
| Pollution | −2.21*** (0.09) | −21.19*** (4.38) |
| Observations | 120 | 120 |
| $R^2$ | 0.15 | 0.39 |
| Adjusted $R^2$ | 0.15 | 0.36 |
| Residual Std. Error | 0.98 (df = 118) | 0.85 (df = 113) |
| F Statistic | 21.58*** (df = 1; 118) | 12.01*** (df = 6; 113) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

Here is **Some interpretation**.

# 7 Outlook

From this analysis it becomes apparent that the introduction of various dummy controls could help explain the spurious relationship between patent and network data. It would hence make sense to control for English speaking countries, as we would suspect the spread of GitHub to be greatest there. Also, we could introduce dummy controls for the overall economic development of the country, assuming that software development is clustered in these locations. In addition, including a map visualization with information on location of the cities in our sample could greatly improve our work.

# References

Blondel, Emmanuel. 2014. *Rsdmx: Tools for Reading SDMX Data and Metadata*. http://CRAN.R-project.org/package=rsdmx.

Couture-Beil, Alex. 2014. *Rjson: JSON for R*. http://CRAN.R-project.org/package=rjson.

Fox, John, and Sanford Weisberg. 2011. *An R Companion to Applied Regression*. Second. Thousand Oaks CA: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Wickham, Hadley. 2014. *Httr: Tools for Working with URLs and HTTP*. http://CRAN.R-project.org/package=httr.

Wickham, Hadley, and Romain Francois. 2014. *Dplyr: Dplyr: A Grammar of Data Manipulation*. http://CRAN.R-project.org/package=dplyr.