# 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 given.

## 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.

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 PCT patent applications per 10,000 inhabitants. From the same database, we also use GDP per capita data and environmental data, as other variables which could prove significant in explaining differences in innovation. From the same OECD database: GDP per capita, pollution levels, greenspace, and employment (as share of national employment) are taken for the relevant cities which github and patent data are obtained for. These other variables are inclued in the analysis as they 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, allowing for an analysis on the regional level. The code used for gathering and cleaning the data can be accessed via this link.

## 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.

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

| 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 1. 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 be endogenous (is this the right word) to patent or github data.

**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

A first examination of the distribution of each variable show distributions which reflect a ..... An examination of the distribution of both the patent applications and github users in each city are taken to examine if a similar distribution can be seen in cities when normalizing for population size. A similar distribution would indicate ... (Olli, question here, if its examining when it is only no following, this isnt giving us total user numbers, just user numbers of those who dont have any following, right?)

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 |

Next, 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 regression model would be most appropriate. Taking into account the clumped distribution of many variables, taking the log regression of several variables when regressing is attempted, to see if this brings about a more normal distribution.

Here is some **Interpretation of scatterplot. Hence log variables.**

Here is some **Interpretation of residual plot**. The relatively random pattern indicates that a linear regression model provids a decent fit to the data.

# 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. The high level of significance between these two variables, rather than demonstrating that one is acting on the other, show that some other 'spur of innovation' is acting on both, but does 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 is exceptionally early analysis and will need to be examined further.

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 Ordinary Least Squares 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 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.

Here is **Some interpretation**.

Here is **Some interpretation**.

# 7 Outlook

**What have learned from preliminary analysis?** Use various dummies. We suspect that country and economic development dummies can explain the spurious relationship between patent and network data. Also

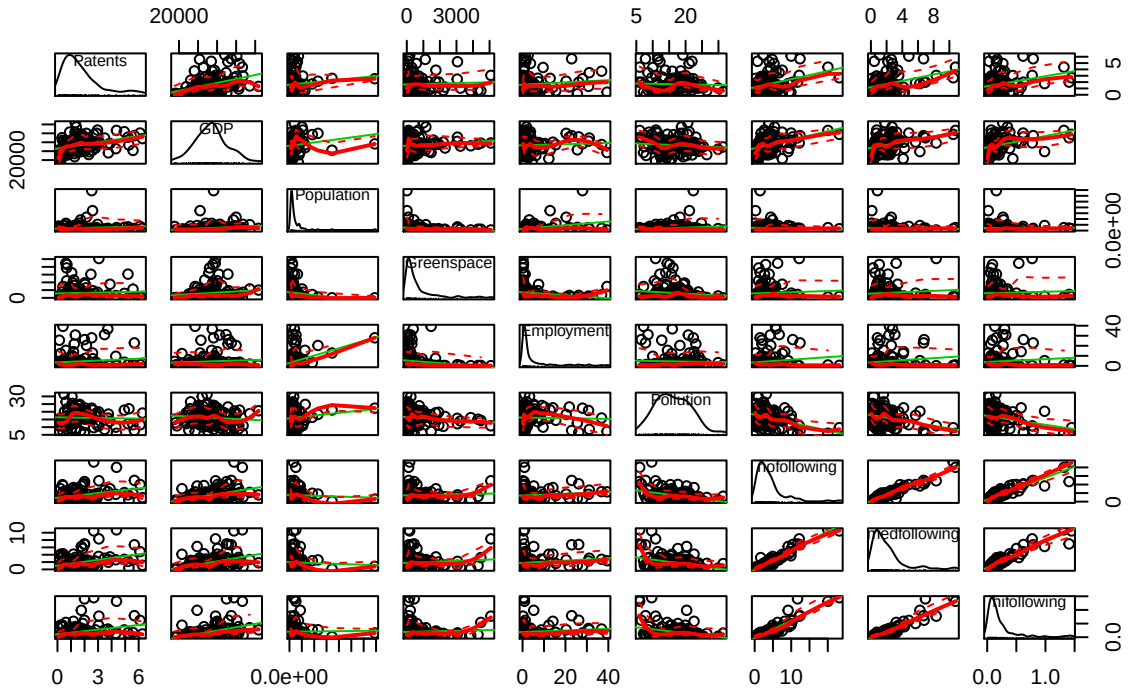


Figure 1: plot of chunk unnamed-chunk-5

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 |

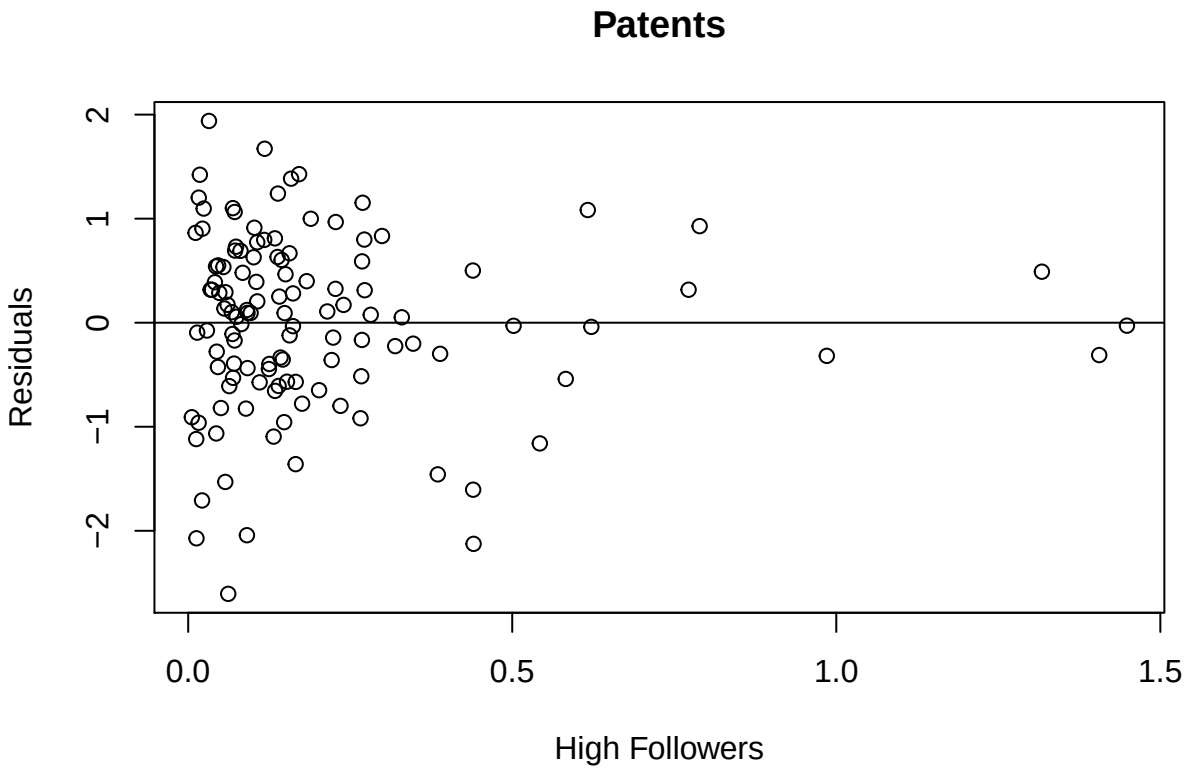


Figure 2: plot of chunk unnamed-chunk-6

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 |

a map would be nice for visualisation.

## 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.